COLLECTED PAPERS OF KENNETH J. ARROW

Applied Economics

The Belknap Press of Harvard University Press
Cambridge, Massachusetts, and London, England 1985

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

This book is printed on acid-free paper,
and its binding materials have been chosen
for strength and durability.

Library of Congress Cataloging in Publication Data

Arrow, Kenneth Joseph, 1921–
Applied economics.

(Collected papers of Kenneth J. Arrow; v. 6)
Includes index.
1. Economics—Addresses, essays, lectures. I. Title.
II. Series: Arrow, Kenneth Joseph, 1921– . Essays;
v. 6.
HB171.A82 1985 330 85-1385
ISBN 0-674-13778-7

Preface

Although my focus and my comparative advantage are certainly on the theoretical side of economic reasoning, I have kept a fairly steady interest in applied problems, some of a relatively specific nature. Perhaps my most important motive was a desire to show by example that economic reasoning could be of real use. (The same motive could be expressed as one of guilt for the enjoyment I had in my work and a feeling that it could only be justified by its social consequences.) Another motive was the need for new problems to stimulate theory. The corpus of neoclassical theory is rich enough to spin off a significant number of research problems endogenously. Improved formulations of general equilibrium theory and of welfare economics were obvious intellectual needs when I started my work. But clearly there is a narrow limit to the range of problems that come out of earlier theorizing; any theoretical subject needs the constant stimulus of new problems drawn from the real world.

The actual applied subjects with which I dealt have been drawn from concerns of the time, as they should be, and are therefore very diverse. The immediate impulses sometimes came from requests by others, sometimes from my own directions. The papers on medical economics have probably been the most consequential in posing new theoretical problems, since many of the issues of the economics of information and particularly what are now called incentive compatibility issues were raised there sharply, though hardly resolved. The studies on racial discrimination in labor markets raised a variety of theoretical issues, some related to the economics of information and some to nonconvexities.

It is doubtful that any of these papers has had a significant effect on actual policy, but perhaps that is too much to hope for.

I should like to thank Janet Marantz for her careful editing and Michael Barclay and Robert Wood for preparation of the index and for help in the proofreading.

Contents

1 The Economic Cost to Western Europe of Restricted Availability of Oil Imports: A Linear Programming Computation

I had spent a number of summers at the RAND Corporation during the 1950s and later. One day, I believe in 1961, Harold Lubell, then a regular staff member of the RAND Corporation, came into my office. He had been editing a volume on the effects of Middle East oil crises on the European economy. (The date reminds us that the events of the 1970s were certainly a foreseen possibility. The Preacher told us, "There is nothing new under the sun.") It had occurred to him that the then fairly new tool of linear programming might have something to say on the subject, and he appealed to me for a quick study. The model is crude and naive in some ways. But the conclusion, that even a drastic reduction in oil supplies is not a fatal blow to an advanced economy, is robust. Some of the hysteria of 1974 might have been avoided if it had been believed. The big lesson is the very considerable degree of substitutability in an economy, a lesson that might also have been learned from the failure of strategic bombing in World War II.

The basic idea of this approach may be summarized as follows. Suppose that Western Europe is maximizing its gross domestic product subject to appropriate constraints. The constraints are taken to be the following: (1) the aggregate amount of value added, taken as an index of capital and labor together, is given; (2) the domestic capacities of the energy

Reprinted from *Middle East Oil Crises and Western Europe's Energy Supplies*, ed. H. Lubell (Baltimore: Johns Hopkins University Press, 1963), appendix D, pp. 214–220.

sector and the food and agriculture sector are limited; (3) the net exports (excess of exports over imports) in any sector are limited (by the willingness of the outside world to import); (4) the aggregate export surplus is limited from below (a balance of payments equilibrium condition); and (5) there is an upper bound on the amount of (net) energy imports (which we are, in effect, identifying with oil imports).

The basic assumptions about the structure of the economy are the following: (1) production is organized into five sectors: energy, food and agriculture, industry and construction, transportation, and other services; (2) in each sector the demand for the services of the other sectors and for value added is governed by fixed input-output coefficients; and (3) domestic final demand for each sector is proportional to total national income.

The bounds used in constraints (1)–(4) are taken to be the actual values in 1953; that is, we assume the absence of any unemployment (in the sense of value added), of any unused capacity in the energy sector and the food and agriculture sector, of any unexploited markets for exports, and of any excess of export surplus over the amount needed to meet obligations. The true bounds can only differ from these by being less restrictive. To the extent that these degrees of freedom are open to the economy in its adjustment to a reduction in oil imports, the cut in national income must be less than that found here.

The upper bound in constraint (5) will be taken as a parameter that can take on different values, ranging from the actual 1953 value down to zero. If our model is correct, then the linear programming solution should yield the actual 1953 value of national income and the actual 1953 values of the different economic variables (such as the outputs of the different sectors) when the upper bound for net imports of energy is set equal to its 1953 value. The difference between the actual national income for 1953 and the optimal value found when the upper bound on energy imports is set at zero is the total cost to Western Europe of a complete cessation of oil imports. The costs of a partial cutting off of imports could be determined similarly by comparing the actual 1953 value with the solution to the linear programming problem when the upper bound on energy imports is set at an intermediate value.

Strictly speaking, it is obvious that these answers, even if otherwise correct, refer only to the change that would have occurred in 1953 had the oil imports been restricted to some lower value than the one actually achieved.

Before going into more detail on the formulation of the problem, let us state the results of the linear programming computation. A complete cessation of oil imports would have led to a 5.2 percent reduction in gross domestic product, or roughly $9 billion. Lesser restrictions from the original net imports of $1.1 billion would have had proportionately smaller effects. This of course leads to an equal reduction in value added used, which, in this model, means an equivalent unemployment of capital and labor. There is a proportionate reduction in domestic final demand for each sector. The foreign trade structure is somewhat altered. The net exports of industry and construction and of other services remain unaltered. Transportation, a major user of energy which actually had exports of $1.1 billion in 1953, would shift to the position of a net importer of $2.6 billion if all oil imports were cut off. The net imports of food and agriculture would fall by $2.6 billion, from $7.3 billion to $4.7 billion. Since net energy imports were $1.1 billion and have changed to zero by assumption, the trade balance remains unchanged, as required.

Because of the rigid input-output coefficients and rigid demand coefficients assumed, it is plausible that the model overstates the loss, especially in the long run. An optimal adjustment policy would use the unemployed value added in various ways that increase gross domestic product and so compensate for the loss in oil imports. For example, exports might be increased by reducing prices in industry and construction; this would be profitable under conditions of unemployment. Or the unemployed resources could be used to increase capacity in the energy and the food and agriculture sectors; or there might be a shift in domestic relative prices, which would induce a shift in domestic consumption toward less oil-using industries. These adjustments are outside the model but would occur in real life. Further evidence that the model overstates the loss is the very large discrepancy between the $9 billion value that the model places on the oil imports and the $1.1 billion actually paid for them. This would suggest that an increase in oil imports would have been very profitable in 1953.

We now spell out the model in detail. Let

t_0 = gross domestic product, as a ratio to the actual 1953 value,
f_i = domestic final demand for sector i,
e_i = net exports for sector i,
x_i = domestic output for sector i,
a_{ij} = input of commodity i for unit domestic output of commodity j,

v_i = value added per unit output of commodity i,
b_{ij} = direct and indirect demand for commodity i for a unit increase in final demand for commodity j (inverse of Leontief matrix),
$\bar{v}$ = actual total value added in 1953,
$\bar{x}_i$ = actual domestic output for sector i in 1953,
$\bar{f}_i$ = actual domestic final demand for sector i in 1953,
$\bar{e}_i$ = actual net exports of sector i in 1953, if positive, and 0 otherwise,
$\bar{e}$ = actual export surplus in 1953.

The aim is to maximize t_0, subject to the constraints set out below.

The condition that domestic final demand be in fixed proportions to national income can be written

(1-1) $$f_i = t_0 \bar{f}_i.$$

The condition that aggregate supply and demand balance for each sector is, under the assumption of constant coefficients,

(1-2) $$\sum_{j=1}^{5} a_{ij} x_j + f_i + e_i = x_i.$$

Equation (1-2) can be solved for the outputs, x_i, in terms of the final demands, $f_j + e_j$:

(1-3) $$\sum_{j=1}^{5} b_{ij}(f_j + e_j) = x_i.$$

We assume capacity constraints on sectors 1 and 2 (energy, and food and agriculture). These can be written $x_i \leqq \bar{x}_i$, where $i = (1, 2)$; or, in view of (1-3),

(1-4) $$\sum_{j=1}^{5} b_{ij} f_j + \sum_{j=1}^{5} b_{ij} e_j \leqq \bar{x}_i \qquad (i = 1, 2).$$

The total value added used is

$$\sum_{i=1}^{5} v_i x_i,$$

which, from (1-3), can be written

$$\sum_{i=1}^{5} v_i \left[\sum_{j=1}^{5} b_{ij}(f_j + e_j) \right] = \sum_{j=1}^{5} \left(\sum_{i=1}^{5} v_i b_{ij} \right) (f_j + e_j).$$

Let us define

(1-5) $$m_j = \sum_i v_i b_{ij},$$

where m_j is the direct and indirect demand for value added for a unit increase in final demand for commodity j. Then the constraint that aggregate value added not exceed that available in 1953 becomes

(1-6) $$\sum_j m_j f_j + \sum_j m_j e_j \leqq \bar{v}.$$

The constraint that exports be limited to their 1953 values implies that in any industry in which net exports are positive they cannot exceed those values. But if net exports were negative in 1953, that is, for a sector that imports, the limit is zero, since this can be achieved by cutting imports, a policy that is under the control of Western Europe, rather than by increasing exports, which are exogenously determined. Thus, in view of the definitions, this constraint becomes

(1-7) $$e_i \leqq \bar{e}_i.$$

The balance of payments constraint requires that total net exports not fall below the 1953 value (which presumably was a minimum necessitated by some constraints):

(1-8) $$\sum_i e_i \geqq \bar{e}.$$

Finally, we are going to impose a limitation on imports of energy. Since we want to consider different possible limitations, we will denote the upper bound by a parameter symbol, γ. The constraint is then

(1-9) $$-e_1 \leqq \gamma.$$

Note that net imports are $-e_1$.

The problem is to choose the variables f_i, t_0, and e_i to maximize t_0 subject to the constraints (1-1), (1-4), (1-6), (1-7), (1-8), and (1-9). This is a linear programming problem. It can be simplified by some change in notation and by substituting from (1-1) into (1-4) and (1-6). First, define a new set of variables:

(1-10) $$t_j = \bar{e}_j - e_j \qquad (j = 1, 2, 3, 4, 5).$$

The condition (1-7) then becomes simply the condition

(1-11) $$t_j \geqq 0,$$

which is automatically handled by a linear programming routine. We eliminate e_j from (1-4), (1-6), (1-8), and (1-9) by the formula $e_j = \bar{e}_j - t_j$. At the same time, we eliminate f_j by use of (1-1). To help, let us define

$$(1\text{-}12) \qquad a_i = \sum_j b_{ij}\bar{f}_j \qquad (i = 1, 2),$$

and

$$(1\text{-}13) \qquad b = \sum_j m_j\bar{f}_j.$$

If the indicated substitutions are performed in (1-4) and (1-6) and use is made of (1-12) and (1-13), respectively, we have

$$(1\text{-}14) \qquad a_i t_0 - \sum_{j=1}^{5} b_{ij}t_j \leqq \bar{x}_i - \sum_{j=1}^{5} b_{ij}\bar{e}_j \qquad (i = 1, 2),$$

and

$$(1\text{-}15) \qquad bt_0 - \sum_{j=1}^{5} m_j t_j \leqq \bar{v} - \sum_{j=1}^{5} m_j\bar{e}_j.$$

Table 1.1 Input-output flows for Western Europe, 1953 (billions of dollars)

	To:						
From:	Energy	Food and agri-culture	Industry and con-struction	Trans-portation	Other services	Domestic final demands	Ex-ports
1. Energy	—	0.7	5.5	1.6	1.0	5.0	0.5
2. Food and agriculture	—	—	5.1	—	—	44.4	2.1
3. Industry and construction	1.2	4.1	—	1.0	4.2	59.5	11.7
4. Transportation	1.1	1.3	2.8	—	0.4	5.9	2.3
5. Other services	0.2	2.5	4.8	0.6	—	62.1	5.6
Imports	1.6	9.4	5.1	1.2	4.2		
Value added	9.6	33.6	54.8	8.9	61.0		
Total output	14.3	51.6	81.7	13.8	75.8		
Domestic output[a]	12.7	42.2	76.6	12.6	71.6		

Source: E. S. Kirschen, *The Structure of the European Economy in 1953* (Paris: OEEC, 1958), pp. 8, 53.
a. Domestic output = total output less imports.

Constraint (1-8) becomes

$$(1\text{-}16) \qquad \sum_{j=1}^{5} t_j \leqq \sum_{j=1}^{5} \bar{e}_j - \bar{e}.$$

Finally, since $\bar{e}_1 = 0$ (because Western Europe was a net importer of energy), $t_1 = \bar{e}_1 - e_1 = -e_1$, and so (1-9) is written

$$(1\text{-}17) \qquad t_1 \leqq \gamma.$$

Thus the problem is that of finding values of the nonnegative variables $t_0, \ldots, t_5$ to maximize t_0 subject to constraints (1-14)–(1-17).

It remains to get the data for this problem. These are derived from the tentative OEEC input-output table for 1953, the relevant parts of which are shown in Table 1.1.

The quantities $\bar{f}_j$ are in the column headed "domestic final demands." The values $\bar{x}_i$ are the domestic outputs (for sectors 1 and 2). The quantity $\bar{v}$ is the sum of the elements in the value added row; $\bar{v} = 167.9$. The quantity $\bar{e}$ is the sum of elements in the exports column *less* the sum of the elements in the imports row; $\bar{e} = 0.7$. The quantity $\bar{e}_j$ is obtained by subtracting the jth import figure from the corresponding export, replacing negative differences by 0, as follows (in billions of dollars):

Sector	1	2	3	4	5
$\bar{e}_j$	0	0	6.6	1.1	1.4

Table 1.2 Input-output coefficients for Western Europe, 1953

	To:				
From:	Energy	Food and agriculture	Industry and construction	Transportation	Other services
1. Energy	—	0.017	0.072	0.127	0.014
2. Food and agriculture	—	—	0.067	—	—
3. Industry and construction	0.094	0.097	—	0.097	0.059
4. Transportation	0.087	0.031	0.037	—	0.006
5. Other services	0.016	0.059	0.063	0.048	—
Value added	0.756	0.796	0.715	0.706	0.852

From Table 1.1 we can derive the table of input-output coefficients by dividing each column by domestic output, as set out in Table 1.2. If in Table 1.2 the diagonal elements are replaced by 1 and the signs of the off-diagonal elements reversed, the inverse of the resulting matrix yields the values of b_{ij}. The m_j's are computed from these and the v_i's (also given in Table 1.2) by (1-5). The quantities a_i ($i = 1, 2$), b can then be computed by (1-12) and (1-13). Thus all the coefficients appearing in the constraints (1-14)–(1-17) have been calculated, except for the upper bound γ in (1-17). The actual 1953 value of t_1, which is net imports of energy, was 1.1, from Table 1.1. Hence, we let γ range down from 1.1 to zero. For $\gamma = 1.1$, it was found that $t_0 = 1$, thereby confirming the model to some extent, since it "predicted" the actual value of gross domestic product (recall that t_0 is the ratio of domestic product to the actual 1953 value). For $\gamma = 0$, t_0 was approximately 0.95.

2 Comment on Duesenberry's "Portfolio Approach to the Demand for Money and Other Assets"

James Duesenberry in his encyclopedic coverage of the movements of holdings of cash and of liquid assets has listed many factors, but three elements appear and reappear: increasing returns to scale in transactions (whether between cash and other liquid assets or between liquid and nonliquid assets); the conveniences of having cash (and, by derivation, the conveniences of liquidity); and risk aversion. I will state some results in theory of choice under uncertainty which bear on the more precise interpretation of these elements and their implications for behavior.

Two branches of the theory of choice under uncertainty will be drawn on here: stochastic multiperiod inventory theory and the theory of risk aversion.

Inventory Theory

There has developed in the last fifteen years a theory of optimal behavior for the holding of inventories, where the firm is facing repeated uncertain demands and has repeated opportunities to purchase inventories.[1] In its simplest form we assume discrete time periods. At the beginning of each period, the firm has a given stock of inventory. It is then faced with a

1. See P. Massé, *Les réserves et la régulation de l'avenir*, Paris, Hermann, 1946; K. J. Arrow, T. E. Harris, and J. Marschak, "Optimal Inventory Policy," *Econometrica*, 19 (1951), pp. 250–272; K. J. Arrow, S. Karlin, and H. Scarf, *Studies in the Mathematical Theory of Inventory and Production*, Palo Alto, Stanford University Press, 1958.

Reprinted from *Review of Economics and Statistics*, 45 suppl. (1963):24–27.

demand, which is a drawing from a probability distribution. The demand is met, at least to the extent possible, and there is a penalty for the shortfall, if any. The firm then places an order for as much more stock as it wishes, and the new order, plus whatever stock may have been left over after meeting the demand, constitutes the stock on hand at the beginning of the next period. The cost for the period is the sum of the penalty and the ordering cost (plus possibly also storage cost); this is a random variable, since both the penalty and the amount ordered may depend on the random demand. The aim of policy is to minimize the sum over time of discounted expected costs.

Two observations may be made on these assumptions. In the first place, since only the expected value of returns is considered, we are assuming risk neutrality. This is an assumption appropriate for a large corporation, where the risk is divided and, in general, small for each stockholder relative to his total wealth, but not for an individual.

In the second place, the whole problem arises because goods cannot be procured instantaneously, at least not without extra cost. If stock could be ordered after the demand is known, there would be no costs beyond the irreducible minimum for ordering the goods. One interpretation of the penalty is the price of immediate delivery of the goods, which is higher than the ordering cost for lagged delivery. The difference between the penalty and ordering costs is a form of transaction cost, which we may term a *timed transaction cost.*

The interpretation of the inventory model relevant for the present discussion is that in which the good in question is cash or possibly liquid assets, and the firm is subject to random demands for cash, which cannot be replenished immediately without additional cost.[2] The optimal policy depends on the nature of the ordering cost function. If cash can be obtained at a cost proportional to the magnitude demanded (constant marginal cost), the optimal policy is to set a target level of cash and then, in each period, after the demand, acquire enough cash to bring the stock up to the target level. The target level depends on the penalty for shortfall, the rate of interest (which is the penalty for excessive holdings of cash), and the probability distribution of demand. The fluctuation in the observed cash holdings, especially if the observation is made just after the demand has occurred, will indeed, as Duesenberry notes, reflect primarily the fluctuations in demand.

2. The model analyzed by D. Patinkin (in *Money, Interest, and Prices,* Evanston, Illinois, Row, Peterson, 1956, chap. 7) is similar to a one-period version of the above model; there may be a whole sequence of demands and payments at random times within a period in which no cash replenishment at all can be carried out.

However, the target may change because of changes in the rate of interest; it may also change because of changes in the subjective probability distribution of demands, which in turn may be influenced by the observations. The latter movement will be relatively slow, since conviction of a change in the probability distribution will be obtained only after a series of observations confirm it.

A second case is that in which the ordering cost contains a fixed component as well as a proportional one. However, the ordering cost is zero if no order is placed. This corresponds to a psychic cost of decision making or to administrative cost or to some other form of economies of scale in handling cash. This is again a transaction cost but of a different kind from timed transaction costs; we may call it a *fixed transaction cost,* in that it does not depend on the magnitude of the transaction.[3]

The optimal policy here is of the two-bin or S,s type, to use the terms in the literature. There are two levels, a target S and a reorder point s; cash is acquired only if reserves fall below s, but when they do, enough cash is acquired to bring the stock up to S.[4] As can be seen, this implies a stickiness in the response to falling cash reserves; only beyond a certain level is there a response, but the response is apt to be large. The difference, $S - s$, which is roughly the size of the order (actually the order is usually somewhat larger), is under certain conditions roughly proportional to the square root of mean cash demands.

Risk Aversion

In this section, I will assume the absence of transaction costs and consider the effects on the demands for cash and for liquid assets of risk aversion. The model is a modified form of that studied by Tobin.[5] It is basically a study of

3. The distinction between transaction costs which depend on the magnitude of the transaction and those which do not appears (in a nonstochastic context) in J. Tobin, "Interest-Elasticity of Transactions Demand for Cash," *Review of Economics and Statistics,* 38, 1956, pp. 241–247.

4. Although the policy had been discussed a good deal in the literature, the first proof of the optimality of the S, s policy under reasonably general conditions is due to my colleague H. Scarf, "The Optimality of the (S, s) Policies in the Dynamic Inventory Problem," in K. J. Arrow, S. Karlin, and P. Suppes, *Mathematical Methods in the Social Sciences, 1959,* Palo Alto, Stanford University Press, 1960, pp. 196–202.

5. J. Tobin, "Liquidity Preference as Behavior toward Risk," *Review of Economic Studies,* 26, 1958, pp. 65–86.

the choice between risky and safe assets; in interpretation it is perhaps more suited to analysis of the margin between liquid and nonliquid assets than between cash and liquid assets, since transaction motives are more significant in the latter choice.

We use the expected utility hypothesis of behavior under uncertainty; that is, the individual makes choices so as to maximize the expected value of a suitably chosen utility function for wealth. One implication of the expected utility hypothesis is not always understood: the utility function must be bounded, for otherwise a version of the St. Petersburg paradox could be found (this point was originally developed by the mathematician Karl Menger). If X is wealth and $U(X)$ is the utility function, then we suppose, of course, that $U(X)$ is strictly increasing; from the boundedness, it must approach a finite upper limit as X approaches infinity and a finite lower limit as X approaches zero. From the first of these, it is clear that on the average there must be risk aversion, i.e., $U''(X)$ (the second derivative) must be negative except, at most, for isolated intervals. It will be assumed here that individuals are risk averters throughout.

The quantity $U''(X)$ is not itself a suitable measure of risk aversion, since it depends on the units in which utility is measured. Two measures will be used here: (1) relative risk aversion, defined as $-XU''(X)/U'(X)$, which is also the elasticity of the marginal utility of income, and (2) absolute risk aversion, defined as $-U''(V)/U'(X)$. The first will be the more important.

The boundedness of the utility function has the following consequence: the relative risk aversion must approach a limit greater than 1 as X approaches infinity and a limit less than 1 as X approaches zero. (This may be made clearer by noting that the logarithmic utility function, which is unbounded at both ends, has a relative risk aversion constantly equal to 1.) If, for simplicity, we assume that the relative risk aversion is monotonic, then it must be monotonic increasing. In a sense, safety is a luxury good.

The choice model is that introduced by Tobin; the notation differs somewhat. Let A be the initial wealth of the individual. He can invest all or some with a random rate of return R; the remainder he leaves in cash with a certain rate of return of zero. The model can easily be extended to the case where there is a secure asset with a positive rate of return; in that case R is interpreted as the difference between the random and the secure rates. This interpretation applies to the choice between liquid and risky assets. Let a be the amount invested. Then the wealth at the end of the period is

$$a(1 + R) + (A - a) = A + aR,$$

and the individual seeks to maximize

$$E[U(A + aR)],$$

where a must lie between 0 and A. It can be shown that the optimal investment a is necessarily positive if and only if $E(R)$ is positive; an individual will always take some part of a favorable risk but, if a risk averter, will never take any part of an unfair risk. The optimum might involve investing the entire initial wealth. If it does not, the optimal investment satisfies the condition

$$E[U'(A + aR)R] = 0.$$

We are interested in the demand for risky investment, a, and its complement $m = A - a$, the demand for cash or liquid assets. First, we consider the effects of initial wealth, A. If absolute risk aversion is increasing, then it can be shown that a decreases as A increases, that is, that risky investment is an inferior good. Since this result is certainly empirically implausible, we must reject the hypothesis of increasing absolute risk aversion. It may be noted that the quadratic utility function, often used for its simplicity, implies increasing absolute risk aversion and so must be rejected. If, on the other hand, we assume decreasing absolute risk aversion, then risky investment becomes a normal good.

Decreasing relative risk aversion, which is a natural assumption as we have seen, has a very interesting implication: the wealth elasticity of the demand for cash or liquid assets is greater than 1, so that money and liquid assets are luxuries. Although the detailed empirical meaning of this implication requires further examination, particularly because of the neglect of transaction costs, it corresponds to the empirical work of Friedman and of Selden.[6]

These results have referred to wealth effects; one might also ask about price effects. Under uncertainty the analogue of a price is the probability distribution of R, and of course there are many possible ways a distribution can change. A simple upward shift by a constant amount (that is, the mean increases while the distribution about the mean remains unchanged) can be shown to increase the demand for risky investment, as might be expected, provided that risky investment is a normal good. It may be noted that an

6. M. Friedman, "The Demand for Money: Some Theoretical and Empirical Results," *Journal of Political Economy*, 67, 1959, pp. 327–351; R. T. Selden, "Monetary Velocity in the United States," in *Studies in the Quantity Theory of Money*, M. Friedman, ed., Chicago, University of Chicago Press, 1956, pp. 179–257.

increase in the rate of interest on the secure asset implies a downward shift in the distribution of the spread, R, and so, again as might be expected, an increase in the demand for liquid assets.

A second type of shift in the distribution of R is a simple multiplicative shift around zero. Here we have the simple and surprising result presented by Tobin; the risky investment is reduced in inverse proportion to the multiplier. Thus an ideal proportional income tax at rate t, which means a multiplier of $1 - t$, will increase investment in risky assets in the proportion $1/(1 - t)$.

Finally, we can consider a multiplicative shift around the mean instead of around zero, which might be thought of as a pure change in dispersion. This can be regarded as a compounding of the two previous shifts, a multiplication around zero followed by an additive shift to restore the mean to its original value. It follows that an increase in dispersion, in this sense, will reduce the demand for risky investment (and increase that for liquid assets), provided that risky investment is a normal good.

3 Uncertainty and the Welfare Economics of Medical Care

This paper is one that I cherish highly. It represented an attempt at understanding an issue to which standard economic theory was clearly only partially applicable, and it led to the beginnings of a new conceptualization. The paper originated in an idea of the Ford Foundation, then in a highly innovative phase, to take three social policy issues and have each surveyed by both a practical specialist and a general theorist. I was asked to be the theorist on medical economics. Then, as earlier in the 1930s and still today, the financing of medical care and in particular the role of insurance was a major political issue. I started the survey in a conscientious catalogue fashion, but felt that the whole study lacked focus. It was obvious enough that uncertainty was a key concept, as indeed it is whenever insurance is involved. It took me a while to see that differential information, between physician and patient and between both and an insurer, was a key component. Because of my obligation to cover the whole field, the special role of differential information does not show as clearly as it might.

This paper is an exploratory and tentative study of the specific differentia of medical care as the object of normative economics. It is contended here, on the basis of comparison of obvious characteristics of the

Reprinted from *American Economic Review*, 53 (1963):941–973. I wish to express thanks for useful comments to F. Bator, R. Dorfman, V. Fuchs, S. Gilson, R. Kessel, S. Mushkin, and C. R. Rorem. This paper was prepared under the sponsorship of the Ford Foundation as part of a series of papers on the economics of health, education, and welfare.

medical care industry with the norms of welfare economics, that the special economic problems of medical care can be explained as adaptations to the existence of uncertainty in the incidence of disease and in the efficacy of treatment.

It should be noted that the subject here is the *medical care industry,* not *health.* The causal factors in health are many, and the provision of medical care is only one. Particularly at low levels of income, other commodities such as nutrition, shelter, clothing, and sanitation may be much more significant. It is the complex of services that center about the physician, private and group practice, hospitals, and public health which I propose to discuss.

The focus of discussion will be on the way the operation of the medical care industry and the efficacy with which it satisfies the needs of society differ from a norm, if at all. The "norm" that the economist usually uses for the purposes of such comparisons is the operation of a competitive model, that is, the services that would be offered and purchased and the prices that would be paid for them if each individual in the market offered or purchased services at the going prices as if his decisions had no influence over them, and if the going prices were such that the amounts of services which were available equaled the total amounts which other individuals were willing to purchase with no imposed restrictions on supply or demand.

The interest in the competitive model stems partly from its presumed descriptive power and partly from its implications for economic efficiency. In particular, we can state the following well-known proposition (First Optimality Theorem). If a competitive equilibrium exists at all, and if all commodities relevant to costs or utilities are in fact priced in the market, then the equilibrium is necessarily *optimal* in the following precise sense (after V. Pareto): there is no other allocation of resources to services which will make all participants in the market better off.

The conditions of this optimality theorem and the definition of optimality both call for comment. A definition is just a definition, but when the *definiendum* is a word already in common use with highly favorable connotations, it is clear that we are really trying to be persuasive; we are implicitly recommending the achievement of optimal states.[1] It is reasonable enough to assert that a change in allocation which makes all participants better off is one that certainly should be made; this is a value judgment, not a descriptive

1. This point has been stressed by I. M. D. Little (1950, pp. 71–74). For the concept of a "persuasive definition," see C. L. Stevenson (1945, pp. 210–17).

proposition, but it is a very weak one. From this it follows that it is not desirable to put up with a nonoptimal allocation. But it does not follow that if we are at an allocation which is optimal in the Pareto sense, we should not change to any other. We cannot indeed make a change that does not hurt someone; but we can still desire to change to another allocation if the change makes enough participants better off and by so much that we feel that the injury to others is not enough to offset the benefits. Such interpersonal comparisons are, of course, value judgments. The change, however, by the previous argument ought to be to an optimal state; of course there are many possible states, each of which is optimal in the sense here used.

However, a value judgment on the desirability of each possible new distribution of benefits and costs corresponding to each possible reallocation of resources is not, in general, necessary. Judgments about the distribution can be made separately, in one sense, from those about allocation if certain conditions are fulfilled. Before stating the relevant proposition, it is necessary to remark that the competitive equilibrium achieved depends in good measure on the initial distribution of purchasing power, which consists of ownership of assets and skills that command a price on the market. A transfer of assets among individuals will, in general, change the final supplies of goods and services and the prices paid for them. Thus, a transfer of purchasing power from the well to the ill will increase the demand for medical services. This will manifest itself in the short run in an increase in the price of medical services and in the long run in an increase in the amount supplied.

With this in mind, the following statement can be made (Second Optimality Theorem): if there are no increasing returns in production, and if certain other minor conditions are satisfied, then every optimal state is a competitive equilibrium corresponding to some initial distribution of purchasing power. Operationally, the significance of this proposition is that if the conditions of the two optimality theorems are satisfied, and if the allocation mechanism in the real world satisfies the conditions for a competitive model, then social policy can confine itself to steps taken to alter the distribution of purchasing power. For any given distribution of purchasing power, the market will, under the assumptions made, achieve a competitive equilibrium which is necessarily optimal, and any optimal state is a competitive equilibrium corresponding to some distribution of purchasing power, so that any desired optimal state can be achieved.

The redistribution of purchasing power among individuals most simply takes the form of money: taxes and subsidies. The implications of such a

transfer for individual satisfactions are, in general, not known in advance. But we can assume that society can judge the distribution of satisfactions ex post facto and, if deemed unsatisfactory, take steps to correct it by subsequent transfers. Thus, by successive approximations, a most preferred social state can be achieved, with resource allocation being handled by the market and public policy confined to the redistribution of money income.[2]

If, on the contrary, the actual market differs significantly from the competitive model, or if the assumptions of the two optimality theorems are not fulfilled, the separation of allocative and distributional procedures becomes, in most cases, impossible.[3]

The first step then in the analysis of the medical care market is the comparison between the actual market and the competitive model. The methodology of this comparison has been a recurrent subject of controversy in economics for over a century. Friedman (1953) has vigorously argued that the competitive or any other model should be tested solely by its ability to predict. In the context of competition, he comes close to arguing that prices and quantities are the only relevant data. This point of view is valuable in stressing that a certain amount of lack of realism in the assumptions of a model is no argument against its value. But the price-quantity implications of the competitive model for pricing are not easy to derive without major—and, in many cases, impossible—econometric efforts.

In this chapter the institutional organization and the observable mores of the medical profession are included among the data to be used in assessing the competitiveness of the medical care market. I shall also examine the presence or absence of the preconditions for the equivalence of competitive equilibria and optimal states. The major competitive preconditions, in the sense used here, are three: the *existence* of competitive equilibrium, the *marketability* of all goods and services relevant to costs and utilities, and *nonincreasing returns*. The first two, as we have seen, ensure that competitive equilibrium is necessarily optimal; the third ensures that every optimal state is the competitive equilibrium corresponding to some distribution of income.[4] The first and third conditions are interrelated; indeed, nonincreasing returns plus some additional conditions not restrictive in a modern economy imply the existence of a competitive equilibrium, that is, imply that there will be some set of prices which will clear all markets.[5]

2. The separation between allocation and distribution even under the above assumptions has glossed over problems in the execution of any desired redistribution policy; in practice it is virtually impossible to find a set of taxes and subsidies that will not have an adverse effect on the achievement of an optimal state. But this discussion would take us even further afield than we have already gone.

3. The basic theorems of welfare economics alluded to so briefly above have been the subject of voluminous literature, but no thoroughly satisfactory statement covering both the theorems themselves and the significance of exceptions to them exists. The positive assertions of welfare economics and their relation to the theory of competitive equilibrium are admirably covered in Koopmans (1957). The best summary of the various ways in which the theorems can fail to hold is probably Bator's (1958).

The concept of marketability is somewhat broader than the traditional divergence between private and social costs and benefits. The latter concept refers to cases in which the organization of the market does not require an individual to pay for costs that he imposes on others as the result of his actions or does not permit him to receive compensation for benefits he confers. In the medical field, the obvious example is the spread of communicable diseases. An individual who fails to be immunized risks not only his own health, a disutility which presumably he has weighed against the utility of avoiding the procedure, but also that of others. In an ideal price system, there would be a price which he would have to pay to anyone whose health is endangered, a price sufficiently high so that the others would feel compensated; or, alternatively, there would be a price which would be paid to him by others to induce him to undergo the immunization procedure. Either system would lead to an optimal state, although the distributional implications would be different. It is, of course, not hard to see that such price systems could not, in fact, be practical; to approximate an optimal state it would be necessary to have collective intervention in the form of subsidy or tax or compulsion.

By the absence of marketability for an action which is identifiable, technologically possible, and capable of influencing some individual's welfare, for better or for worse, is meant here the failure of the existing market to provide a means whereby the services can be both offered and demanded upon payment of a price. Nonmarketability may be due to intrinsic technological characteristics of the product which prevent a suitable price from being enforced, as in the case of communicable diseases, or it may be due to social or historical controls, such as those prohibiting an individual from selling himself into slavery. This distinction is, in fact, difficult to make precise, although it is obviously of importance for policy; for the present

4. There are further minor conditions, for which see Koopmans (1957, pp. 50–55).

5. For a more precise statement of the existence conditions, see Koopmans (1957, pp. 56–60) or Debreu (1959, chap. 5).

purposes, it will be sufficient to identify nonmarketability with the observed absence of markets.

The instance of nonmarketability with which we shall be most concerned is that of risk bearing. The relevance of risk bearing to medical care seems obvious; illness is to a considerable extent an unpredictable phenomenon. The ability to shift the risks of illness to others is worth a price which many are willing to pay. Because of pooling and of superior willingness and ability, others are willing to bear the risks. Nevertheless, as we shall see in greater detail, a great many risks are not covered, and indeed the markets for the services of risk coverage are poorly developed or nonexistent. Why this should be so is explained in more detail below; briefly, it is impossible to draw up insurance policies which will sufficiently distinguish among risks, particularly since observation of the results will be incapable of distinguishing between avoidable and unavoidable risks, so that incentives to avoid losses are diluted.

The optimality theorems discussed above are usually presented in the literature as referring only to conditions of certainty, but there is no difficulty in extending them to the case of risks, provided the additional services of risk bearing are included with other commodities.[6]

However, the variety of possible risks in the world is really staggering. The relevant commodities include, in effect, bets on all possible occurrences in the world which impinge upon utilities. In fact, many of these "commodities," that is, desired protection against many risks, are simply not available. Thus, a wide class of commodities is nonmarketable, and a basic competitive precondition is not satisfied.[7]

There is a still more subtle consequence of the introduction of risk-bearing considerations. When there is uncertainty, information or knowledge becomes a commodity. Like other commodities, it has a cost of production and a cost of transmission, and so it is naturally not spread out over the entire population but concentrated among those who can profit most from it. (These costs may be measured in time or disutility as well as money.) But

6. The theory, in variant forms, seems to have been first worked out by Allais (1953), Arrow (1953), and Baudier (1959). For further generalization, see Debreu (1959, chap 7; 1960).

7. It should also be remarked that in the presence of uncertainty, indivisibilities that are sufficiently small to create little difficulty for the existence and viability of competitive equilibrium may nevertheless give rise to a considerable range of increasing returns because of the operation of the law of large numbers. Since most objects of insurance (lives, fire hazards, etc.) have some element of indivisibility, insurance companies have to be above a certain size. But it is not clear that this effect is sufficiently great to create serious obstacles to the existence and viability of competitive equilibrium in practice.

the demand for information is difficult to discuss in the rational terms usually employed. The value of information is frequently not known in any meaningful sense to the buyer; if, indeed, he knew enough to measure the value of information, he would know the information itself. But information, in the form of skilled care, is precisely what is being bought from most physicians and, indeed, from most professionals. The elusive character of information as a commodity suggests that it departs considerably from the usual marketability assumptions about commodities.[8]

That risk and uncertainty are, in fact, significant elements in medical care hardly needs argument. I will hold that virtually all the special features of this industry in fact stem from the prevalence of uncertainty.

The nonexistence of markets for the bearing of some risks in the first instance reduces welfare for those who wish to transfer those risks to others for a certain price, as well as for those who would find it profitable to take on the risk at such prices. But it also reduces the desire to render or consume services which have risky consequences; in technical language, these commodities are complementary to risk bearing. Conversely, the production and consumption of commodities and services with little risk attached act as substitutes for risk bearing and are encouraged by market failure there with respect to risk bearing. Thus, the observed commodity pattern will be affected by the nonexistence of other markets.

The failure of one or more of the competitive preconditions has as its most immediate and obvious consequence a reduction in welfare below that obtainable from existing resources and technology, in the sense of a failure to reach a Pareto-optimal state. But more can be said. I propose here the view that when the market fails to achieve an optimal state, society will, to some extent at least, recognize the gap, and nonmarket social institutions will arise attempting to bridge it.[9] Certainly this process is not necessarily conscious; nor is it uniformly successful in approaching more closely to optimality when the entire range of consequences is considered. It has

8. One form of production of information is research. Not only does the product have unconventional aspects as a commodity, but it is also subject to increasing returns in use, since new ideas, once developed, can be used over and over without being consumed, and to difficulties of market control, since the cost of reproduction is usually much less than that of production. Hence, it is not surprising that a free enterprise economy will tend to underinvest in research; see Nelson (1959) and Arrow (1962).

9. An important current situation in which normal market relations have had to be greatly modified in the presence of great risks is the production and procurement of modern weapons; see Peck and Scherer (1962), pp. 581–82). I am indebted for this reference to V. Fuchs and to Alchian, Arrow, and Capron (1958, pp. 71–75).

always been a favorite activity of economists to point out that actions which on their face achieve a desirable goal may have less obvious consequences, particularly over time, which more than offset the original gains.

But it is contended here that the special structural characteristics of the medical care market are largely attempts to overcome the lack of optimality that result from the nonmarketability of the bearing of suitable risks and the imperfect marketability of information. These compensatory institutional changes, with some reinforcement from usual profit motives, largely explain the observed noncompetitive behavior of the medical care market, behavior which, in itself, interferes with optimality. The social adjustment toward optimality thus puts obstacles in its own path.

The doctrine that society will seek to achieve optimality by nonmarket means if it cannot achieve them in the market is not novel. Certainly, the government, at least in its economic activities, is usually implicitly or explicitly held to function as the agency which substitutes for the market's failure.[10] I am arguing here that in some circumstances other social institutions will step into the optimality gap, and that the medical care industry, with its variety of special institutions, some ancient, some modern, exemplifies this tendency.

It may be useful to remark here that a good part of the preference for redistribution expressed in government taxation and expenditure policies and private charity can be reinterpreted as desire for insurance. It is noteworthy that virtually nowhere is there a system of subsidies that has as its aim simply an equalization of income. The subsidies or other governmental help go to those who are disadvantaged in life by events the incidence of which is popularly regarded as unpredictable: the blind, dependent children, the medically indigent. Thus, optimality, in a context which includes risk bearing, includes much that appears to be motivated by distributional value judgments when looked at in a narrower context.[11]

This methodological background gives rise to the following plan for this paper. I present a catalogue of stylized generalizations about the medical care market which differentiate it from the usual commodity markets, followed by a comparison of the behavior of the market with that of the competitive model which disregards the fact of uncertainty. Then, I compare the medical care market, both as to behavior and as to preconditions,

10. For an explicit statement of this view, see Baumol (1952). But I believe this position is implicit in most discussions of the functions of government.

11. Since writing the above, I find that Buchanan and Tullock (1962, chap. 13) have argued that all redistribution can be interpreted as "income insurance."

with the ideal competitive market that takes account of uncertainty, making an attempt to demonstrate that the characteristics outlined previously can be explained either as the result of deviations from the competitive preconditions or as attempts to compensate by other institutions for these failures. The discussion is not designed to be definitive, but provocative. In particular, I have been chary about drawing policy inferences; to a considerable extent, they depend on further research, for which the present paper is intended to provide a framework.

The Special Characteristics of the Medical Care Market

Some characteristics of medical care distinguish it from the usual commodity of economics textbooks.[12] This list is not exhaustive, and I do not claim that the characteristics listed are individually unique to this market. But, taken together, they do establish a special place for medical care in economic analysis.

The Nature of Demand

The most obvious distinguishing characteristic of an individual's demand for medical services is that it is not steady in origin like, for example, for food or clothing, but irregular and unpredictable. Medical services, apart from preventive services, afford satisfaction only in the event of illness, a departure from the normal state of affairs. It is hard, indeed, to think of another commodity of significance in the average budget of which this is true. A portion of legal services, devoted to defense in criminal trials or to lawsuits, might fall in this category, but the incidence is surely very much lower (and, of course, there are, in fact, strong institutional similarities between the legal and medical care markets).[13]

In addition, the demand for medical services is associated, with a considerable probability, with an assault on personal integrity. There is some risk of death and a more considerable risk of impairment of full functioning. In particular, there is a major potential for loss or reduction of earning ability. The risks are not by themselves unique; food is also a necessity, but avoidance of deprivation of food can be guaranteed with sufficient income,

12. For an illuminating survey to which I am much indebted, see Mushkin (1958).

13. In governmental demand, military power is an example of a service used only irregularly and unpredictably. Here too, special institutional and professional relations have emerged, although the precise social structure is different for reasons that are not hard to analyze.

whereas the same cannot be said of avoidance of illness. Illness is thus not only risky but a costly risk in itself, apart from the cost of medical care.

Expected Behavior of the Physician

It is clear from everyday observation that the behavior expected of sellers of medical care is different from that of business people in general. These expectations are relevant because medical care belongs to the category of commodities for which the product and the activity of production are identical. In all such cases, the customer cannot test the product before consuming it, and there is an element of trust in the relationship.[14] But the ethically understood restrictions on the activities of a physician are much more severe than on those of, say, a barber. His behavior is supposed to be governed by a concern for the customer's welfare which would not be expected of a salesman. In Talcott Parsons's terms, there is a "collectivity orientation" which distinguishes medicine and other professions from business, where self-interest on the part of participants is the accepted norm.[15]

A few illustrations will indicate the degree of difference between the behavior expected of physicians and that expected of the typical businessperson.[16] Advertising and overt price competition are virtually eliminated among physicians. Advice given by physicians as to further treatment by himself or others is supposed to be completely divorced from self-interest. It is at least claimed that treatment is dictated by the objective needs of the case and not limited by financial considerations.[17] While the ethical compulsion is surely not absolute in fact as it is in theory, we can hardly suppose that it

14. Even with material commodities, testing is never so adequate that all elements of implicit trust can be eliminated. Of course, over the long run, experience with the quality of product of a given seller provides a check on the possibility of trust.

15. See Parsons (1951, p. 463). The whole of chap. 10 of this work is a most illuminating analysis of the social role of medical practice; even though Parsons's interest lies in different areas from mine, I must acknowledge here my indebtedness to his work.

16. I am indebted to Herbert Klarman of Johns Hopkins University for some of the points discussed in this and the following paragraph.

17. The belief that the ethics of medicine demands treatment independent of the patient's ability to pay is strongly ingrained. Such a perceptive observer as Rene Dubos has made the remark that the high cost of anticoagulants restricts their use and may contradict classical medical ethics, as though this were an unprecedented phenomenon. See Dubos (1959, p. 419): "A time *may come* when medical ethics will have to be considered in the harsh light of economics" (emphasis added). Of course, this expectation amounts to ignoring the scarcity of medical resources; one has only to have been poor to realize the error. We may confidently assume that price and income do have some consequences for medical expenditures.

has no influence over resource allocation in this area. Charity treatment in one form or another does exist because of this tradition about human rights to adequate medical care.[18] The physician is relied upon as an expert in certifying the existence of illnesses and injuries for various legal and other purposes. It is socially expected that his concern for the correct conveying of information will, when appropriate, outweigh his desire to please his customers.[19]

Departure from the profit motive is strikingly manifested by the overwhelming predominance of nonprofit over proprietary hospitals.[20] The hospital per se offers services not too different from those of a hotel, and it is certainly not obvious that the profit motive will not lead to a more efficient supply. The explanation may lie either on the supply side or on that of demand. The simplest explanation is that public and private subsidies decrease the cost to the patient in nonprofit hospitals. A second possibility is that the association of profit making with the supply of medical services arouses suspicion and antagonism on the part of patients and referring physicians, so they prefer nonprofit institutions. Either explanation implies a preference on the part of some group, whether donors or patients, against the profit motive in the supply of hospital services.[21]

Conformity to collectivity-oriented behavior is especially important since it is a commonplace that the physician-patient relationship affects the quality of the medical care product. A pure cash nexus would be inadequate; if nothing else, the patient expects that the same physician will normally treat him on successive occasions. This expectation is strong enough to persist even in the Soviet Union, where medical care is nominally removed from the market place (1957, pp. 194–196). That purely psychic interac-

18. A needed piece of research is a study of the exact nature of the variations of medical care received and medical care paid for as income rises. (The relevant income concept also needs study.) For this purpose, some disaggregation is needed; differences in hospital care which are essentially matters of comfort should, in the above view, be much more responsive to income than, for example, drugs.

19. This role is enhanced in a socialist society, where the state itself is actively concerned with illness in relation to work; see Field (1957, chap. 9).

20. About 3 percent of beds were in proprietary hospitals in 1958, against 30 percent in voluntary nonprofit and the remainder in federal, state, and local hospitals. See Somers and Somers (1961, chart 4-2, p. 60).

21. C. R. Rorem has pointed out to me some further factors in this analysis. (1) Given the social intention of helping all patients without regard to immediate ability to pay, economies of scale would dictate a predominance of community-sponsored hospitals. (2) Some proprietary hospitals will tend to control total costs to the patient more closely, including the fees of physicians, who will therefore tend to prefer community-sponsored hospitals.

tions between physician and patient have effects which are objectively indistinguishable in kind from the effects of medication is evidenced by the use of the placebo as a control in medical experimentation; see Shapiro (1960).

Product Uncertainty

Uncertainty as to the quality of the product is perhaps more intense here than in any other important commodity. Recovery from disease is as unpredictable as its incidence. With most commodities the possibility of learning from one's own experience or that of others is strong because there is an adequate number of trials. In the case of severe illness that is generally not true; the uncertainty resulting from inexperience is added to the intrinsic difficulty of prediction. Further, the amount of uncertainty, measured in terms of utility variability, is certainly much greater for medical care in severe cases than for, say, houses or automobiles, even though these are also expenditures sufficiently infrequent that there may be considerable residual uncertainty.

Further, there is a special quality to the uncertainty; it is very different on the two sides of the transaction. Because medical knowledge is so complicated, the information possessed by the physician as to the consequences and possibilities of treatment is necessarily very much greater than that of the patient, or at least so it is believed by both parties.[22] Further, both parties are aware of this informational inequality, and their relationship is colored by this knowledge.

To avoid misunderstanding, observe that the difference in information relevant here is a difference in information as to the consequence of a purchase of medical care. There is always an inequality of information as to production methods between the producer and the purchaser of any commodity, but in most cases the customer may well have as good or nearly as good an understanding of the utility of the product as the producer.

Supply Conditions

In competitive theory the supply of a commodity is governed by the net return from its production compared with the return derivable from the use

22. Without trying to assess the present situation, it is clear in retrospect that at some point in the past the actual differential knowledge possessed by physicians may not have been much. But from the economic point of view it is subjective belief of both parties, as manifested in their market behavior, that is relevant.

of the same resources elsewhere. There are several significant departures from this theory in the case of medical care.

Most obviously, entry to the profession is restricted by licensing. Licensing, of course, restricts supply and therefore increases the cost of medical care. It is defended as guaranteeing a minimum of quality. Restriction of entry by licensing occurs in most professions, including barbering and undertaking.

A second feature is perhaps even more remarkable. The cost of medical education today is high and, according to the usual figures, is borne only to a minor extent by the student. Thus, the private benefits to the entering student considerably exceed the costs. (It is, however, possible that research costs, not properly chargeable to education, swell the apparent difference.) This subsidy should, in principle, cause a fall in the price of medical services, but it is offset by rationing through limited entry to schools and through elimination of students during the medical school career. These restrictions basically render superfluous the licensing, except in regard to graduates of foreign schools.

The special role of educational institutions in simultaneously subsidizing and rationing entry is common to all professions requiring advanced training.[23] It is a striking and insufficiently remarked phenomenon that such an important part of resource allocation should be performed by non-profit-oriented agencies.

Since this last phenomenon goes well beyond the purely medical aspect, I will not dwell on it longer here except to note that the anomaly is most striking in the medical field. Educational costs tend to be far higher there than in any other branch of professional training. While tuition is the same, or only slightly higher, so that the subsidy is much greater, at the same time the earnings of physicians rank highest among professional groups, so there would not at first blush seem to be any necessity for special inducements to enter the profession. Even if we grant that, for reasons unexamined here, there is a social interest in subsidized professional education, it is not clear why the rate of subsidization should differ among professions. One might expect that the tuition of medical students would be higher than that of other students.

The high cost of medical education in the United States is itself a reflection of the quality standards imposed by the American Medical Association since the Flexner Report, and it is, I believe, only since then that the

23. The degree of subsidy in different branches of professional education is worthy of a major research effort.