

21 Most Influential Economic Papers Of All Time



The word 'economics' comes from two Greek words, 'eco' meaning home and 'nomos' meaning accounts – which is a social science concerned with the production, distribution, and consumption of goods and services. It studies how economic agents and economies make choices on allocating resources to satisfy their wants and needs, trying to determine how these groups should organize and coordinate efforts to achieve maximum output.

Contents

1. A Theory of Production
2. The Use of Knowledge In Society
3. Economic Growth and Income Inequality
4. The Cost of Capital, Corporation Finance and the Theory of Investment
5. A Theory of Optimum Currency Areas
6. Uncertainty and the Welfare Economics of Medical Care

7. National Debt In A Neoclassical Growth Model
8. The Role of Monetary Policy
9. Migration, Unemployment and Development
10. Optimal Taxation and Public Production I: Production Efficiency
11. Optimal Taxation and Public Production II: Tax Rules
12. Production, Information Costs, and Economic Organization
13. The Economic Theory of Agency: The Principal's Problem
14. Some International Evidence on Output-Inflation Tradeoffs
15. The Political Economy of the Rent-Seeking Society
16. Monopolistic Competition and Optimum Product Diversity
17. An Almost Ideal Demand System
18. On the Impossibility of Informationally Efficient Markets
19. Scale Economies, Product Differentiation, and the Pattern of Trade
20. Do Stock Prices Move Too Much to be justified by Subsequent Changes in Dividends?
21. Capital Theory and Investment Behavior

A THEORY OF PRODUCTION¹

By

CHARLES W. COBB, *Amherst College*

PAUL H. DOUGLAS, *University of Chicago*

1. *Introduction.*—The progressive refinement during the recent years in the measurement of the volume of physical production in manufacturing suggests the possibility of attempting (1) to measure the changes in the amount of labor and capital which have been used to turn out this volume of goods, and (2) to determine what relationships existed between the three factors of labor, capital, and product.

If the relative supply from year to year of labor and capital were thus even approximately ascertained, a number of further problems would inevitably present themselves for solution of which the following are typical. (1) Can we estimate, within limits, whether this increase in production was purely fortuitous, whether it was primarily caused by technique, and the degree, if any, to which it responded to changes in the quantity of labor and capital?

(2) May it be possible to determine, again within limits, the *relative* influence upon production of labor as compared with capital?

(3) As the proportions of labor to capital changed from year to year, may it be possible to deduce the relative amount added to the total physical product by each unit of labor and capital and what is more important still by the *final* units of labor and capital in these respective years?

Might at least an historical approach to the theories of decreasing imputed productivity (diminishing increment to total product) be afforded and the way opened for further attempts to secure quantitative approximations to these assumed tendencies, if indeed there should turn out to be historical validity to them?

(4) Can we measure the probable slopes of the curves of incremental product which are imputed to labor and to capital and thus give greater definiteness to what is at present purely an hypothesis with no quantitative values attached?

(5) Finally from such a study of (a) the imputed physical product from year to year of a unit of labor and capital when joined with (b) a study of the relative exchange value of a physical unit of manufactured goods in these years and compared with (c) the actual movement of "real" wages in manufacturing and of real interest (if the latter can be ascertained), may we secure light upon the question as

¹ Mr. Douglas has been responsible for sections 1-5 and 8-10, of this paper and Mr. Cobb for sections 6 and 7.

to whether or not the processes of distribution are modeled at all closely upon those of the production of values?

The paper which follows attempts to deal with these questions and to throw some light upon them. But before this is done, it is of course necessary to construct indexes of the relative amounts of labor and capital which have been used and it is this which is dealt with in the two succeeding sections, leaving the later sections for the treatment of the interrelationships which may be discovered.

2. The Growth of Fixed Capital in Manufacturing in the United States, 1889-1922.—The census of manufactures has periodically included a question on the amount of capital invested in the various manufacturing enterprises and has tabulated the returns. This, however, includes in addition to fixed capital in the form of machinery and buildings, working capital including raw materials, goods in process of manufacture and finished goods in warehouses. It also includes land. Since we are attempting to measure the capital which aids in the production of goods, we should exclude working capital, for this is the result and not a cause of the process of manufacture.² We should also exclude land values since these are largely composed of the unearned increment. We shall therefore attempt to measure the changes in the physical quantity of (1) machinery, tools, and equipment and (2) factory buildings.

Unfortunately while statistics of total capital are given virtually every census year, they were only segregated for these specific groups in 1889, 1899, and 1904.³ The Census Bureau in its 1922 report on *Wealth, Public Debt and Taxation* estimated that manufacturing machinery, tools, and equipment formed 30 per cent of the total amount of manufacturing capital.⁴ Since it set the latter at 52,610 millions, this would give a figure for machinery, etc., of 15,783 million.

Year	Value of Factory Buildings (in millions of dollars)	Percentage of Manufacturing Capital	Value of Machinery, Implements and Equipment (in millions of dollars)	Percentage of Total Manufacturing Capital
1889.....	879	13.4	1,584	24.3
1899.....	1,450	14.8	2,543	25.9
1904.....	1,996	15.8	3,490	27.5
1922.....			15,783	30.0 ⁵

² Working capital of course normally "produces" value for its owner but we are here not concerned with value but with physical production.

³ See *13th Census* (1900), VI, xcvi, and the *Census of Manufactures*, 1904, Part I, pp. lxiv-lxv.

⁴ Bureau of the Census, *Estimated National Wealth* (1925), pp. 9-10.

⁵ Estimate of the Census Bureau.

The amounts which have thus been ascribed to each of these forms of capital and the percentages which they formed of total capital for the given years were as shown on page 140.

These statistics furnish a basis for estimating the probable value of these forms of manufacturing capital in those years when no such segregation of items was carried out. Not only was the total amount of capital increasing but fixed capital was coming to form a larger percentage of this greater sum.

It seems undeniable that buildings and machinery did not increase as rapidly in comparison with working capital during the eighties as they did during the fifteen years which followed 1889 when buildings advanced from 13.4 to 15.8 per cent, or an increase of 2.4 points, and machinery, etc., from 24.3 to 27.5, or a gain of 3.2 points. This was an advance of .16 and .21 points a year, respectively. We have assumed the growth in the proportions which buildings formed of the total was at approximately only one-quarter of the rate of speed of the nineties and for machinery at only one-fifth. This would give 13.0 per cent as the probable figure for buildings in 1879 and 24.0 per cent as that for machinery, tools, and equipment.

If we accept the census estimate of 30 per cent as the proportion which machinery formed of the total in the terminal year of 1922, we may then distribute the 2.5 per cent increase from 27.5 per cent in 1904, according to a fairly even ratio. The rate of growth from 1914 on was, however, undoubtedly somewhat more rapid than during the previous decade and allowance should be made for this fact.

The growth in the relative importance of buildings since 1904 is more problematical since we have no end value on which to build. While the absolute increases have been enormous, it has not seemed to us that the relative importance of buildings in comparison with other forms of capital has advanced at the same rate as during the years 1889-1904. Because of this and the results of a Missouri investigation, we have estimated the percentage at 16.5 for 1922 and have distributed this over the preceding years but providing for a more rapid growth after 1914 than before. Table I gives the estimated percentage of each of these forms of manufacturing capital in the various years and the amounts in terms of dollars.

There is some evidence to indicate that the estimated total for buildings and machinery at 46.5 per cent is not far from correct. Thus the Missouri State Bureau of Labor Statistics shows that in 1923, \$334.7 millions were invested in that state in manufacturing buildings, machinery, etc., and \$58.7 millions in "grounds."⁶ The amount of the working capital is not given but this was set for the country as a

* *Forty-fourth Annual Report Missouri Bureau of Labor* (1923), p. 155.

whole by the Federal Trade Commission at 45.7 per cent of the total capital.⁷ Since this is based upon the returns of 54,862 corporations with a total capital of 33.65 billions, it may be accepted as the best nation-wide estimate which we have. If we apply this ratio to Missouri, we would get 331.1 millions or a total with other items of 724.6 millions. Now buildings, machinery, and equipment were, as stated, evaluated independently by the Missouri study at 334.7 millions and this would be 46.2 per cent of the total. This is in almost exact agreement with the estimate of 46.5 per cent which we have made for these forms of capital in 1922. Since the types of manufacturing in Missouri are not unrepresentative⁸ of conditions in the country as a whole, our estimate can be considered to be buttressed and until better statistics are collected to be probably the best which can be made.

TABLE I

ESTIMATED VALUES OF MANUFACTURING BUILDINGS AND MACHINERY, TOOLS AND EQUIPMENT AND PERCENTAGES WHICH THEY FORMED OF TOTAL MANUFACTURING CAPITAL, 1879-1922

Year	Percentage of Total Manufacturing Capital		Value in Millions of Dollars		
	Buildings	Machinery and Equipment	Buildings	Machinery and Equipment	Total
1879.....	13.0	24.0	363	670	1,033
1889.....	13.4	24.3	879	1,584	2,463
1899.....	14.8	25.9	1,450	2,543	3,993
1904.....	15.8	27.5	1,996	3,490	5,486
1909.....	16.0	28.1	2,948	5,178	8,126
1914.....	16.2	28.7	3,692	6,541	10,233
1919.....	16.4	29.5	7,293	13,118	20,411
1922.....	16.5	30.0	8,681	15,783	24,464

There remains however the natural query as to what these census returns mean and how much the original data are worth. In recent years, the Census Bureau has instructed its agents to see that these statistics be taken "at the amounts carried on the books." Does this book value then mean the original cost of the buildings, machinery, etc., or the cost of reproduction? Mr. La Verne Beals, the chief statistician for manufactures, who is probably the ablest expert in this general field, has stated⁹ that the "manufacturers have as a rule reported capital on the basis of original cost rather than cost of reproduction."

⁷ Federal Trade Commission, *National Wealth and Income*, p. 135. (Senate Doc. 126, 69th Congress, 1st Session.)

⁸ Thus while Missouri does not have any textile industries and but a small clothing industry, it does have a considerable amount of capital invested in printing, foundries, automobile manufacture, meat packing, smelting, and brick and lime works. There is also a fast growing shoe industry.

⁹ Letter to author, October 28, 1925.

It is true that the Census Bureau has frequently issued cautions against accepting too implicitly its total for manufacturing capital and has indeed omitted such a question from its schedule for the 1921, 1923, and 1925 censuses. But, if the difficulties created by the fact that the investments are computed in terms of the price levels of the different years in which they were originally made can be overcome, and if the capital index can then be reduced to dollars of constant purchasing power, there would then seem to be no good reason why the resulting data should not be taken as a fairly accurate index of the *relative growth* of fixed capital, if not of its absolute amount. Moreover the proper correction of the distortions produced by changing price levels would remove most of the objections which can be leveled against such figures as a measurement of the total amount of capital. There remain two further problems before we can construct a continuous and comparable index: (1) finding the probable increments in each of the intervening years and (2) reducing these various increments of savings in terms of the value of a common price level.

Since the statistics are based upon original cost, the first problem consists in finding the annual increments of capital in terms of the prices of that year and of adding these to the values of the preceding year. The method followed was, in brief: (1) To ascertain the quantities of the following commodities produced in each year from 1899-1922: pig iron, rolled and forged steel, lumber, coke, cement, bricks and copper.¹⁰ It will be noted that these commodities are the most important of those which are used in the construction of machinery and of buildings. In those few cases where it was impossible to secure actual figures of production for a given year, these were estimated from other years on the basis of the relative change in Professor Day's index of physical production for that group of manufactured products in which the commodity in question was included.¹¹ For the period 1880-1889, the quantities of pig iron, steel, cement, copper and coke were used. (2) The quantity produced of each commodity in each year was multiplied by its current price per unit of product.¹² The prices for the period from 1890-1922 were those collected and published by the United States Bureau of Labor Statistics¹³ while those used for the decade from 1880-1890 were those published in the reports of the

¹⁰ The raw data were secured from the United States Statistical Abstract for the various years. Also *Mineral Resources of the United States* 1921 Part I, pp. 235-82; 565-98; Part II, pp. 371-440.

¹¹ E. E. Day, "An Index of the Physical Volume of Production," *Review of Economic Statistics*, Vol. II (1920) pp. 328-29. Day, "The Physical Volume of Production in the United States for 1923." *Ibid.* Vol. VI (1924) p. 201.

¹² The average of the prices of spruce and maple was used for lumber.

¹³ Bulletin 335 of the United States Bureau of Labor Statistics, *Wholesale Prices*, 1890-1922, pp. 126-56.

Aldrich Committee.¹⁴ In some cases, it was possible directly to derive the value of the total product without multiplying the physical product by the price per unit and wherever this was the case the directly quoted total was used. (3) The values of each commodity produced in a given year were then added together to obtain the total values of these producers goods turned out in each year. (4) The values of these capital goods which were produced between two census years were then totaled (e.g., 1880 to 1889 inclusive) and the value for each year was divided by the total for the period in order to get the percentage which it formed of the total value produced during the period as a whole. These percentages were then applied to the total increase in the value of buildings and machinery over the same period and estimated yearly increases in the value of these items were thus obtained.

This process may be illustrated by the following example. The increase in the value of buildings and machinery between 1879 and 1889 was 1430 millions. The total money values in each of the years of these capital goods and the per cent which each of these yearly totals formed of the total for the period were as follows:

Year	Value of Specified Capital Goods (in millions of dollars)	Percent of Total Value for Decade
1880.....	200	9.6
1881.....	210	10.0
1882.....	216	10.3
1883.....	184	8.8
1884.....	148	7.1
1885.....	141	6.7
1886.....	211	10.0
1887.....	282	13.5
1888.....	241	11.5
1889.....	263	12.5
Total.....	2,096	100.0

The increase in the value of buildings and machinery during the decade, 1430 millions, was then multiplied by each of these percentages and the probable amounts of the yearly increases in value were obtained. These amounts when totaled and added to the total for 1879 would of necessity equal the 1889 value. The basic assumption is of course that the capital values in terms of original cost grew from year to year as the money value of the capital goods produced.

But since these estimated additions to capital are reckoned in terms of the dollars of the given years, if we are to secure an index of rela-

¹⁴ Report of Senate Committee on Whole Prices, on Wages and on Transportation, Appendix A. The criticisms of the index of prices do not apply here since the absolute prices quoted were used.

tive real capital, it is necessary to eliminate the effect of changing price levels. A capital cost index was accordingly computed which was based on three sets of relative prices: (1) the wholesale prices of metals and metal products, (2) the wholesale prices of building materials and (3) money wages. The Aldrich Committee report was used to obtain prices for the first two groups of products from 1880 to 1889¹⁵ while the indexes of the Bureau of Labor Statistics were used for the years 1890 to 1922.¹⁶ For wages, the index previously computed by one of the authors was used for the period from 1890 on¹⁷ while the average wages computed by Dr. R. P. Falkner for the Aldrich report were taken to show the movement during the eighties. These three series were then reduced to relatives with 1880 serving as 100 and were

TABLE II

ESTIMATED ANNUAL ADDITIONS TO FIXED CAPITAL IN MANUFACTURING TOGETHER WITH
CUMULATIVE TOTAL CAPITAL AS EXPRESSED IN TERMS OF COST AND 1880 PRICES
(Millions of dollars), 1899-1922

Year	Annual Increase in Terms of Cost Price (1)	Cost Index (1880=100) (2)	Annual Increase in Terms of 1880 dollars (3)	Total Fixed Capital in 1880 dollars (4)	Relative Total Capital 1899=100 (5)
1899.....	339	88	387	4449	100
1900.....	264	89	297	4746	107
1901.....	277	88	315	5061	114
1902.....	342	89	383	5444	122
1903.....	328	91	362	5806	131
1904.....	282	87	326	6132	138
1905.....	457	92	494	6626	149
1906.....	612	100	611	7237	163
1907.....	629	106	595	7832	176
1908.....	373	94	397	8229	185
1909.....	569	96	591	8820	198
1910.....	422	100	420	9240	208
1911.....	379	99	384	9624	216
1912.....	457	103	443	10067	226
1913.....	497	110	453	10520	236
1914.....	356	101	353	10873	244
1915.....	1017	105	967	11840	266
1916.....	1899	135	1402	13242	298
1917.....	2891	173	1673	14915	335
1918.....	2473	183	1350	16265	366
1919.....	1898	196	969	17234	387
1920.....	2096	237	884	18118	407
1921.....	780	184	424	18542	417
1922.....	1177	181	650	19192	431

¹⁵ Report of Senate Committee on Wholesale Prices, etc., pp. 92-99. The celebrated twenty-five varieties of jack-knives were subtracted from the metal index before using it.

¹⁶ Bulletin 335, Wholesale Prices, 1890-1922, pp. 8-9.

¹⁷ Paul H. Douglas, "The Recent Movement of Real Wages and Its Economic Significance," Supplement, *American Economic Review*, March, 1926, p. 30.

combined into a weighted average. The weights used were metals and metal products, 4; building materials, 2; and wages, 3.

Each yearly increase in the value of manufacturing buildings and machinery was then divided by the relative cost index for that year and a series of "deflated" increases were thus obtained, or rather a series of increases which were expressed in terms of the 1880 price level for capital goods. The next and final step was to add these deflated yearly increases to the estimated total for buildings and machinery for 1879 and thereafter to the total for each preceding year. Table 2 shows all this material. Since our other data only extend from 1899-1922, the years prior to 1899 are omitted from this table. Values given are in millions of dollars.

The index is defective in that it does not allow for the replacement of original capital at differing price levels. The census statistics of book value, undoubtedly include replacements made at different and generally higher prices than those which prevailed when the original capital was invested. Consequently the advance from year to year is not solely the result of the saving of additional increments of capital but includes in part the replacement at other price levels of the old capital as it wore out. The consequence is that our index is throughout most of its course somewhat higher than it should be. It is hoped to publish a revision of this index in the not distant future in which this error will be eliminated. In the meantime this is offered as a first approximation.

The index does not of course measure the short-time fluctuations in the amount of capital used. Thus, no allowance is made for the capital which is allowed to be idle during periods of business depression nor for the greater than normal intensity of use in the form of second shifts, etc., which characterizes the periods of prosperity.

The validity of this index of growth is somewhat strengthened, however, when we compare the increase in terms of book value which we have estimated for the United States¹⁸ during the years 1910-1920 with

Year	Massachusetts (Total Capital)	Estimated for United States (Fixed Capital)
1911.....	105	104
1912.....	110	110
1913.....	113	116
1914.....	130	120
1915.....	130	132
1916.....	150	154
1917.....	188	188
1918.....	210	217
1919.....	248	239
1920.....	250	263

¹⁸ This column was omitted from Table II because of lack of space.

the growth of total capital in Massachusetts when computed upon a similar basis.¹⁹ Using 1910 as a base, the relative increases were as shown on page 146.

The coincidence between these two indexes is very striking and this becomes even more the case when we remember that most of the greater increase shown for the United States as a whole was due to the fact that the fixed capital was increasing at a more rapid rate than was the supply of total capital in manufacturing.

It may be remarked that this index shows a truly unprecedented growth in the volume of fixed capital. Thus the amount virtually doubled during the decade from 1899-1909. This was a compounded average yearly rate of increase of 7 per cent. This same rate of increase was virtually maintained during the succeeding decade. From 1919 on the rate of growth slackened during the three succeeding years but while we have not computed the growth since 1922 it has beyond question increased greatly since then. Taken as a whole this period showed an approximate doubling in the quantity during every decade, which would probably be scaled down to about 6 per cent per year compounded if deductions were made for the increased cost of replacing the old capital. This is a rate of growth which it is believed has not been matched by any other country.²⁰ It will be remembered that Cassel estimates the rate of growth of capital in Western Europe at 3 per cent a year. If this is true, the rate of industrial capital growth in the United States has been twice as great while if the growth be reckoned on a per capita basis, the disparity is even greater.

3. *The Growth in the Labor Supply, 1899-1922.*—The various censuses of manufactures give the average number of wage-earners employed in each of the census years.²¹ Using these as the bases, we can find the probable numbers employed in the intercensus years by using an index of relative employment. This index was constructed for the years 1899-1904 by combining statistics of the relative number employed from year to year in Massachusetts²² and Pennsylvania.²³ From 1904 to 1914, figures for New Jersey²⁴ were substituted for those of Pennsylvania. In both periods, the relative index for each state was then weighted by the number shown by the census to be employed in that state at the beginning of the period and a combined index was

¹⁹ See Annual reports of Massachusetts Bureau of Statistics, *Statistics of Manufactures*, 1910-20.

²⁰ Our index shows a more than doubling from 1879 to 1899 and an increase of approximately 90 per cent during the nineties.

²¹ I.e. namely 1899, 1904, 1909, 1914, 1919, and 1921.

²² See Annual reports on *Statistics of Manufactures*, Massachusetts, 1900-1905.

²³ See Reports Pennsylvania State Department of Internal Affairs.

²⁴ Annual volume of New Jersey Bureau of Labor and Industries, *Statistics of Manufactures* (1904-1914).

thus secured. The assumption was then made that the volume of employment of the country as a whole followed a similar course to that in these two states. When the rate of change in these two states differed over a census period from the country-wide figures, then it was assumed that this greater or less degree of change had been distributed evenly over the intervening years, and the percentage changes for the two states were scaled down or up to conform to this standard.²⁵ Thus the increase in the number employed in 1904 over 1899 was as shown by the Census 1,066,000, or 21 per cent. If the increase shown for Massachusetts and New Jersey was 24 per cent, then it was assumed that the differences between the rate of growth for the country and for the two states increased annually at the rate of one-fifth of 3 per cent or .6 per cent. Then if the increase shown in Massachusetts and New Jersey for 1900 over 1899 was 4.6 per cent, this was scaled down to 4.0 per cent. Similar methods were used for the subsequent years.

From 1914 to 1919 the index was secured by combining that of the Bureau of Labor Statistics²⁶ for a number of industries with that for New York. In doing this, the Bureau's index was given a weight of 3 and that of New York a weight of 1.²⁷ From 1919 on, the index of the Federal Reserve Board was used which in turn was largely based upon the index of the Bureau of Labor Statistics. A substantially similar method was used to find the probable number employed in each

TABLE III
THE PROBABLE AVERAGE NUMBER OF WAGE-EARNERS EMPLOYED IN MANUFACTURING
1899-1922

Year	Average Number Employed (in thousands)	Relative Number (1899 = 100)	Year	Average Number Employed (in thousands)	Relative Number (1899 = 100)
1899.....	4713	100	1911	6855	145
1900.....	4968	105	1912	7167	152
1901.....	5184	110	1913	7277	154
1902.....	5554	118	1914	7026	149
1903.....	5784	123	1915	7269	154
1904.....	5468	116	1916	8601	182
1905.....	5906	125	1917	9218	196
1906.....	6251	133	1918	9446	200
1907.....	6483	138	1919	9096	193
1908.....	5714	121	1920	9110	193
1909.....	6615	140	1921	6947	147
1910.....	6807	144	1922	7602	161

* This is the identical method which I have followed in interpolating average annual earnings in the intercensal years from the statistics of earnings of the various states.

^a See files of *Monthly Labor Review*.

^b See *New York Labor Market Bulletin*.

of the intercensal years up to and including 1922.²⁸ Table III gives these estimated numbers from 1899 on and also expresses them in terms of relatives.

This index is defective in a number of respects as a perfect measure of the working force. (1) It does not include clerical employees who have been increasing in number at approximately double the rate of the wage-earners. (2) It is based on man-years rather than "standard" man hours. The average number of hours constituting the standard week's work has declined during this period, so that an increase in the number of men would be necessary merely to offset this decrease. One of the authors has computed a tentative index of standard man hours by multiplying the number of workers in each year by the average number of hours in the "normal" week. There is reason to believe, however, that this index is not yet perfected and so man-years have been used instead. It is hoped to include total "standard" hours in later studies. (3) It does not measure deviations from this standard week whether they take the form of short-time in periods of depression or overtime in the years of prosperity.

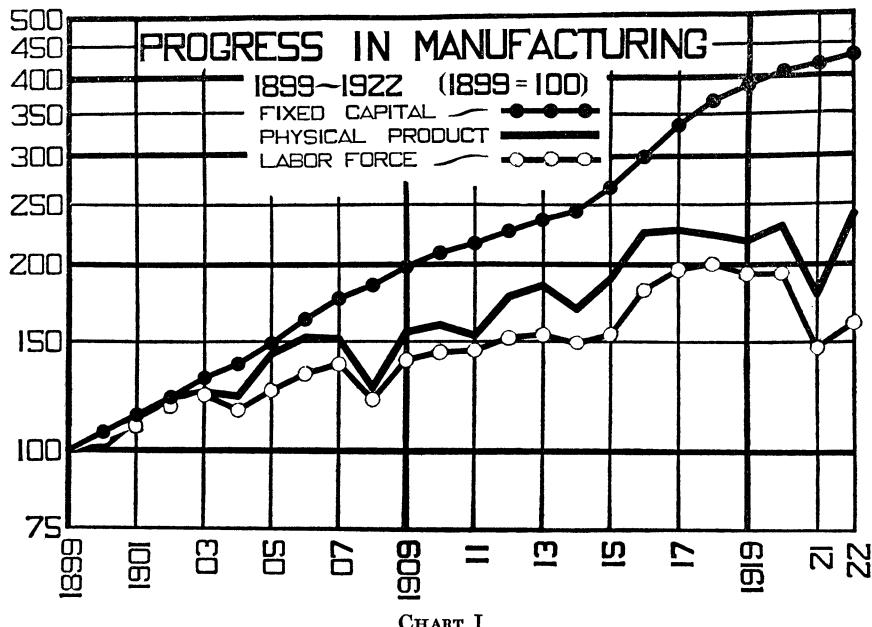
Such an index of course makes no allowance for possible changes in the quality of the laborers or in the intensity of their work. These factors may be of considerable importance but at present they certainly cannot be measured quantitatively and until they can be, it is better for any statistical study to ignore them than to make necessarily fantastic estimates as to their importance. When they can be measured, then they should be included.

4. *The Growth of Physical Production, 1899-1922.*—For this, we have used E. E. Day's well-known index of the physical volume of

TABLE IV
INDEX OF PHYSICAL VOLUME OF MANUFACTURES IN THE UNITED STATES

Year	Index of Manufactures	Year	Index of Manufactures
1899.....	100	1911.....	158
1900.....	101	1912.....	177
1901.....	112	1913.....	184
1902.....	122	1914.....	169
1903.....	124	1915.....	189
1904.....	122	1916.....	225
1905.....	143	1917.....	227
1906.....	152	1918.....	223
1907.....	151	1919.....	218
1908.....	126	1920.....	231
1909.....	155	1921.....	179
1910.....	159	1922.....	240

* Since these statistics of employment did not begin until July, 1914, the yearly average was secured by projecting them back for the preceding six months according to the monthly fluctuations in employment shown by the 1914 census of manufactures.



production for the years 1899-1922, since at the time we were carrying through our studies the later index given by Dr. Thomas was not available.²⁹

Chart I shows on a logarithmic scale the relative growth in manufacturing during this period of fixed capital, of the labor force, and of the

TABLE V
THE RELATIVE PROPORTIONS OF LABOR AND CAPITAL WHICH WERE COMBINED IN MANUFACTURING 1899-1922 (1899=100)

Year	Relative Amount of Labor to Capital	Year	Relative Amount of Labor to Capital
1899.....	100	1911.....	67
1900.....	98	1912.....	67
1901.....	96	1913.....	65
1902.....	97	1914.....	61
1903.....	94	1915.....	58
1904.....	84	1916.....	61
1905.....	84	1917.....	59
1906.....	82	1918.....	55
1907.....	78	1919.....	50
1908.....	65	1920.....	47
1909.....	71	1921.....	35
1910.....	69	1922.....	37

²⁹ For a description of the methods and sources used in computing the index of production for manufactures, see E. E. Day and W. M. Persons, "An Index of the Physical Volume of Production," *Review of Economic Statistics*, II (1920) 309-37; 361-67. See also Ada M. Mathews, "The Physical Volume of Production in the United States in 1924," *Ibid.*, VII. (1925), 215.

physical product. It will be noted that by 1922 the supply of capital had more than quadrupled as compared with 1899, while the labor force was only 61 per cent greater. The ratio of capital to the working force was indeed 2.67 times as great in 1922 as it had been in 1899. The increase in the physical product during this period was 140 per cent or an increase of approximately 50 per cent per worker.³⁰

5. *The Ratio of Labor to Capital.*—The changing ratio between labor and capital as compared with 1899 can be found by dividing the relative index of the labor supply by the relative index of fixed capital (L/C). This is shown in Table V. We thus have a measure of the changing proportions of the two factors throughout the years of this period.

It will be noted that since our index of labor measures the decline in the number of wage-earners employed during periods of depression while our index of capital does not show the unused capital, that during such years, the proportion of labor to capital drops sharply, with a tendency to rise during the succeeding years. The general drift is, however, of course downward because of the much more rapid increase of capital.

6. *Theory of Production.*—Relative to the indices of Production, Labor, and Capital, and the period 1899–1922 the function of Labor and Capital alone

$$P' = 1.01 L^{3/4} C^{1/4}$$

has the following properties:

- 1) To say that P' represents the actual Production P is to give particular expression to a well-known theory.
- 2) P' approaches zero as either L or C approaches zero.
- 3) P' approximates P over the period.
- 4) The deviations of P' from P are individually significant.
- 5) P' correlates closely with P when we include secular trends.
- 6) P' correlates closely with P when we eliminate secular trends.

In the sense of the foregoing let us call P' a *norm* for P over the period, and proceed to examine its properties in more detail.

(1) The theory referred to (due to J. B. Clark, Wicksteed et al.) states that Production, Labor and Capital are so related that if we multiply both Labor and Capital by a factor m then Production will be increased m times, that is Production is a first degree homogeneous function of Labor and Capital. Now P' is taken to be such a function.

(2) Among such functions the further theoretical restriction is placed upon P' that it should approach zero as either L or C approaches zero.

³⁰ As Dr. Thomas shows, the most remarkable increase in productivity has come since 1921 and is scarcely included in the above statistics.

Among functions with these properties (1) and (2) let us make a definite choice³¹ and examine the consequences of that choice, reserving the right to make other choices if we wish. Let us choose the function

$$P' = bL^kC^{1-k}$$

and find such numerical values of b and k that P' will "best" approximate P in the sense of the Theory of Least Squares. Then relative to the indices and the period we have the norm

$$P' = 1.01L^{3/4}C^{1/4}$$

(3) Given the indices of L and C , the function P' may be computed and may be compared with P in Table VI and Chart II as follows:

TABLE VI

RELATION BETWEEN (1) PRODUCT CALCULATED FROM RECORDED VALUES OF L AND C (TABLES II AND III) BY MEANS OF THE FORMULA $P' = 1.01 L^{3/4} C^{1/4}$ AND (2) RECORDED VALUES OF PRODUCT (TABLE IV)

Year	P' Product Calculated (1)	P Product Recorded (2)	Percent Deviation (2)-(1) (2)	Business Annals ³²
1899	101	100	-1	Prosperity
1900	107	101	-6	Prosperity; brief recession
1901	112	112	0	Prosperity
1902	121	122	+0.8	Prosperity
1903	126	124	-1.6	Prosperity; recession
1904	123	122	-0.8	Mild depression; revival
1905	133	143	+7.	Prosperity
1906	141	152	+7.	Prosperity
1907	148	151	+2.	Prosperity, panic, recession, depression
1908	137	126	-9.	Depression
1909	155	155	0	Revival, mild prosperity
1910	160	159	-0.6	Recession
1911	163	153	-6.5	Mild depression
1912	170	177	+4.	Revival; prosperity
1913	174	184	+5.5	Prosperity; recession
1914	171	169	-1.2	Depression
1915	179	189	+5.	Revival; prosperity
1916	209	225	+7.2	Prosperity
1917	227	227	0.	Prosperity; war activity
1918	236	223	-6.	War activity; recession
1919	233	218	-7.	Revival; prosperity
1920	236	231	-2.2	Prosperity; recession, depres- sion
1921	194	179	-8.4	Depression
1922	209	240	+18.	Revival; prosperity

³¹ This amounts to an assumption that the marginal productivity of labor is proportional to the production per unit labor and the marginal productivity of capital is proportional to production per unit capital. These properties are derived from the "chosen" function in a later section.

³² W. L. Thorp *Business Annals*, p. 138 ff.

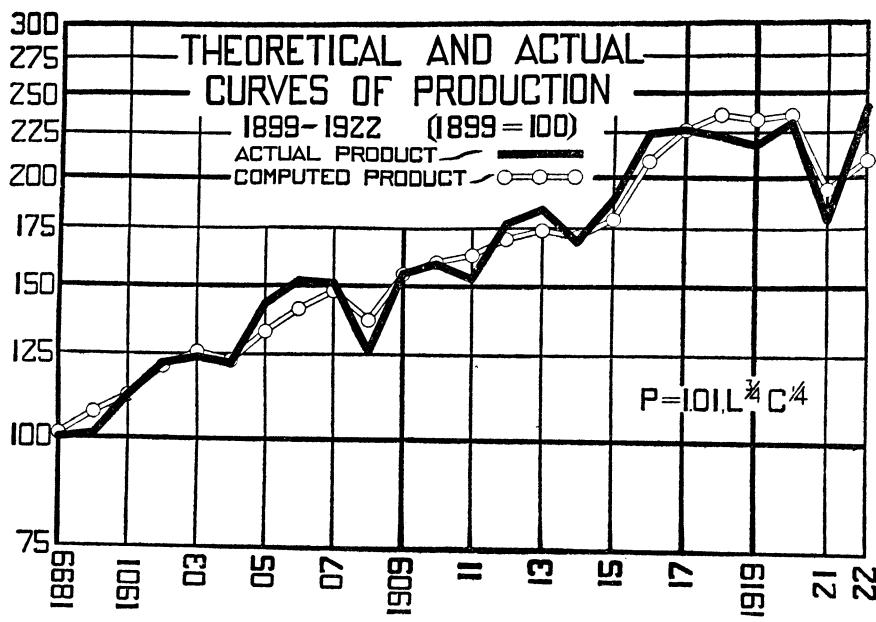


CHART II

The average percentage deviation of P' from P without regard to sign is 4.2 per cent. In fact, P lies nearer to P' than to its own moving three year average, the corresponding standard deviations being 8.7 and 11.7 respectively.

TABLE VII
DEVIATIONS FROM TREND OF P AND P' (TRENDS ARE MOVING 3 YEAR AVERAGES)

Year	Deviation of P from Trend of P	Deviation of P' from Trend of P'	Year	Deviation of P from Trend of P	Deviation of P' from Trend of P'
1900.....	-3	0	1911.....	-10	-1
1901.....	0	-1	1912.....	6	1
1902.....	3	1	1913.....	7	2
1903.....	1	3	1914.....	-12	-4
1904.....	-8	-4	1915.....	-5	-7
1905.....	4	1	1916.....	11	4
1906.....	3	0	1917.....	2	3
1907.....	8	6	1918.....	0	4
1908.....	-18	-10	1919.....	-6	-2
1909.....	7	4	1920.....	22	15
1910.....	3	1	1921.....	-38	-19

(4) It is evident from the foregoing Table VI and Chart II that the trends of P and P' (say the moving three year averages) are much alike, in fact P' was constructed so that they should be. A study of Table VII and Chart III will show also that in general P' fluctuates with the business

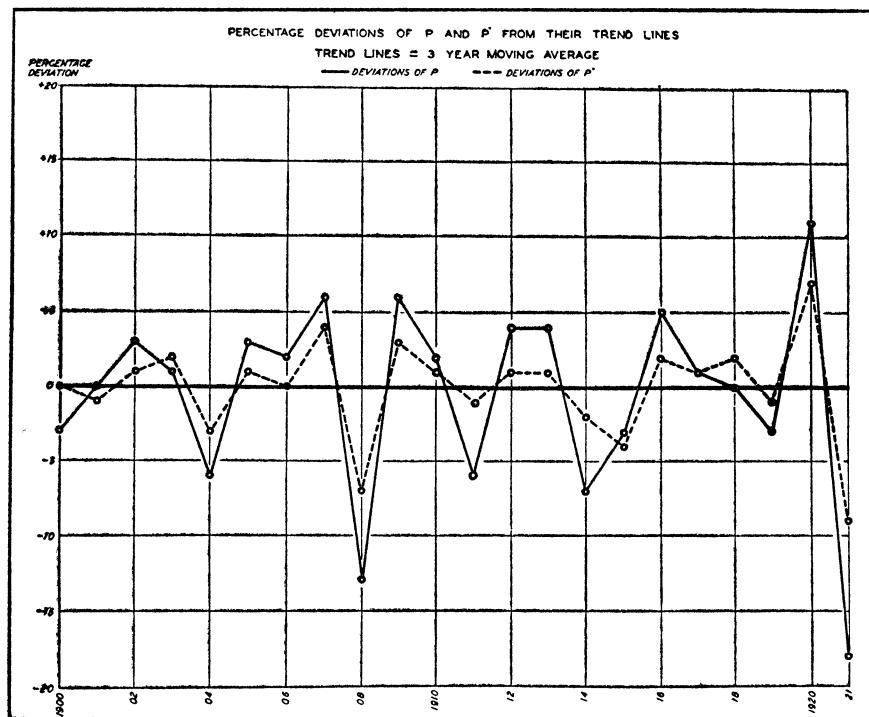


CHART III

cycle in the same direction as does P , with this difference that the oscillations of P' (relative to trend) are not as great as those of P , due to the steady influence of the steadily increasing C .

When we consider also the Business Annals as given by W. L. Thorp it is evident (Table VI and Chart IV)³³ that in general we compute too little in times of prosperity and too much in times of depression. Then not only does P' follow the business cycle but also the deviations of P' from P follow the business cycle.

(5) and (6). The coefficient of correlation between P and P' with trends included is .97 and with trends eliminated is .94.

So far we have been taking for granted that the "normal" production P' would have been produced with given quantities of labor and capital under "normal" conditions. These normal conditions are fictitious. For example the productive power of the "average" worker or of the dollar of constant purchasing power is supposed to remain constant over the period. For normal conditions management would not be more or less efficient at different times. There would be no booms nor depressions, no wars and

³³ Note that the algebraic signs of the percentages in the chart are opposite from those in the table.

**PERCENTAGE DEVIATIONS
OF COMPUTED FROM ACTUAL PRODUCT
1899 - 1922**

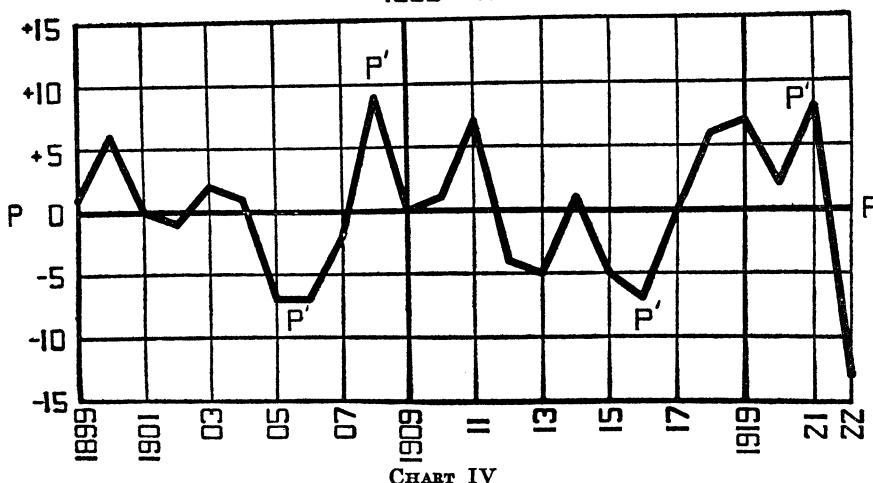


CHART IV

so on, under normal conditions. The differences between production under normal conditions and production under actual conditions may be compared as in (4) above with the *Business Annals* of the period, year by year.

Now it is possible to apply mathematical analysis to the fictitious production P' but it is not possible to apply such analysis to the actual production P unless we make (or conceal) certain further assumptions. Let us choose the following assumptions and let their justification rest on what we deduce from them.

(A) The Physical Volume of Production is proportional to the Volume of Production due to manufacturing alone.

(B) Any departure of P from P' may be represented by a change in the value of the coefficient of $L^{3/4}C^{1/4}$ so that always

$$P = bL^{3/4}C^{1/4}$$

where the value of b is independent of L and C .

These two assumptions are made in accordance with a general policy to ignore the quantitative effects of any force for which we have no quantitative data. The coefficient b is thus made a catch-all for the effects of such forces.

Making these assumptions it follows at once by mathematical analysis that:

- I. The marginal productivity of labor is $3/4 P/L$.
- II. The marginal productivity of capital is $1/4 P/C$.
- III. The productivity of total labor is $3/4 P$.
- IV. The productivity of total capital is $1/4 P$.

This imputes three-fourths of the product to labor and one-fourth to capital for the period in question.

V. The elasticity of the product with respect to small changes in labor alone is $3/4$.

VI. The elasticity of the product with respect to small changes in capital alone is $1/4$.

This means that a small percentage change in labor alone has three times the effect that would be made by the same small percentage change in capital alone.

These six theorems will be proved in the next section. It should be born in mind, however, that our results have been given exact numerical values for the sake of fixing the ideas. But the numbers themselves are fixed tentatively relative to a certain period and to certain indices. When the indices are refined or the period is changed it may be that the constant $3/4$ will appear as a constant .7 or .6 or perhaps as a variable. Even the form of the function P' may have to be changed.

It is the purpose of this paper, then, not to state results but to illustrate a method of attack. In choosing a definite Norm for Production as a first approximation it is not at all certain that we have arrived immediately at the best possible. The advantage in choosing a norm at all seems to be that it involves us in logical consequences which may be compared with the facts as we get the facts. It enables us to talk rightly or wrongly with more precision and to draw conclusions which become hypotheses.

7. *Mathematical Analysis.*—Given the function

$$P = bL^kC^{1-k}$$

where b is independent of L and C and (to fix the ideas) k is supposed to be constant and equal to $3/4$. Then the six theorems of the preceding section may be proved by the six equations which follow:

$$(1) \quad \frac{\partial P}{\partial L} = \frac{3}{4} \frac{P}{L} \quad (4) \quad C \quad \frac{\partial P}{\partial C} = \frac{1}{4}P$$

$$(2) \quad \frac{\partial P}{\partial C} = \frac{1}{4} \frac{P}{C} \quad (5) \quad \frac{\partial(\log P)}{\partial(\log L)} = \frac{3}{4}$$

$$(3) \quad \frac{L\partial P}{\partial L} = \frac{3}{4}P \quad (6) \quad \frac{\partial(\log P)}{\partial(\log C)} = \frac{1}{4}$$

If b is taken equal to 1.01 say, then

$$(7) \quad \frac{\partial P}{\partial L} = 1.01 \times \frac{3}{4} \times \left(\frac{L}{C} \right)^{-1/4}; \quad b = 1.01$$

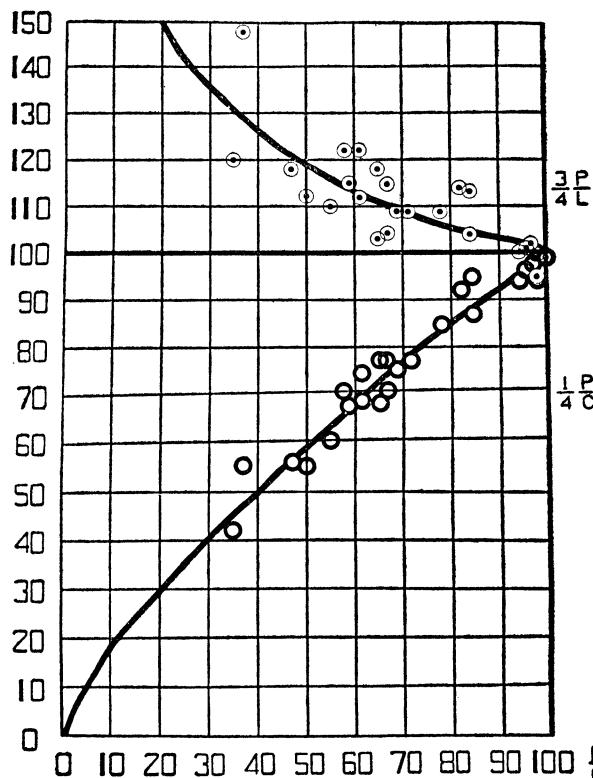
$$(8) \quad \frac{\partial P}{\partial C} = 1.01 \times \frac{1}{4} \times \left(\frac{L}{C} \right)^{3/4}; \quad b = 1.01$$

From (7) and (8) it follows that just as Production has a norm which it approximates, so the marginal productivities of labor and capital have norms which they approximate; namely, the curves $y = 1.01(L/C)^{-1/4}$ and $y = 1.01(L/C)^{3/4}$ respectively.

The three norms and the corresponding quantities are so related that if one quantity, say production, rises above its norm by 5 per cent then each of the other two quantities rises above its norm by 5 per cent. This is due to the algebraic identity

Chart V³⁴

RELATIVE FINAL PRODUCTIVITIES OF
LABOR AND CAPITAL



* In the chart the "normal" curves are taken without the coefficient 1.01 and the indices of marginal productivity are depressed proportionally.

$$\frac{P}{L} : \left(\frac{L}{C} \right)^{-1/4} = \frac{P}{C} : \left(\frac{L}{C} \right)^{3/4} = P : L^{3/4} C^{1/4} = b : 1.$$

We may now find the rates of change of the marginal productivities and total productivities by taking derivatives of equations (1) to (4) replacing the constant $3/4$ by the indefinite k and remembering that k is to be constant, positive and less than 1.

$$(9) \quad \frac{\partial}{\partial C} \left[\frac{\partial P}{\partial L} \right] = k(1-k) \frac{P}{LC}$$

$$(10) \quad \frac{\partial}{\partial L} \left[\frac{\partial P}{\partial C} \right] = k(1-k) \frac{P}{LC}$$

and hence:

The productivity of unit labor *increases* per unit increase in capital alone. The productivity of unit capital *increases* per unit increase in labor alone. These rates of increase (which are equal for fixed values of L and C) are given by the expression on the right hand side of equations (7) and (8).

$$(11) \quad \frac{\partial}{\partial L} \left[\frac{\partial P}{\partial L} \right] = k(k-1) \frac{P}{L^2}$$

and hence (diminishing returns):

The productivity of unit labor *decreases* per unit increase in labor alone (since $k-1$ is negative) at a rate given by the right hand side of equation (11).

Similarly:

$$(12) \quad \frac{\partial}{\partial C} \left[\frac{\partial P}{\partial C} \right] = k(k-1) \frac{P}{C^2}$$

and hence (diminishing returns):

The productivity of unit capital *decreases* per unit increase in capital alone at a rate given by the right hand side of equation (12).

$$(13) \quad \frac{\partial}{\partial L} \left[L \frac{\partial P}{\partial L} \right] = k^2 \frac{P}{L}$$

and hence:

The productivity of total labor *increases* per unit increase in labor alone, at a rate given by the right hand side of equation (13).

$$(14) \quad \frac{\partial}{\partial C} \left[C \frac{\partial P}{\partial C} \right] = (1-k)^2 \frac{P}{C}$$

and hence:

The productivity of total capital *increases* per unit increase in capital alone at a rate given by the right hand side of equation (14).

$$(15) \quad \frac{\partial}{\partial L} \left[C \frac{\partial P}{\partial C} \right] = k(1-k) \frac{P}{L}$$

and hence:

The productivity of total capital *increases* per unit increase in labor alone at a rate given by the right hand side of equation (15).

$$(16) \quad \frac{\partial}{\partial C} \left[L \frac{\partial P}{\partial L} \right] = k(1-k) \frac{P}{C}$$

and hence:

The productivity of total labor *increases* per unit increase in capital alone at a rate given by the right hand side of equation (16).

Finally, if k is supposed to vary then P' becomes a function of three variables, and we have a new batch of theorems for example: "If k increases while L and C remain fixed then P' increases if L/C is greater than 1, and P' decreases if L/C is less than 1."

Thus if we choose a smaller k than $3/4$ (say $2/3$ for the whole period) the P' curve thus computed will lie above the P' curve computed with $k=3/4$ whenever L/C is less than 1, that is over most of the period. The relation between P and the new $P' = 1.01 L^{2/3} C^{1/3}$ is given in the following table.

TABLE VIII
RELATION BETWEEN P AND $P' = 1.01 L^{2/3} C^{1/3}$

	P	P'	$\frac{P-P'}{P} \times 100$		P	P'	$\frac{P-P'}{P} \times 100$
1899.....	100	101	-1	1911.....	153	166	-8
1900.....	101	106	-5	1912.....	177	173	+2
1901.....	112	111	+1	1913.....	184	178	+3
1902.....	122	119	+3	1914.....	169	176	-4
1903.....	124	125	-1	1915.....	189	185	+2
1904.....	122	123	-1	1916.....	225	214	+5
1905.....	143	133	+7	1917.....	227	234	-3
1906.....	152	142	+7	1918.....	223	244	-9
1907.....	151	149	+1	1919.....	218	243	-11
1908.....	126	139	-10	1920.....	231	247	-7
1909.....	155	157	-1	1921.....	179	208	-16
1910.....	159	163	-3	1922.....	240	223	+7

8. *What Indications are there That the Theory Outlined is Valid?*—That the equation $P' = 1.01 L^{2/3} C^{1/3}$ describes in a fairly accurate manner the actual processes of production in manufacturing during this period as indicated by:

(1) The close consilience between P and P' as shown in Table VI and Chart II with a coefficient of correlation of +.97. When three year moving averages of P and P' are taken the agreement is even closer, the percentage deviation of P' and P amounting on the average (without regard to sign) to only 2.6 per cent per year instead of 4.3 per cent on the year to year observations. The cumulative error of the three year moving average of P' from the three year moving average of P is in turn only -.1 per cent.

(2) The close degree to which the theoretical curves of imputed productivity of unit labor, i.e., $y = (L/C)^{-1/4}$ and of unit capital ($y = (L/C)^{3/4}$) form the curves of best fit to the "recorded" values of unit productivity of labor and of capital.

(3) It has some times been charged that the relationship discovered between capital, labor, and manufacturing product is purely fortuitous and that equally good results would be secured by comparing the relative movement of hogs in Wisconsin, cattle in Wisconsin, with the physical product in manufacturing. But there is a logical and economic connection between labor, capital, and product which is not present in the attempted *reductio ad absurdum*. Moreover the fact that the deviations of P' and P from their respective three year moving averages move closely together as is shown by Chart III and that they have a correlation coefficient of +.94 indicates that the relationship is not merely one between factors whose secular trend happens to be upward.

(4) The fact is that the deviations of P' from P are in nearly every case precisely what one would expect. Thus during depressions, large amounts of capital are of necessity allowed to lie idle but our index of capital growth makes no allowance for this. Similarly because of the practise of short-time, the number of man hours worked decreases by a greater ratio than that of the number of men employed. Our computed index P' would therefore be expected to be greater than the actual index P . Note then that in the depression years of 1908, 1911, 1914, 1920, and 1921, P' was 9, 7, 1, 2, and 8 per cent respectively higher than P , and that during the years marked by some recession or a slight depression such as 1900, 1903, 1904, and 1910, P' was also higher than P by 6, 2, 1, and 1 per cent respectively.

Conversely, since our index of labor does not take into account overtime hours nor our index of capital the greater intensity of use which prosperity brings, it would be expected that P' would be less than P during this phase of the cycle. This is born out in practice. For the prosperous years of 1905 and 1906, P' was 7 per cent below P , and for 1907, the first three-quarters of which displayed great activity, it was 2 per cent lower. In the prosperous years of 1912 and 1913, P' was in turn 4 and 6 per cent below P and in 1915 and 1916, again 5 and 7 per cent less than P . In 1922, P' was no less than 13 per cent below P .

The only two years which constitute an exception to what we would thus expect are 1918 and 1919. These were years of business prosperity yet P'

instead of being lower than P was actually higher by 6 and 7 per cent. This may, however, have been caused by the dilution and reduced efficiency of labor which made each unit of labor actually less productive than normally.

9. *Does the Process of Distribution Approximate the Apparent Laws of Production?*—We have attempted to check this theory to see whether the processes of distribution have followed in any degree the laws of production which we believe we have traced. By the methods which have previously been described (Sections 6 and 7) the relative final physical productivities of labor for each of the years were found in terms of 1899 to be as follows:

1899.....	100	1907.....	110	1915.....	123
1900.....	96	1908.....	104	1916.....	123
1901.....	102	1909.....	110	1917.....	116
1902.....	103	1910.....	110	1918.....	111
1903.....	101	1911.....	105	1919.....	113
1904.....	105	1912.....	116	1920.....	119
1905.....	114	1913.....	119	1921.....	121
1906.....	115	1914.....	113	1922.....	149

These relative physical productivities were then multiplied by the relative exchange value of a composite unit of manufactured goods and thus the relative *value product* per laborer in each of the years as distinguished from the relative physical product was secured. It is then possible to compare the movement of this value product of final labor with the relative movement of the real wages of the workers during this period in order to determine the degree of correspondence between them.

Before entering upon such a comparison, however, it is appropriate to describe how the exchange ratio of each unit of manufactured *goods* and of manufacturing as a whole was found. This was secured by multiplying the index of physical production by the ratio between the price level of manufactured goods and the relative general price level.

$$\frac{\text{Price Index of}}{\text{Index of physical production} \times \frac{\text{Manufactured Goods}}{\text{General Price Level}}}$$

This ratio of the prices of manufactured commodities to the general price level was computed from the statistics of wholesale prices collected by the Bureau of Labor Statistics and is shown in the following table.

This index shows that when measured in terms of 1899, a unit of manufactured goods had less purchasing power in ten subsequent years, reaching its lowest point of 85 in 1910. Its exchange value was somewhat higher in the subsequent years and it rose somewhat in 1922 when it was still 10 per cent below 1899. This in turn reduced the total

value product from 240 to 217 which although also shared by 1920, was the highest point for the period.

TABLE IX
RELATIVE VALUE PRODUCT OF MANUFACTURED GOODS AND TOTAL VALUE PRODUCT OF
MANUFACTURING 1899-1922 (1899=100)

Year	Price of all Mfg Commodities 1	All Commodity Index 2	Ratio Mfg Commodities to all Commodities 3*	Total Value Product (Physical product times column 3)
1899.....	100	100	100	100
1900.....	105	108	98	99
1901.....	101	106	96	107
1902.....	103	113	91	111
1903.....	104	114	91	113
1904.....	103	114	90	109
1905.....	106	115	92	132
1906.....	112	118	95	144
1907.....	119	125	95	144
1908.....	110	120	91	115
1909.....	112	129	87	134
1910.....	115	135	85	136
1911.....	111	124	90	137
1912.....	116	132	88	156
1913.....	117	134	88	162
1914.....	113	131	86	146
1915.....	119	135	88	167
1916.....	156	169	92	207
1917.....	210	237	89	201
1918.....	226	259	87	194
1919.....	242	276	89	191
1920.....	284	302	94	217
1921.....	186	196	95	170
1922.....	179	199	90	217

* Column 3 equals column 1 divided by column 2.

The relative physical productivities of the final units of labor successive years were then multiplied by the relative exchange ratio of a unit of physical product for the appropriate year and an index of relative *value* productivity for the final units of labor in the various years was thus obtained. This was as shown in Table X with the average for the years from 1899 to 1908 taken as 100.³⁵

This was then compared with the index of real wages for manufacturing computed by one of the authors.³⁶ To avoid the assumption that correlation was perfect in the year 1899, the average for the years

³⁵ The price statistics were taken from Bulletin 390 of the Bureau of Labor Statistics. The groups of commodities included to form the price index of manufactured goods were: (1) food, (2) cloths and clothing, (3) chemicals and drugs, (4) metals and metal products, (5) building materials, (6) house furnishings, (7) among the miscellaneous commodities, leather, paper and pulp, soap, and tobacco.

³⁶ Paul H. Douglas, "The Recent Movement of Real Wages and Its Economic Significance." Supplement, *American Economic Review*, March, 1926, p. 33.

1899-1908 was instead taken as the base. The comparative table is shown on page 164.

TABLE X
RELATIVE VALUE PRODUCTIVITY PER UNIT LABOR 1899-1922

Year	Relative Value Productivity per Unit of Labor	Year	Relative Value Productivity per Unit of Labor
1899.....	101	1911.....	96
1900.....	95	1912.....	103
1901.....	99	1913.....	106
1902.....	95	1914.....	98
1903.....	93	1915.....	110
1904.....	96	1916.....	115
1905.....	106	1917.....	104
1906.....	111	1918.....	98
1907.....	105	1919.....	102
1908.....	96	1920.....	114
1909.....	97	1921.....	117
1910.....	95	1922.....	136

The coefficient of correlation between these series is + .69 with a probable error of $\pm .072$ and if a comparison is made between the seven year moving averages of the two, the coefficient is + .89 with a probable error of $\pm .03$. There is virtually no relationship, however, between the short time movements of the two, since the correlation of the deviations of each from its trend gives a coefficient of only .12.

The degree of correspondence discovered is however sufficient to give a considerable degree of corroboration to the law of production which has been worked out and to indicate that the processes of distribution follow in large measure the processes of production if sufficient time is allowed.

A further interesting comparison is afforded by the studies of the National Bureau of Economic Research into the proportion of the manufacturing product which went to labor during the decade 1909-1918. They found that wages and salaries formed on the average 74 per cent of the total value added by manufactures during these years.³⁷ We have found in our formula that when we attribute to labor 75 per cent of the product, we get a close consilience to the actual normal course of production.

There is apparently therefore a decided tendency for distribution to follow the laws of imputed productivity. Lest some be led however hastily to conclude that this lends an ethical justification to the existing social and economic order, it should be pointed out that

³⁷ National Bureau of Economic Research, *Income in the United States*, Vol. 2, p. 98. The percentages by years were as follows:

1909	72.2	1911	76.4	1913	74.5	1915	75.4	1917	71.0
1910	71.6	1912	74.5	1914	77.8	1916	68.7	1918	78.1

TABLE XI

RELATIVE MOVEMENT IN MANUFACTURING OF IMPUTED VALUE PRODUCT PER WORKER AND REAL WAGES (1899-1922) (1899-1908=100)

	(1) Value Product Unit Labor (Average 1899-1908 = 100)	(2) Real Wages (Average 1899-1908 = 100)	(3) Per Cent Deviation of (2) from (1). (2)-(1)	Business Annals (Abbreviated)
1899.....	101	99	-2	
1900.....	95	98	+3	
1901.....	99	101	+2	
1902.....	95	102	+7	
1903.....	93	100	+8	
1904.....	96	99	+3	
1905.....	106	103	-3	
1906.....	111	101	-9	Brief Recessions
1907.....	105	99	-6	
1908.....	96	94	-2	
1909.....	97	102	+5	
1910.....	95	104	+9	
1911.....	96	97	+1	Mild Depression
1912.....	103	99	-4	
1913.....	106	100	-6	
1914.....	98	99	+1	
1915.....	110	99	-10	Depression
1916.....	115	104	-10	
1917.....	104	103	-1	
1918.....	98	107	+9	War
1919.....	102	111	+9	War
1920.....	114	114	0	
1921.....	117	115	-2	
1922.....	136	119	-13	Depression

(1) Sum of deviations, without regard to sign = 125 per cent

(2) Average deviation = $\frac{125}{24} = 5.2$ per cent

(3) Sum of deviations with regard to sign = $68 + 57 = -11$ per cent

(4) Average deviations with regard to sign = $\frac{-11}{24} = -.5$ per cent

even if there were precise correspondence, it would not furnish any light upon the question as to whether capital for example should be privately owned to the degree to which it is in our society. For while capital may be "productive," it does not follow that the capitalist always is. Capital would still be "productive" even though its ownership were changed. Nor does it follow that the uses to which the capitalists put the income which they receive are on the whole socially the best. One may therefore be still a supporter of socialism, communism, or individualism and still square his social philosophy with the theory of production which we have developed.

10. *A Program for Further Work.*—In closing, it should be made clear that we do not claim to have actually solved the law of production,

but merely that we have made an approximation to it and suggested a method of attack. Future progress will be assisted by developing more refined series, by using different mathematical techniques, and by analyzing other sets of data.

Thus we may hope for: (1) An improved index of labor supply which will approximate more closely the relative actual number of hours worked not only by manual workers but also by clerical workers as well; (2) a better index of capital growth; (3) an improved index of production which will be based upon the admirable work of Dr. Thomas; (4) a more accurate index of the relative exchange value of a unit of manufactured goods.

In analyzing this data, we should (1) be prepared to devise formulas which will not necessarily be based upon constant relative "contributions" of each factor to the total product but which will allow for variations from year to year, and (2) will eliminate so far as possible the time element from the process.

We have developed our theory from the movement of labor, capital, production, value, and wages for the manufacturing industries of this country considered as a whole. There is opportunity to apply the same, or an improved method of analysis, to other lines of industry such as transportation, mining, public utilities, etc., in this country and to similar data for other countries. When this is done, we shall have most interesting material on the slope of the curves of imputed productivity for a wide variety of industries and may be able to frame combined curves for a country as a whole and from this frame interesting international comparisons.

Finally, we should ultimately look forward toward including the third factor of natural resources in our equations and of seeing to what degree this modifies our conclusions and what light it throws upon the laws of rent.

These are tasks which will require much time to complete but we submit that they are necessary if the precise relationships which probably lurk within economic phenomena are to be detected and measured.

The American Economic Review

VOLUME XXXV

SEPTEMBER, 1945

NUMBER FOUR

THE USE OF KNOWLEDGE IN SOCIETY

*By F. A. HAYEK**

I

What is the problem we wish to solve when we try to construct a rational economic order?

On certain familiar assumptions the answer is simple enough. *If* we possess all the relevant information, *if* we can start out from a given system of preferences and *if* we command complete knowledge of available means, the problem which remains is purely one of logic. That is, the answer to the question of what is the best use of the available means is implicit in our assumptions. The conditions which the solution of this optimum problem must satisfy have been fully worked out and can be stated best in mathematical form: put at their briefest, they are that the marginal rates of substitution between any two commodities or factors must be the same in all their different uses.

This, however, is emphatically *not* the economic problem which society faces. And the economic calculus which we have developed to solve this logical problem, though an important step toward the solution of the economic problem of society, does not yet provide an answer to it. The reason for this is that the "data" from which the economic calculus starts are never for the whole society "given" to a single mind which could work out the implications, and can never be so given.

The peculiar character of the problem of a rational economic order is determined precisely by the fact that the knowledge of the circumstances of which we must make use never exists in concentrated or integrated form, but solely as the dispersed bits of incomplete and frequently contradictory knowledge which all the separate individuals possess. The economic problem of society is thus not merely a problem

* The author is Tooke professor of political economy and statistics at the University of London (London School of Economics and Political Science).

of how to allocate "given" resources—if "given" is taken to mean given to a single mind which deliberately solves the problem set by these "data." It is rather a problem of how to secure the best use of resources known to any of the members of society, for ends whose relative importance only these individuals know. Or, to put it briefly, it is a problem of the utilization of knowledge not given to anyone in its totality.

This character of the fundamental problem has, I am afraid, been rather obscured than illuminated by many of the recent refinements of economic theory, particularly by many of the uses made of mathematics. Though the problem with which I want primarily to deal in this paper is the problem of a rational economic organization, I shall in its course be led again and again to point to its close connections with certain methodological questions. Many of the points I wish to make are indeed conclusions toward which diverse paths of reasoning have unexpectedly converged. But as I now see these problems, this is no accident. It seems to me that many of the current disputes with regard to both economic theory and economic policy have their common origin in a misconception about the nature of the economic problem of society. This misconception in turn is due to an erroneous transfer to social phenomena of the habits of thought we have developed in dealing with the phenomena of nature.

II

In ordinary language we describe by the word "planning" the complex of interrelated decisions about the allocation of our available resources. All economic activity is in this sense planning; and in any society in which many people collaborate, this planning, whoever does it, will in some measure have to be based on knowledge which, in the first instance, is not given to the planner but to somebody else, which somehow will have to be conveyed to the planner. The various ways in which the knowledge on which people base their plans is communicated to them is the crucial problem for any theory explaining the economic process. And the problem of what is the best way of utilizing knowledge initially dispersed among all the people is at least one of the main problems of economic policy—or of designing an efficient economic system.

The answer to this question is closely connected with that other question which arises here, that of *who* is to do the planning. It is about this question that all the dispute about "economic planning" centers. This is not a dispute about whether planning is to be done or not. It is a dispute as to whether planning is to be done centrally, by one authority for the whole economic system, or is to be divided

among many individuals. Planning in the specific sense in which the term is used in contemporary controversy necessarily means central planning—direction of the whole economic system according to one unified plan. Competition, on the other hand, means decentralized planning by many separate persons. The half-way house between the two, about which many people talk but which few like when they see it, is the delegation of planning to organized industries, or, in other words, monopoly.

Which of these systems is likely to be more efficient depends mainly on the question under which of them we can expect that fuller use will be made of the existing knowledge. And this, in turn, depends on whether we are more likely to succeed in putting at the disposal of a single central authority all the knowledge which ought to be used but which is initially dispersed among many different individuals, or in conveying to the individuals such additional knowledge as they need in order to enable them to fit their plans in with those of others.

III

It will at once be evident that on this point the position will be different with respect to different kinds of knowledge; and the answer to our question will therefore largely turn on the relative importance of the different kinds of knowledge; those more likely to be at the disposal of particular individuals and those which we should with greater confidence expect to find in the possession of an authority made up of suitably chosen experts. If it is today so widely assumed that the latter will be in a better position, this is because one kind of knowledge, namely, scientific knowledge, occupies now so prominent a place in public imagination that we tend to forget that it is not the only kind that is relevant. It may be admitted that, so far as scientific knowledge is concerned, a body of suitably chosen experts may be in the best position to command all the best knowledge available—though this is of course merely shifting the difficulty to the problem of selecting the experts. What I wish to point out is that, even assuming that this problem can be readily solved, it is only a small part of the wider problem.

Today it is almost heresy to suggest that scientific knowledge is not the sum of all knowledge. But a little reflection will show that there is beyond question a body of very important but unorganized knowledge which cannot possibly be called scientific in the sense of knowledge of general rules: the knowledge of the particular circumstances of time and place. It is with respect to this that practically every individual has some advantage over all others in that he possesses unique information of which beneficial use might be made, but of

which use can be made only if the decisions depending on it are left to him or are made with his active coöperation. We need to remember only how much we have to learn in any occupation after we have completed our theoretical training, how big a part of our working life we spend learning particular jobs, and how valuable an asset in all walks of life is knowledge of people, of local conditions, and special circumstances. To know of and put to use a machine not fully employed, or somebody's skill which could be better utilized, or to be aware of a surplus stock which can be drawn upon during an interruption of supplies, is socially quite as useful as the knowledge of better alternative techniques. And the shipper who earns his living from using otherwise empty or half-filled journeys of tramp-steamers, or the estate agent whose whole knowledge is almost exclusively one of temporary opportunities, or the *arbitrageur* who gains from local differences of commodity prices, are all performing eminently useful functions based on special knowledge of circumstances of the fleeting moment not known to others.

It is a curious fact that this sort of knowledge should today be generally regarded with a kind of contempt, and that anyone who by such knowledge gains an advantage over somebody better equipped with theoretical or technical knowledge is thought to have acted almost disreputably. To gain an advantage from better knowledge of facilities of communication or transport is sometimes regarded as almost dishonest, although it is quite as important that society make use of the best opportunities in this respect as in using the latest scientific discoveries. This prejudice has in a considerable measure affected the attitude toward commerce in general compared with that toward production. Even economists who regard themselves as definitely above the crude materialist fallacies of the past constantly commit the same mistake where activities directed toward the acquisition of such practical knowledge are concerned—apparently because in their scheme of things all such knowledge is supposed to be "given." The common idea now seems to be that all such knowledge should as a matter of course be readily at the command of everybody, and the reproach of irrationality leveled against the existing economic order is frequently based on the fact that it is not so available. This view disregards the fact that the method by which such knowledge can be made as widely available as possible is precisely the problem to which we have to find an answer.

IV

If it is fashionable today to minimize the importance of the knowledge of the particular circumstances of time and place, this is closely connected with the smaller importance which is now attached to change

as such. Indeed, there are few points on which the assumptions made (usually only implicitly) by the "planners" differ from those of their opponents as much as with regard to the significance and frequency of changes which will make substantial alterations of production plans necessary. Of course, if detailed economic plans could be laid down for fairly long periods in advance and then closely adhered to, so that no further economic decisions of importance would be required, the task of drawing up a comprehensive plan governing all economic activity would appear much less formidable.

It is, perhaps, worth stressing that economic problems arise always and only in consequence of change. So long as things continue as before, or at least as they were expected to, there arise no new problems requiring a decision, no need to form a new plan. The belief that changes, or at least day-to-day adjustments, have become less important in modern times implies the contention that economic problems also have become less important. This belief in the decreasing importance of change is, for that reason, usually held by the same people who argue that the importance of economic considerations has been driven into the background by the growing importance of technological knowledge.

Is it true that, with the elaborate apparatus of modern production, economic decisions are required only at long intervals, as when a new factory is to be erected or a new process to be introduced? Is it true that, once a plant has been built, the rest is all more or less mechanical, determined by the character of the plant, and leaving little to be changed in adapting to the ever-changing circumstances of the moment?

The fairly widespread belief in the affirmative is not, so far as I can ascertain, borne out by the practical experience of the business man. In a competitive industry at any rate—and such an industry alone can serve as a test—the task of keeping cost from rising requires constant struggle, absorbing a great part of the energy of the manager. How easy it is for an inefficient manager to dissipate the differentials on which profitability rests, and that it is possible, with the same technical facilities, to produce with a great variety of costs, are among the commonplaces of business experience which do not seem to be equally familiar in the study of the economist. The very strength of the desire, constantly voiced by producers and engineers, to be able to proceed untrammeled by considerations of money costs, is eloquent testimony to the extent to which these factors enter into their daily work.

One reason why economists are increasingly apt to forget about the constant small changes which make up the whole economic picture is probably their growing preoccupation with statistical aggregates, which

show a very much greater stability than the movements of the detail. The comparative stability of the aggregates cannot, however, be accounted for—as the statisticians seem occasionally to be inclined to do—by the “law of large numbers” or the mutual compensation of random changes. The number of elements with which we have to deal is not large enough for such accidental forces to produce stability. The continuous flow of goods and services is maintained by constant deliberate adjustments, by new dispositions made every day in the light of circumstances not known the day before, by *B* stepping in at once when *A* fails to deliver. Even the large and highly mechanized plant keeps going largely because of an environment upon which it can draw for all sorts of unexpected needs; tiles for its roof, stationery for its forms, and all the thousand and one kinds of equipment in which it cannot be self-contained and which the plans for the operation of the plant require to be readily available in the market.

This is, perhaps, also the point where I should briefly mention the fact that the sort of knowledge with which I have been concerned is knowledge of the kind which by its nature cannot enter into statistics and therefore cannot be conveyed to any central authority in statistical form. The statistics which such a central authority would have to use would have to be arrived at precisely by abstracting from minor differences between the things, by lumping together, as resources of one kind, items which differ as regards location, quality, and other particulars, in a way which may be very significant for the specific decision. It follows from this that central planning based on statistical information by its nature cannot take direct account of these circumstances of time and place, and that the central planner will have to find some way or other in which the decisions depending on them can be left to the “man on the spot.”

V

If we can agree that the economic problem of society is mainly one of rapid adaptation to changes in the particular circumstances of time and place, it would seem to follow that the ultimate decisions must be left to the people who are familiar with these circumstances, who know directly of the relevant changes and of the resources immediately available to meet them. We cannot expect that this problem will be solved by first communicating all this knowledge to a central board which, after integrating *all* knowledge, issues its orders. We must solve it by some form of decentralization. But this answers only part of our problem. We need decentralization because only thus can we ensure that the knowledge of the particular circumstances of time and place will be promptly used. But the “man on the spot” cannot decide

solely on the basis of his limited but intimate knowledge of the facts of his immediate surroundings. There still remains the problem of communicating to him such further information as he needs to fit his decisions into the whole pattern of changes of the larger economic system.

How much knowledge does he need to do so successfully? Which of the events which happen beyond the horizon of his immediate knowledge are of relevance to his immediate decision, and how much of them need he know?

There is hardly anything that happens anywhere in the world that *might* not have an effect on the decision he ought to make. But he need not know of these events as such, nor of *all* their effects. It does not matter for him *why* at the particular moment more screws of one size than of another are wanted, *why* paper bags are more readily available than canvas bags, or *why* skilled labor, or particular machine tools, have for the moment become more difficult to acquire. All that is significant for him is *how much more or less* difficult to procure they have become compared with other things with which he is also concerned, or how much more or less urgently wanted are the alternative things he produces or uses. It is always a question of the relative importance of the particular things with which he is concerned, and the causes which alter their relative importance are of no interest to him beyond the effect on those concrete things of his own environment.

It is in this connection that what I have called the economic calculus proper helps us, at least by analogy, to see how this problem can be solved, and in fact is being solved, by the price system. Even the single controlling mind, in possession of all the data for some small, self-contained economic system, would not—every time some small adjustment in the allocation of resources had to be made—go explicitly through all the relations between ends and means which might possibly be affected. It is indeed the great contribution of the pure logic of choice that it has demonstrated conclusively that even such a single mind could solve this kind of problem only by constructing and constantly using rates of equivalence (or "values," or "marginal rates of substitution"), *i.e.*, by attaching to each kind of scarce resource a numerical index which cannot be derived from any property possessed by that particular thing, but which reflects, or in which is condensed, its significance in view of the whole means-end structure. In any small change he will have to consider only these quantitative indices (or "values") in which all the relevant information is concentrated; and by adjusting the quantities one by one, he can appropriately rearrange his dispositions without having to solve the whole puzzle *ab initio*, or without needing at any stage to survey it at once in all its ramifications.

Fundamentally, in a system where the knowledge of the relevant facts is dispersed among many people, prices can act to coördinate the separate actions of different people in the same way as subjective values help the individual to coördinate the parts of his plan. It is worth contemplating for a moment a very simple and commonplace instance of the action of the price system to see what precisely it accomplishes. Assume that somewhere in the world a new opportunity for the use of some raw material, say tin, has arisen, or that one of the sources of supply of tin has been eliminated. It does not matter for our purpose—and it is very significant that it does not matter—which of these two causes has made tin more scarce. All that the users of tin need to know is that some of the tin they used to consume is now more profitably employed elsewhere, and that in consequence they must economize tin. There is no need for the great majority of them even to know where the more urgent need has arisen, or in favor of what other needs they ought to husband the supply. If only some of them know directly of the new demand, and switch resources over to it, and if the people who are aware of the new gap thus created in turn fill it from still other sources, the effect will rapidly spread throughout the whole economic system and influence not only all the uses of tin, but also those of its substitutes and the substitutes of these substitutes, the supply of all the things made of tin, and their substitutes, and so on; and all this without the great majority of those instrumental in bringing about these substitutions knowing anything at all about the original cause of these changes. The whole acts as one market, not because any of its members survey the whole field, but because their limited individual fields of vision sufficiently overlap so that through many intermediaries the relevant information is communicated to all. The mere fact that there is one price for any commodity—or rather that local prices are connected in a manner determined by the cost of transport, etc.—brings about the solution which (it is just conceptually possible) might have been arrived at by one single mind possessing all the information which is in fact dispersed among all the people involved in the process.

VI

We must look at the price system as such a mechanism for communicating information if we want to understand its real function—a function which, of course, it fulfills less perfectly as prices grow more rigid. (Even when quoted prices have become quite rigid, however, the forces which would operate through changes in price still operate to a considerable extent through changes in the other terms of the contract.) The most significant fact about this system is the economy of knowledge

with which it operates, or how little the individual participants need to know in order to be able to take the right action. In abbreviated form, by a kind of symbol, only the most essential information is passed on, and passed on only to those concerned. It is more than a metaphor to describe the price system as a kind of machinery for registering change, or a system of telecommunications which enables individual producers to watch merely the movement of a few pointers, as an engineer might watch the hands of a few dials, in order to adjust their activities to changes of which they may never know more than is reflected in the price movement.

Of course, these adjustments are probably never "perfect" in the sense in which the economist conceives of them in his equilibrium analysis. But I fear that our theoretical habits of approaching the problem with the assumption of more or less perfect knowledge on the part of almost everyone has made us somewhat blind to the true function of the price mechanism and led us to apply rather misleading standards in judging its efficiency. The marvel is that in a case like that of a scarcity of one raw material, without an order being issued, without more than perhaps a handful of people knowing the cause, tens of thousands of people whose identity could not be ascertained by months of investigation, are made to use the material or its products more sparingly; *i.e.*, they move in the right direction. This is enough of a marvel even if, in a constantly changing world, not all will hit it off so perfectly that their profit rates will always be maintained at the same constant or "normal" level.

I have deliberately used the word "marvel" to shock the reader out of the complacency with which we often take the working of this mechanism for granted. I am convinced that if it were the result of deliberate human design, and if the people guided by the price changes understood that their decisions have significance far beyond their immediate aim, this mechanism would have been acclaimed as one of the greatest triumphs of the human mind. Its misfortune is the double one that it is not the product of human design and that the people guided by it usually do not know why they are made to do what they do. But those who clamor for "conscious direction"—and who cannot believe that anything which has evolved without design (and even without our understanding it) should solve problems which we should not be able to solve consciously—should remember this: The problem is precisely how to extend the span of our utilization of resources beyond the span of the control of any one mind; and, therefore, how to dispense with the need of conscious control and how to provide inducements which will make the individuals do the desirable things without anyone having to tell them what to do.

The problem which we meet here is by no means peculiar to economics but arises in connection with nearly all truly social phenomena, with language and most of our cultural inheritance, and constitutes really the central theoretical problem of all social science. As Alfred Whitehead has said in another connection, "It is a profoundly erroneous truism, repeated by all copy-books and by eminent people when they are making speeches, that we should cultivate the habit of thinking what we are doing. The precise opposite is the case. Civilization advances by extending the number of important operations which we can perform without thinking about them." This is of profound significance in the social field. We make constant use of formulas, symbols and rules whose meaning we do not understand and through the use of which we avail ourselves of the assistance of knowledge which individually we do not possess. We have developed these practices and institutions by building upon habits and institutions which have proved successful in their own sphere and which have in turn become the foundation of the civilization we have built up.

The price system is just one of those formations which man has learned to use (though he is still very far from having learned to make the best use of it) after he had stumbled upon it without understanding it. Through it not only a division of labor but also a coördinated utilization of resources based on an equally divided knowledge has become possible. The people who like to deride any suggestion that this may be so usually distort the argument by insinuating that it asserts that by some miracle just that sort of system has spontaneously grown up which is best suited to modern civilization. It is the other way round: man has been able to develop that division of labor on which our civilization is based because he happened to stumble upon a method which made it possible. Had he not done so he might still have developed some other, altogether different, type of civilization, something like the "state" of the termite ants, or some other altogether unimaginable type. All that we can say is that nobody has yet succeeded in designing an alternative system in which certain features of the existing one can be preserved which are dear even to those who most violently assail it—such as particularly the extent to which the individual can choose his pursuits and consequently freely use his own knowledge and skill.

VII

It is in many ways fortunate that the dispute about the indispensability of the price system for any rational calculation in a complex society is now no longer conducted entirely between camps holding different political views. The thesis that without the price system we

could not preserve a society based on such extensive division of labor as ours was greeted with a howl of derision when it was first advanced by von Mises twenty-five years ago. Today the difficulties which some still find in accepting it are no longer mainly political, and this makes for an atmosphere much more conducive to reasonable discussion. When we find Leon Trotsky arguing that "economic accounting is unthinkable without market relations"; when Professor Oscar Lange promises Professor von Mises a statue in the marble halls of the future Central Planning Board; and when Professor Abba P. Lerner rediscovers Adam Smith and emphasizes that the essential utility of the price system consists in inducing the individual, while seeking his own interest, to do what is in the general interest, the differences can indeed no longer be ascribed to political prejudice. The remaining dissent seems clearly to be due to purely intellectual, and more particularly methodological, differences.

A recent statement by Professor Joseph Schumpeter in his *Capitalism, Socialism and Democracy* provides a clear illustration of one of the methodological differences which I have in mind. Its author is pre-eminent among those economists who approach economic phenomena in the light of a certain branch of positivism. To him these phenomena accordingly appear as objectively given quantities of commodities impinging directly upon each other, almost, it would seem, without any intervention of human minds. Only against this background can I account for the following (to me startling) pronouncement. Professor Schumpeter argues that the possibility of a rational calculation in the absence of markets for the factors of production follows for the theorist "from the elementary proposition that consumers in evaluating ('demanding') consumers' goods *ipso facto* also evaluate the means of production which enter into the production of these goods."¹

Taken literally, this statement is simply untrue. The consumers do nothing of the kind. What Professor Schumpeter's "*ipso facto*" presumably means is that the valuation of the factors of production is

¹ J. Schumpeter, *Capitalism, Socialism, and Democracy* (New York, Harper, 1942), p. 175. Professor Schumpeter is, I believe, also the original author of the myth that Pareto and Barone have "solved" the problem of socialist calculation. What they, and many others, did was merely to state the conditions which a rational allocation of resources would have to satisfy, and to point out that these were essentially the same as the conditions of equilibrium of a competitive market. This is something altogether different from showing how the allocation of resources satisfying these conditions can be found in practice. Pareto himself (from whom Barone has taken practically everything he has to say), far from claiming to have solved the practical problem, in fact explicitly denies that it can be solved without the help of the market. See his *Manuel d'économie pure* (2nd ed., 1927), pp. 233-34. The relevant passage is quoted in an English translation at the beginning of my article on "Socialist Calculation: The Competitive 'Solution,'" in *Economica*, New Series, Vol. VIII, No. 26 (May, 1940), p. 125.

implied in, or follows necessarily from, the valuation of consumers' goods. But this, too, is not correct. Implication is a logical relationship which can be meaningfully asserted only of propositions simultaneously present to one and the same mind. It is evident, however, that the values of the factors of production do not depend solely on the valuation of the consumers' goods but also on the conditions of supply of the various factors of production. Only to a mind to which all these facts were simultaneously known would the answer necessarily follow from the facts given to it. The practical problem, however, arises precisely because these facts are never so given to a single mind, and because, in consequence, it is necessary that in the solution of the problem knowledge should be used that is dispersed among many people.

The problem is thus in no way solved if we can show that all the facts, *if* they were known to a single mind (as we hypothetically assume them to be given to the observing economist), would uniquely determine the solution; instead we must show how a solution is produced by the interactions of people each of whom possesses only partial knowledge. To assume all the knowledge to be given to a single mind in the same manner in which we assume it to be given to us as the explaining economists is to assume the problem away and to disregard everything that is important and significant in the real world.

That an economist of Professor Schumpeter's standing should thus have fallen into a trap which the ambiguity of the term "datum" sets to the unwary can hardly be explained as a simple error. It suggests rather than there is something fundamentally wrong with an approach which habitually disregards an essential part of the phenomena with which we have to deal: the unavoidable imperfection of man's knowledge and the consequent need for a process by which knowledge is constantly communicated and acquired. Any approach, such as that of much of mathematical economics with its simultaneous equations, which in effect starts from the assumption that people's *knowledge* corresponds with the objective *facts* of the situation, systematically leaves out what is our main task to explain. I am far from denying that in our system equilibrium analysis has a useful function to perform. But when it comes to the point where it misleads some of our leading thinkers into believing that the situation which it describes has direct relevance to the solution of practical problems, it is time that we remember that it does not deal with the social process at all and that it is no more than a useful preliminary to the study of the main problem.

The American Economic Review

VOLUME XLV

MARCH, 1955

NUMBER ONE

ECONOMIC GROWTH AND INCOME INEQUALITY*

By SIMON KUZNETS

The central theme of this paper is the character and causes of long-term changes in the personal distribution of income. Does inequality in the distribution of income increase or decrease in the course of a country's economic growth? What factors determine the secular level and trends of income inequalities?

These are broad questions in a field of study that has been plagued by looseness in definitions, unusual scarcity of data, and pressures of strongly held opinions. While we cannot completely avoid the resulting difficulties, it may help to specify the characteristics of the size-of-income distributions that we want to examine and the movements of which we want to explain.

Five specifications may be listed. First, the units for which incomes are recorded and grouped should be family-expenditure units, properly adjusted for the number of persons in each—rather than income recipients for whom the relations between receipt and use of income can be widely diverse. Second, the distribution should be complete, *i.e.*, should cover all units in a country rather than a segment either at the upper or lower tail. Third, if possible we should segregate the units whose main income earners are either still in the learning or already in the retired stages of their life cycle—to avoid complicating the picture by including incomes *not* associated with full-time, full-fledged participation in economic activity. Fourth, income should be defined as it is now for national income in this country, *i.e.*, received by individuals, including income in kind, before and after direct taxes, excluding capital gains. Fifth, the units should be grouped by *secular* levels of income, free of cyclical and other transient disturbances.

For such a distribution of mature expenditure units by secular levels

* Presidential address delivered at the Sixty-seventh Annual Meeting of the American Economic Association, Detroit, Michigan, December 29, 1954.



Lenn August

Number 56 of a series of photographs of past presidents of the Association.

of income per capita, we should measure shares of some fixed ordinal groups—percentiles, deciles, quintiles, etc. In the underlying array the units should be classified by average income levels for a sufficiently long span so that they form income-status groups—say a generation or about 25 years. Within such a period, even when classified by secular income levels, units may shift from one ordinal group to another. It would, therefore, be necessary and useful to study separately the relative share of units that, throughout the generation period of reference, were continuously within a specific ordinal group, and the share of the units that moved into that specific group; and this should be done for the shares of “residents” and “migrants” within all ordinal groups. Without such a long period of reference and the resulting separation between “resident” and “migrant” units at different relative income levels, the very distinction between “low” and “high” income classes loses its meaning, particularly in a study of long-term changes in shares and in inequalities in the distribution. To say, for example, that the “lower” income classes gained or lost during the last twenty years in that their share of total income increased or decreased has meaning only if the units have been classified as members of the “lower” classes throughout those 20 years—and for those who have moved into or out of those classes recently such a statement has no significance.

Furthermore, if one may add a final touch to what is beginning to look like a statistical economist’s pipe dream, we should be able to trace secular income levels not only through a single generation but at least through two—connecting the incomes of a given generation with those of its immediate descendants. We could then distinguish units that, throughout a given generation, remain within one ordinal group and whose children—through *their* generation—are also within that group, from units that remain within a group through their generation but whose children move up or down on the relative economic scale in their time. The number of possible combinations and permutations becomes large; but it should not obscure the main design of the income structure called for—the classification by long-term income status of a given generation and of its immediate descendants. If living members of society—as producers, consumers, savers, decision-makers on secular problems—react to long-term changes in income levels and shares, data on such an income structure are essential. An economic society can then be judged by the secular level of the income share that it provides for a given generation and for its children. The important corollary is that the study of long-term changes in the income distribution must distinguish between changes in the shares of resident groups—resident within either one or two generations—and changes in the income shares of

groups that, judged by their secular levels, migrate upward or downward on the income scale.

Even if we had data to approximate the income structure just outlined, the broad question posed at the start—how income inequality changes in the process of a country's economic growth—could be answered only for growth under defined economic and social conditions. And, in fact, we shall deal with this question in terms of the experience of the now developed countries which grew under the aegis of the business enterprise. But even with this limitation, there are no statistics that can be used directly for the purpose of measuring the *secular* income structure. Indeed, I have difficulty in visualizing how such information could practicably be collected—a difficulty that may be due to lack of familiarity with the studies of our colleagues in demography and sociology who have concerned themselves with problems of generation or intergeneration mobility and status. But although we now lack data directly relevant to the secular income structure, the setting up of reasonably clear and yet difficult specifications is not merely an exercise in perfectionism. For if these specifications do approximate, and I trust that they do, the real core of our interest when we talk about shares of economic classes or long-term changes in these shares, then proper disclosure of our meaning and intentions is vitally useful. It forces us to examine and evaluate critically the data that are available; it prevents us from jumping to conclusions based on these inadequate data; it reduces the loss and waste of time involved in mechanical manipulations of the type represented by Pareto-curve-fitting to groups of data whose meaning, in terms of income concept, unit of observation, and proportion of the total universe covered, remains distressingly vague; and most important of all, it propels us toward a deliberate construction of testable bridges between the available data and the income structure that is the real focus of our interest.

I. *Trends in Income Inequality*

Forewarned of the difficulties, we turn now to the available data. These data, even when relating to complete populations, invariably classify units by income for a given year. From our standpoint, this is their major limitation. Because the data often do not permit many size-groupings, and because the difference between annual income incidence and longer-term income status has less effect if the number of classes is small and the limits of each class are wide, we use a few wide classes. This does not resolve the difficulty; and there are others due to the scantiness of data for long periods, inadequacy of the unit used—which is, at best, a family and very often a reporting unit—errors in the

data, and so on through a long list. Consequently, the trends in the income structure can be discerned but dimly, and the results considered as preliminary informed guesses.

The data are for the United States, England, and Germany—a scant sample, but at least a starting point for some inferences concerning long-term changes in the presently developed countries. The general conclusion suggested is that the relative distribution of income, as measured by annual income incidence in rather broad classes, has been moving toward equality—with these trends particularly noticeable since the 1920's but beginning perhaps in the period before the first world war.

Let me cite some figures, all for income before direct taxes, in support of this impression. In the United States, in the distribution of income among families (excluding single individuals), the shares of the two lowest quintiles rise from 13½ per cent in 1929 to 18 per cent in the years after the second world war (average of 1944, 1946, 1947, and 1950); whereas the share of the top quintile declines from 55 to 44 per cent, and that of the top 5 per cent from 31 to 20 per cent. In the United Kingdom, the share of the top 5 per cent of units declines from 46 per cent in 1880 to 43 per cent in 1910 or 1913, to 33 per cent in 1929, to 31 per cent in 1938, and to 24 per cent in 1947; the share of the lower 85 per cent remains fairly constant between 1880 and 1913, between 41 and 43 per cent, but then rises to 46 per cent in 1929 and 55 per cent in 1947. In Prussia income inequality increases slightly between 1875 and 1913—the shares of the top quintile rising from 48 to 50 per cent, of the top 5 per cent from 26 to 30 per cent; the share of the lower 60 per cent, however, remains about the same. In Saxony, the change between 1880 and 1913 is minor: the share of the two lowest quintiles declines from 15 to 14½ per cent; that of the third quintile rises from 12 to 13 per cent, of the fourth quintile from 16½ to about 18 per cent; that of the top quintile declines from 56½ to 54½ per cent, and of the top 5 per cent from 34 to 33 per cent. In Germany as a whole, relative income inequality drops fairly sharply from 1913 to the 1920's, apparently due to decimation of large fortunes and property incomes during the war and inflation; but then begins to return to prewar levels during the depression of the 1930's.¹

¹ The following sources were used in calculating the figures cited:

United States. For recent years we used *Income Distribution by Size, 1944-1950* (Washington, 1953) and Selma Goldsmith and others, "Size Distribution of Income Since the Mid-Thirties," *Rev. Econ. Stat.*, Feb. 1954, XXXVI, 1-32; for 1929, the Brookings Institution data as adjusted in Simon Kuznets, *Shares of Upper Groups in Income and Savings* (New York, 1953), p. 220.

United Kingdom. For 1938 and 1947, Dudley Seers, *The Levelling of Income Since 1938*

Even for what they are assumed to represent, let alone as approximations to shares in distributions by secular income levels, the data are such that differences of two or three percentage points cannot be assigned significance. One must judge by the general weight and consensus of the evidence—which unfortunately is limited to a few countries. It justifies a tentative impression of constancy in the relative distribution of income before taxes, followed by some narrowing of relative income inequality after the first world war—or earlier.

Three aspects of this finding should be stressed. First, the data are for income before direct taxes and exclude contributions by government (e.g., relief and free assistance). It is fair to argue that both the proportion and progressivity of direct taxes and the proportion of total income of individuals accounted for by government assistance to the less privileged economic groups have grown during recent decades. This is certainly true of the United States and the United Kingdom, but in the case of Germany is subject to further examination. It follows that the distribution of income after direct taxes and including free contributions by government would show an even greater narrowing of inequality in developed countries with size distributions of pretax, ex-government-benefits income similar to those for the United States and the United Kingdom.

Second, such stability or reduction in the inequality of the percentage shares was accompanied by significant rises in real income per capita. The countries now classified as developed have enjoyed rising per capita incomes except during catastrophic periods such as years of active world conflict. Hence, if the shares of groups classified by their annual income position can be viewed as approximations to shares of groups classified by their secular income levels, a constant percentage share of a given group means that its per capita real income is rising at the same rate as the average for all units in the country; and a reduction in inequality of the shares means that the per capita income of the lower-income groups is rising at a more rapid rate than the per capita income of the upper-income groups.

The third point can be put in the form of a question. Do the distribu-

(Oxford, 1951) p. 39; for 1929, Colin Clark, *National Income and Outlay* (London, 1937) Table 47, p. 109; for 1880, 1910, and 1913, A. Bowley, *The Change in the Distribution of the National Income, 1880-1913* (Oxford, 1920).

Germany. For the constituent areas (Prussia, Saxony and others) for years before the first world war, based on S. Prokopovich, *National Income of Western European Countries* (published in Moscow in the 1920's). Some summary results are given in Prokopovich, "The Distribution of National Income," *Econ. Jour.*, March 1926, XXXVI, 69-82. See also, "Das Deutsche Volkseinkommen vor und nach dem Kriege," *Einzelschrift zur Stat. des Deutschen Reichs*, no. 24 (Berlin, 1932), and W. S. and E. S. Woytinsky, *World Population and Production* (New York, 1953) Table 192, p. 709.

tions by annual incomes properly reflect trends in distribution by secular incomes? As technology and economic performance rise to higher levels, incomes are less subject to transient disturbances, not necessarily of the cyclical order that can be recognized and allowed for by reference to business cycle chronology, but of a more irregular type. If in the earlier years the economic fortunes of units were subject to greater vicissitudes—poor crops for some farmers, natural calamity losses for some nonfarm business units—if the over-all proportion of individual entrepreneurs whose incomes were subject to such calamities, more yesterday but some even today, was larger in earlier decades, these earlier distributions of income would be more affected by transient disturbances. In these earlier distributions the temporarily unfortunate might crowd the lower quintiles and depress their shares unduly, and the temporarily fortunate might dominate the top quintile and raise its share unduly—proportionately more than in the distributions for later years. If so, distributions by longer-term average incomes might show less reduction in inequality than do the distributions by annual incomes; they might even show an opposite trend.

One may doubt whether this qualification would upset a narrowing of inequality as marked as that for the United States, and in as short a period as twenty-five years. Nor is it likely to affect the persistent downward drift in the spread of the distributions in the United Kingdom. But I must admit a strong element of judgment in deciding how far this qualification modifies the finding of long-term stability followed by reduction in income inequality in the few developed countries for which it is observed or is likely to be revealed by existing data. The important point is that the qualification is relevant; it suggests need for further study if we are to learn much from the available data concerning the secular income structure; and such study is likely to yield results of interest in themselves in their bearing upon the problem of trends in temporal instability of income flows to individual units or to economically significant groups of units in different sectors of the national economy.

II. *An Attempt at Explanation*

If the above summary of trends in the secular income structure of developed countries comes perilously close to pure guesswork, an attempt to explain these dimly discernible trends may surely seem foolhardy. Yet it is necessary to do so if only to bring to the surface some factors that may have been at play; induce a search for data bearing upon these factors; and thus confirm or revise our impressions of the trends themselves. Such preliminary speculations are useful

provided it is recognized that we are at a relatively early stage in a long process of interplay among tentative summaries of evidence, preliminary hypotheses, and search for additional evidence that might lead to reformulation and revisions—as bases for new analysis and further search.

The present instalment of initial speculation may be introduced by saying that a long-term constancy, let alone reduction, of inequality in the secular income structure is a puzzle. For there are at least two groups of forces in the long-term operation of developed countries that make for *increasing* inequality in the distribution of income before taxes and excluding contributions by governments. The first group relates to the concentration of savings in the upper-income brackets. According to all recent studies of the apportionment of income between consumption and savings, only the upper-income groups save; the total savings of groups below the top decile are fairly close to zero. For example, the top 5 per cent of units in the United States appear to account for almost two-thirds of individuals' savings; and the top decile comes close to accounting for all of it. What is particularly important is that the inequality in distribution of savings is greater than that in the distribution of property incomes, and hence of assets.² Granted that this finding is based on distribution of annual income, and that a distribution by secular levels would show less inequality in income and correspondingly less concentration of savings, the inequality in savings would still remain fairly sharp, perhaps more so than in holdings of assets. Other conditions being equal, the cumulative effect of such inequality in savings would be the concentration of an *increasing* proportion of income-yielding assets in the hands of the upper groups—a basis for larger income shares of these groups and their descendants.

The second source of the puzzle lies in the industrial structure of the income distribution. An invariable accompaniment of growth in developed countries is the shift away from agriculture, a process usually referred to as industrialization and urbanization. The income distribution of the total population, in the simplest model, may therefore be viewed as a combination of the income distributions of the rural and of the urban populations. What little we know of the structures of these two component income distributions reveals that: (a) the average per capita income of the rural population is usually lower than that of the urban;³ (b) inequality in the percentage shares within the

² See Kuznets, *op. cit.*, particularly Chapters 2 and 6.

³ The lower level of per capita income of the agricultural or rural population compared with that of urban is fairly well established, for this country by states, and for many

distribution for the rural population is somewhat narrower than in that for the urban population—even when based on annual income; and this difference would probably be wider for distributions by secular income levels.⁴ Operating with this simple model, what conclusions do we reach? First, all other conditions being equal, the increasing weight of urban population means an increasing share for the more unequal of the two component distributions. Second, the relative difference in per capita income between the rural and urban populations does not necessarily drift downward in the process of economic growth: indeed, there is some evidence to suggest that it is stable at best, and tends to widen because per capita productivity in urban pursuits increases more rapidly than in agriculture. If this is so, inequality in the total income distribution should increase.

Two questions then arise: First, why does the share of the top-income groups show no rise over time if the concentration of savings has a cumulative effect? Second, why does income inequality decline and particularly why does the share of the lower-income groups rise if both the weight of the more unequal urban income distribution and the relative difference between per capita urban and per capita rural incomes increase?

The first question has been discussed elsewhere, although the results are still preliminary hypotheses,⁵ and it would be impossible to do more here than summarize them briefly.

Factors Counteracting the Concentration of Saving

One group of factors counteracting the cumulative effect of con-

other countries (see, e.g., a summary table of closely related measures of product and workers engaged, for various divisions of the productive system, in Colin Clark, *Conditions of Economic Progress*, 2nd ed. [London 1951], pp. 316-18). The same table suggests, for the countries with sufficiently long records, a stable or increasing relative difference between per-worker product in agriculture and per-worker product in other sectors of the economy.

⁴ This is true of the U. S. distributions prior to the second world war (see sources cited in footnote 1); in the years after the second world war the difference seems to have disappeared. It is true of the distributions for Prussia, cited by Prokopovich; and most conspicuous for India today as shown in the rough distributions by M. Mukherjee and A. K. Ghosh in "The Pattern of Income and Expenditures in the Indian Union: A Tentative Study," *International Statistical Conferences*, December 1951, Calcutta, India, Part III, pp. 49-68.

⁵ Some elements of the discussion appeared in "Proportion of Capital Formation to National Product," a paper submitted to the annual meeting of the American Economic Association in 1951 and published in *Am. Econ. Rev.*, Proceedings, May 1952, XLII, 507-26. A more elaborate statement is presented in "International Differences in Capital Formation and Financing" (particularly Appendix C, Levels and Trends in Income Shares of Upper Income Groups), a paper submitted to a Conference on Capital Formation and Economic Growth held in 1953 under the auspices of the Universities-National Bureau Committee for Economic Research. It is now in press as part of the volume of proceedings of that conference.

centration of savings upon upper-income shares is legislative interference and "political" decisions. These may be aimed at limiting the cumulation of property directly through inheritance taxes and other explicit capital levies. They may produce similar effects indirectly, e.g., by government-permitted or -induced inflation which reduces the economic value of accumulated wealth stored in fixed-price securities or other properties not fully responsive to price changes; or by legal restriction of the *yield* on accumulated property, as happened recently in the form of rent controls or of artificially low long-term interest rates maintained by the government to protect the market for its own bonds.

To discuss this complex of processes is beyond the competence of this paper, but its existence and possible wide effect should be noted and one point emphasized. All these interventions, even when not directly aimed at limiting the effects of accumulation of past savings in the hands of the few, do reflect the view of society on the long-term utility of wide income inequalities. This view is a vital force that would operate in democratic societies even if there were no other counteracting factors. This should be borne in mind in connection with *changes* in this view even in developed countries, which result from the process of growth and constitute a re-evaluation of the need for income inequalities as a source of savings for economic growth. The result of such changes would be an increasing pressure of legal and political decisions on upper-income shares—increasing as a country moves to higher economic levels.

We turn to three other, less obvious groups of factors countervailing the cumulative effects of concentration of savings. The first is demographic. In the presently developed countries there have been differential rates of increase between the rich and the poor—family control having first spread to the former. Hence, even disregarding migration, one can argue that the top 5 per cent of 1870 and its descendants would account for a significantly smaller percentage of the population in 1920. This is even more likely in a country like the United States with its substantial immigration—usually entering the income distribution at the lower-income levels; and may be less likely in a country from which the poor have emigrated. The top 5 per cent of population in 1920 is, therefore, comprised only partly of the descendants of the top 5 per cent of 1870; perhaps half or a larger fraction must have originated in the lower-income brackets of 1870. This means that the period during which effects of concentration of savings can be assumed to have cumulated to raise the income share of any given fixed ordinal group (whether it be the top 1, 5, or 10 per cent of the population) is much shorter than the fifty years in the span; and hence these effects are much weaker than they would have

been if the top 5 per cent of 1870 had, through their descendants, filled completely the ranks of the top 5 per cent of the population of 1920. Although the cumulative effect of savings may be to raise the relative income of a *progressively diminishing* top proportion of total population, their effect on the relative share of a *fixed* top proportion of the population is much reduced.

The second group of forces resides in the very nature of a dynamic economy with relative freedom of individual opportunity. In such a society technological change is rampant and property assets that originated in older industries almost inevitably have a diminishing proportional weight in the total because of the more rapid growth of younger industries. Unless the descendants of a high-income group manage to shift their accumulating assets into new fields and participate with new entrepreneurs in the growing share of the new and more profitable industries, the long-range returns on their property holdings are likely to be significantly lower than those of the more recent entrants into the class of substantial asset holders. "From shirt-sleeves to shirt-sleeves in three generations" probably exaggerates the effects of this dynamism of a growing economy: there are, among the upper-income groups of today, many descendants of the upper-income groups of more than three or even four generations ago. But the adage is realistic in the sense that a *long unbroken* sequence of connection with rising industries and hence with major sources of continued large property incomes is exceedingly rare; that the successful great entrepreneurs of today are rarely sons of the great and successful entrepreneurs of yesterday.

The third group of factors is suggested by the importance, even in the upper-income brackets, of service income. At any given time, only a limited part of the income differential of a top group is accounted for by the concentration of property yields: much of it comes from the high level of service income (professional and entrepreneurial earnings and the like). The secular rise in the upper incomes due to this source is likely to be less marked than in the service incomes of lower brackets, and for two somewhat different reasons. First, in so far as high levels of service incomes of given upper units are due to individual excellence (as is true of many professional and entrepreneurial pursuits), there is much less incentive for and possibility of keeping such incomes at continued high relative levels. Hence, the service incomes of the descendants of an *initially high* level unit are not likely to show as strong an upward trend as the incomes for the large body of population at lower-income levels. Second, a substantial part of the rising trend in per capita income is due to interindustry shift, *i.e.*, a shift of workers from lower-income to higher-income industries. The possibilities of rise

due to such interindustry shifts in the service incomes of the initially high-income groups are much more limited than for the population as a whole: they are already in high-income occupations and industries and the range for them toward higher paid occupations is more narrowly circumscribed.

These three groups of factors, even disregarding such legislative and political intervention as is indicated above, are all characteristics of a dynamic growing economy. The differentials in rate of natural increase between the upper- and the lower-income groups are true only of a rapidly growing population—with or without immigration—but accompanied by declining death rates and declining birth rates, a demographic pattern associated in the past only with the growing Western economies. The impact of new industries on obsolescence of already established wealth as a source of property income is clearly a function of rapid growth, and the more rapid the growth the greater the impact will be. The effect of interindustry shifts on the rise of per capita income, particularly of lower-income groups, is also a function of growth since only in a growing economy is there much shift in the relative importance of the several industrial sectors. One can then say, in general, that the basic factor militating against the rise in upper-income shares that would be produced by the cumulative effects of concentration of savings, is the dynamism of a growing and free economic society.

Yet while the discussion answers the original question, it yields no determinate answer as to whether the trend in income shares of upper groups is upward, downward, or constant. Even for the specific question discussed, a determinate answer depends upon the relative balance of factors—continuous concentration of savings making for an increasing share, and the offsetting forces tending to cancel this effect. To tell what the trend of upper-income shares is likely to be, we need to know much more about the weights of these conflicting pressures. Moreover, the discussion has brought to the surface factors that, in and of themselves, may cause either an upward or a downward trend in the share of upper-income groups and hence in income inequality—in distributions of annual or of secular income. For example, the new entrants into the upper groups—the upward “migrants”—who rise either because of exceptional ability or attachment to new industries or for a variety of other reasons—may be entering the fixed upper group of say the top 5 per cent with an income differential—either annual or long-term—that may be relatively *greater* than that of entrants in the preceding generation. Nothing in the argument so far excludes this possibility—which would mean a rise in the share of upper-income groups, even if the share of the old “resident” part remains constant or

even declines. Even disregarding other factors that will be noted in the next section, no firm conclusion as to trends of upper-income shares can be derived from the bare model discussed. Search for further data might yield evidence that would permit a reasonably rough but determinate conclusion; but I have no such evidence at hand.

The Shift from Agricultural to Nonagricultural Sectors

What about the trend toward greater inequality due to the shift from the agricultural to the nonagricultural sectors? In view of the importance of industrialization and urbanization in the process of economic growth, their implications for trends in the income distribution should be explored—even though we have neither the necessary data nor a reasonably complete theoretical model.

The implications can be brought out most clearly with the help of a numerical illustration (see Table I). In this illustration we deal with two sectors: agriculture (A) and all others (B). For each sector we assume percentage distributions of total sector income among sector deciles: one distribution (E) is of moderate inequality, with the shares starting at 5.5 per cent for the lowest decile and rising 1 percentage point from decile to decile to reach 14.5 per cent for the top decile; the other distribution (U) is much more unequal, the shares starting at 1 per cent for the lowest decile, and rising 2 percentage points from decile to decile to reach 19 per cent for the top decile. We assign per capita incomes to each sector: 50 units to A and 100 units to B in case I (lines 1-10 in the illustration); 50 to A and 200 to B in case II (lines 11-20). Finally, we allow the proportion of the numbers in sector A in the total number to decline from 0.8 to 0.2.

The numerical illustration is only a partial summary of the calculations, showing the shares of the lowest and highest quintiles in the income distribution for the total population under different assumptions.⁶ The basic assumptions used throughout are that the per capita income of sector B (nonagricultural) is always higher than that of sector A; that the proportion of sector A in the total number declines; and that the inequality of the income distribution within sector A may be as wide as that within sector B but not wider. With the assumptions con-

⁶ The underlying calculations are quite simple. For each case we distinguish 20 cells within the total distribution—sets of ten deciles for each sector. For each cell we compute the percentage shares of both number and income in the number and income of total population, and hence also the relative per capita income of each cell. The cells are then arrayed in increasing order of their relative per capita income and cumulated. In the resulting cumulative distributions of number and countrywide income we establish, by arithmetic interpolation, if interpolation is needed, the percentage shares in total income of the successive quintiles of the country's population.

TABLE I.—PERCENTAGE SHARES OF 1ST AND 5TH QUINTILES IN THE INCOME DISTRIBUTION FOR TOTAL POPULATION UNDER VARYING ASSUMPTIONS CONCERNING PER CAPITA INCOME WITHIN THE SECTORS, PROPORTIONS OF SECTORS IN TOTAL NUMBER, AND INTRASECTOR INCOME DISTRIBUTIONS

	Proportion of Number in Sector A to Total Number						
	0.8 (1)	0.7 (2)	0.6 (3)	0.5 (4)	0.4 (5)	0.3 (6)	0.2 (7)
I. Per Capita Income of Sector A=50; of Sector B=100							
1. Per capita income of total pop- ulation Distribution (<i>E</i>) for Both Sec- tors	60	65	70	75	80	85	90
2. Share of 1st quintile	10.5	9.9	9.6	9.3	9.4	9.8	10.2
3. Share of 5th quintile	34.2	35.8	35.7	34.7	33.2	31.9	30.4
4. Range (3-2) Distribution (<i>U</i>) for Both Sec- tors	23.7	25.9	26.1	25.3	23.9	22.1	20.2
5. Share of 1st quintile	3.8	3.8	3.7	3.7	3.8	3.8	3.9
6. Share of 5th quintile	40.7	41.9	42.9	42.7	41.5	40.2	38.7
7. Range (6-5) Distribution (<i>E</i>) for Sector A, (<i>U</i>) for Sector B	36.8	38.1	39.1	39.0	37.8	36.4	34.8
8. Share of 1st quintile	9.3	8.3	7.4	6.7	6.0	5.4	4.9
9. Share of 5th quintile	37.7	41.0	42.9	42.7	41.5	40.2	38.7
10. Range (9-8)	28.3	32.7	35.4	36.0	35.5	34.8	33.8
II. Per Capita Income of Sector A=50; of Sector B=200							
11. Per capita income of total pop- ulation Distribution (<i>E</i>) for Both Sec- tors	80	95	110	125	140	155	170
12. Share of 1st quintile	7.9	6.8	6.1	5.6	5.4	5.4	5.9
13. Share of 5th quintile	50.0	49.1	45.5	41.6	38.0	35.0	32.2
14. Range (13-12) Distribution (<i>U</i>) for Both Sec- tors	42.1	42.3	39.4	36.0	32.6	29.6	26.3
15. Share of 1st quintile	3.1	2.9	2.7	2.6	2.6	2.7	3.1
16. Share of 5th quintile	52.7	56.0	54.5	51.2	47.4	44.1	40.9
17. Range (16-15) Distribution (<i>E</i>) for Sector A, (<i>U</i>) for Sector B	49.6	53.1	51.8	48.6	44.8	41.4	37.9
18. Share of 1st quintile	7.4	6.2	5.4	4.7	4.2	3.9	3.8
19. Share of 5th quintile	51.6	56.0	54.6	51.2	47.4	44.1	40.9
20. Range (19-18)	44.2	49.8	49.2	46.5	43.2	40.2	37.2

For methods of calculating the shares of quintiles, see text (p. 12 and fn. 6). Some differences will not check because of rounding.

cerning three sets of factors—intersector differences in per capita income, intrasector distributions, and sector weights—varying within the limitations just indicated, the following conclusions are suggested:

First, if the per capita income differential increases, or if the income distribution is more unequal for sector B than for sector A, or if both conditions are present, the rise over time in the relative weight of sector B causes a marked increase in inequality in the countrywide income distribution. We have here a demonstration of the effects upon trends in income inequality of interindustry shifts away from agriculture discussed above (pp. 7-8).

Second, if the intrasector income distribution is the same for both sectors, and the widening inequality in the countrywide income distribution is due only to the increasing per capita income differential in favor of sector B, such widening is greater when the intrasector income distributions are characterized by moderate rather than wide inequality. Thus, if the intrasector distributions are of the *E* type, the range in the countrywide distribution widens from 23.7 to 26.3 as proportion of A drops from 0.8 to 0.2 and as the ratio of per capita income of sector B to that of sector A changes from 2 to 4 (see line 4, col. 1, and line 14, col. 7). If the *U* distributions are used, the range, under identical conditions, widens only from 36.8 to 37.9 (see line 7, col. 1, and line 17, col. 7). This difference is revealed more clearly by the change in the share of the 1st quintile, which bears the brunt of widening inequality: for the *E* distribution, the share drops from 10.5 (line 2, col. 1) to 5.9 (line 12, col. 7); for the *U* distribution, from 3.8 (line 5, col. 1) to 3.1 (line 15, col. 7).

Third, if the per capita income differential between sectors is constant, but the intrasector distribution of B is more unequal than that of A, the widening inequality in the countrywide distribution is the greater, the lower the assumed per capita income differential. Thus for a differential of 2 to 1, the range widens from 28.3 when the proportion of A is 0.8 (line 10, col. 1) to 36.0 at the peak when the proportion of A is 0.5 (line 10, col. 4) and is still 33.8 when the proportion of A drops to 0.2 (line 10, col. 7). For a per capita income differential of 4 to 1, the widening of the range at the maximum is only from 44.2 (line 20, col. 1) to 49.8 (line 20, col. 2) and then the range declines to 37.2 (line 20, col. 7), well below the initial level.

Fourth, the assumptions utilized in the numerical illustration—of a rise in proportions of total number in section B, of greater inequality in the distribution within sector B, and of the growing excess of per capita income in B over that in A—yield a decline in the share of the 1st quintile that is much more conspicuous than the rise in the share of the 5th quintile. Thus the share of the 1st quintile, with the proportion of A at 0.8, distribution in B more unequal than in A, and a per capita income differential of 2 to 1, is 9.3 (line 8, col. 1). As we shift to a proportion of A of 0.2, and a per capita income differential of 4 to 1, the

share of the 1st quintile drops to 3.8 (line 18, col. 7). Under the same conditions, the share of the 5th quintile changes from 37.7 (line 9, col. 1) to 40.9 (line 19, col. 7).

Fifth, even if the differential in per capita income between the two sectors remains constant and the intrasector distributions are identical for the two sectors, the mere shift in the proportions of numbers produces slight but significant changes in the distribution for the country as a whole. In general, as the proportion of A drifts from 0.8 downwards, the range tends first to widen and then to diminish. When the per capita income differential is low (2 to 1), the widening of the range reaches a peak close to middle of the series, *i.e.*, at a proportion of A equal to 0.6 (lines 4 and 7); and the movements in the range tend to be rather limited. When the per capita income differential is large (4 to 1), the range contracts as soon as the proportion of A passes the level of 0.7, and the decline in the range is quite substantial (lines 14 and 17).

Sixth, of particular bearing upon the shares of upper-income groups is the finding that the share of the top quintile declines as the proportion of A falls below a certain, rather high fraction of total numbers. There is not a single case in the illustration in which the share of the 5th quintile fails to decline, either throughout or through a substantial segment of the sequence in the downward movement of the proportion of A from 0.8 to 0.2. In lines 6 and 9, the share of the 5th quintile declines beyond the point at which the proportion of A is 0.6; and in all other relevant lines the downward trend in the share of the 5th quintile sets in earlier. The reason lies, of course, in the fact that with increasing industrialization, the growing weight of the nonagricultural sector, with its higher per capita income, raises the per capita income for the whole economy; and yet per capita income within each sector and the intrasector distributions are kept constant. Under such conditions, the upper shares would fail to decline only if there were either a greater rise in per capita income of sector B than in that of sector A; or increasing inequality in the intrasector distribution of sector B.

Several other conjectural conclusions could be drawn with additional variations in assumptions, and multiplication of sectors beyond the two distinguished in the numerical illustration. But even in the simple model illustrated the variety of possible patterns is impressive; and one is forced to the view that much more empirical information is needed to permit a proper choice of specific assumptions and constants. Granted that several of the conclusions could be generalized in formal mathematical terms, useful inferences would be within our reach only if we knew more about the specific sector distributions and the levels and trends in per capita income differentials among the sectors.

If then we limit ourselves to what is known or can be plausibly as-

sumed, the following inferences can be suggested. We know that per capita income is greater in sector B than in sector A; that, at best, the per capita income differential between sectors A and B has been fairly constant (*e.g.*, in the United States) and has perhaps more often increased; that the proportion of sector A in total numbers has diminished. Then, if we start with intrasector distribution of B more unequal than for A, we would expect results suggested by either lines 8-10 or 18-20. In the former case, the range widens as the proportion of A drops from 0.8 to 0.5, and then narrows. In the latter case, the range declines beyond the point at which the proportion of A is 0.7. But in both cases, the share of the 1st quintile declines, and fairly appreciably and continuously (see lines 8 and 18). The magnitude and continuity of the decline are partly the result of the specific assumptions made; but one would be justified in arguing that within the broad limits suggested by the illustration, the assumption of greater inequality in the intrasector distribution for sector B than for sector A, yields a downward trend in the share of the lower-income groups. Yet we find no such trend in the empirical evidence that we have. Can we assume that in the earlier periods the internal distribution for sector B was not more unequal than for sector A, despite the more recent indications that urban income distribution is more unequal than the rural?

There is, obviously, room for conjecture. It seems most plausible to assume that in earlier periods of industrialization, even when the nonagricultural population was still relatively small in the total, its income distribution was more unequal than that of the agricultural population. This would be particularly so during the periods when industrialization and urbanization were proceeding apace and the urban population was being swelled, and fairly rapidly, by immigrants—either from the country's agricultural areas or from abroad. Under these conditions, the urban population would run the full gamut from low-income positions of recent entrants to the economic peaks of the established top-income groups. The urban income inequalities might be assumed to be far wider than those for the agricultural population which was organized in relatively small individual enterprises (large-scale units were rarer then than now).

If we grant the assumption of wider inequality of distribution in sector B, the shares of the lower-income brackets should have shown a downward trend. Yet the earlier summary of empirical evidence indicates that during the last 50 to 75 years there has been no widening in income inequality in the developed countries but, on the contrary, some narrowing within the last two to four decades. It follows that the intra-sector distribution—either for sector A or for sector B—must have shown sufficient narrowing of inequality to offset the increase called

for by the factors discussed. Specifically, the shares of the *lower* income groups in sectors A and/or B must have increased sufficiently to offset the decline that would otherwise have been produced by a combination of the elements shown in the numerical illustration.

This narrowing in inequality, the offsetting rise in the shares of the lower brackets, most likely occurred in the income distribution for the urban groups, in sector B. While it may also have been present in sector A, it would have had a more limited effect on the inequality in the countrywide income distribution because of the rapidly diminishing weight of sector A in the total. Nor was such a narrowing of income inequality in agriculture likely: with industrialization, a higher level of technology permitted larger-scale units and, in the United States for example, sharpened the contrast between the large and successful business farmers and the subsistence sharecroppers of the South. Furthermore, since we accept the assumption of *initially* narrower inequality in the internal distribution of income in sector A than in sector B, any significant reduction in inequality in the former is less likely than in the latter.

Hence we may conclude that the major offset to the widening of income inequality associated with the shift from agriculture and the countryside to industry and the city must have been a rise in the income share of the lower groups within the nonagricultural sector of the population. This provides a lead for exploration in what seems to me a most promising direction: consideration of the pace and character of the economic growth of the urban population, with particular reference to the relative position of lower-income groups. Much is to be said for the notion that once the early turbulent phases of industrialization and urbanization had passed, a variety of forces converged to bolster the economic position of the lower-income groups within the urban population. The very fact that after a while, an increasing proportion of the urban population was "native," *i.e.*, born in cities rather than in the rural areas, and hence more able to take advantage of the possibilities of city life in preparation for the economic struggle, meant a better chance for organization and adaptation, a better basis for securing greater income shares than was possible for the newly "immigrant" population coming from the countryside or from abroad. The increasing efficiency of the older, established urban population should also be taken into account. Furthermore, in democratic societies the growing political power of the urban lower-income groups led to a variety of protective and supporting legislation, much of it aimed to counteract the worst effects of rapid industrialization and urbanization and to support the claims of the broad masses for more adequate shares of the growing income of the country. Space does not permit the discussion of demographic, political, and social considerations that could be brought

to bear to explain the offsets to any declines in the shares of the lower groups, declines otherwise deducible from the trends suggested in the numerical illustration.

III. Other Trends Related to Those in Income Inequality

One aspect of the conjectural conclusion just reached deserves emphasis because of its possible interrelation with other important elements in the process and theory of economic growth. The scanty empirical evidence suggests that the narrowing of income inequality in the developed countries is relatively recent and probably did not characterize the earlier stages of their growth. Likewise, the various factors that have been suggested above would explain stability and narrowing in income inequality in the later rather than in the earlier phases of industrialization and urbanization. Indeed, they would suggest widening inequality in these early phases of economic growth, especially in the older countries where the emergence of the new industrial system had shattering effects on long-established pre-industrial economic and social institutions. This timing characteristic is particularly applicable to factors bearing upon the lower-income groups: the dislocating effects of the agricultural and industrial revolutions, combined with the "swarming" of population incident upon a rapid decline in death rates and the maintenance or even rise of birth rates, would be unfavorable to the relative economic position of lower-income groups. Furthermore, there may also have been a preponderance in the earlier periods of factors favoring maintenance or increase in the shares of top-income groups: in so far as their position was bolstered by gains arising out of new industries, by an unusually rapid rate of creation of new fortunes, we would expect these forces to be relatively stronger in the early phases of industrialization than in the later when the pace of industrial growth slackens.

One might thus assume a long swing in the inequality characterizing the secular income structure: widening in the early phases of economic growth when the transition from the pre-industrial to the industrial civilization was most rapid; becoming stabilized for a while; and then narrowing in the later phases. This long secular swing would be most pronounced for older countries where the dislocation effects of the earlier phases of modern economic growth were most conspicuous; but it might be found also in the "younger" countries like the United States, if the period preceding marked industrialization could be compared with the early phases of industrialization, and if the latter could be compared with the subsequent phases of greater maturity.

If there is some evidence for assuming this long swing in relative inequality in the distribution of income before direct taxes and exclud-

ing free benefits from government, there is surely a stronger case for assuming a long swing in inequality of income net of direct taxes and including government benefits. Progressivity of income taxes and, indeed, their very importance characterize only the more recent phases of development of the presently developed countries; in narrowing income inequality they must have accentuated the downward phase of the long swing, contributing to the reversal of trend in the secular widening and narrowing of income inequality.

No adequate empirical evidence is available for checking this conjecture of a long secular swing in income inequality;⁷ nor can the phases be dated precisely. However, to make it more specific, I would place the early phase in which income inequality might have been widening, from about 1780 to 1850 in England; from about 1840 to 1890, and particularly from 1870 on in the United States; and, from the 1840's to the 1890's in Germany. I would put the phase of narrowing income inequality somewhat later in the United States and Germany than in England—perhaps beginning with the first world war in the former and in the last quarter of the 19th century in the latter.

Is there a possible relation between this secular swing in income inequality and the long swing in other important components of the growth process? For the older countries a long swing is observed in the rate of growth of population—the upward phase represented by acceleration in the rate of growth reflecting the early reduction in the death rate which was not offset by a decline in the birth rate (and in some cases was accompanied by a rise in the birth rate); and the downward phase represented by a shrinking in the rate of growth reflecting the more pronounced downward trend in the birth rate. Again, in the older countries, and also perhaps in the younger, there may have been a secular swing in the rate of urbanization, in the sense that the proportional additions to urban population and the measures of internal migration that produced this shift of population probably increased for a while—from the earlier much lower levels; but then tended to diminish as urban population came to dominate the country and as the rural reservoirs of migration became proportionally much smaller. For old, and perhaps for young countries also, there must have been a secular swing in the proportions of savings or capital formation to total economic product. Per capita product in pre-industrial times was not large enough to permit as high a nationwide rate of saving or capital formation as was attained in the course of industrial development: this is

⁷ Prokopovich's data on Prussia, from the source cited in footnote 1, indicate a substantial widening in income inequality in the early period. The share of the lower 90 per cent of the population declines from 73 per cent in 1854 to 65 per cent in 1875; the share of the top 5 per cent rises from 21 to 25 per cent. But I do not know enough about the data for the early years to evaluate the reliability of the finding.

suggested by present comparisons between net capital formation rates of 3 to 5 per cent of national product in underdeveloped countries and rates of 10 to 15 per cent in developed countries. If then, at least in the older countries, and perhaps even in the younger ones—prior to initiation of the process of modern development—we begin with low secular levels in the savings proportions, there would be a rise in the early phases to appreciably higher levels. We also know that during recent periods the net capital formation proportion and even the gross, failed to rise and perhaps even declined.

Other trends might be suggested that would possibly trace long swings similar to those for inequality in income structure, rate of growth of population, rate of urbanization and internal migration, and the proportion of savings or capital formation to national product. For example, such swings might be found in the ratio of foreign trade to domestic activities; in the aspects, if we could only measure them properly, of government activity that bear upon market forces (there must have been a phase of increasing freedom of market forces, giving way to greater intervention by government). But the suggestions already made suffice to indicate that the long swing in income inequality must be viewed as part of a wider process of economic growth, and interrelated with similar movements in other elements. The long alternation in the rate of growth of population can be seen partly as a cause, partly as an effect of the long swing in income inequality which was associated with a secular rise in real per capita income levels. The long swing in income inequality is also probably closely associated with the swing in capital formation proportions—in so far as wider inequality makes for higher, and narrower inequality for lower, country-wide savings proportions.

IV. Comparison of Developed and Underdeveloped Countries

What is the bearing of the experience of the developed countries upon the economic growth of underdeveloped countries? Let us examine briefly the data on income distribution in the latter, and speculate upon some of the implications.

As might have been expected, such data for underdeveloped countries are scanty. For the present purpose distributions of family income for India in 1949-50, for Ceylon in 1950, and for Puerto Rico in 1948 were used. While the coverage is narrow and the margin of error wide, the data show that income distribution in these underdeveloped countries is somewhat *more* unequal than in the developed countries during the period after the second world war. Thus the shares of the lower 3 quintiles are 28 per cent in India, 30 per cent in Ceylon, and 24 per cent in Puerto Rico—compared with 34 per cent in the United States and 36

per cent in the United Kingdom. The shares of the top quintile are 55 per cent in India, 50 per cent in Ceylon, and 56 per cent in Puerto Rico, compared with 44 per cent in the United States and 45 per cent in the United Kingdom.⁸

This comparison is for income before direct taxes and excluding free benefits from governments. Since the burden and progressivity of direct taxes are much greater in developed countries, and since it is in the latter that substantial volumes of free economic assistance are extended to the lower-income groups, a comparison in terms of income net of direct taxes and including government benefits would only accentuate the wider inequality of income distributions in the underdeveloped countries. Is this difference a reliable reflection of wider inequality also in the distribution of *secular* income levels in underdeveloped countries? Even disregarding the margins of error in the data, the possibility raised earlier in this paper that transient disturbances in income levels may be more conspicuous under conditions of primitive material and economic technology would affect the comparison just made. Since the distributions cited reflect the annual income levels, a greater allowance should perhaps be made for transient disturbances in the distributions for the underdeveloped than in those for the developed countries. Whether such a correction would obliterate the difference is a matter on which I have no relevant evidence.

Another consideration might tend to support this qualification. Underdeveloped countries are characterized by low average levels of income per capita, low enough to raise the question how the populations manage to survive. Let us assume that these countries represent fairly unified population groups, and exclude, for the moment, areas that combine large native populations with small enclaves of nonnative, privileged minorities, e.g., Kenya and Rhodesia, where income inequality, because of the excessively high income shares of the privileged minority, is appreciably wider than even in the underdeveloped countries cited above.⁹ On this assumption, one may infer that in countries

⁸ For sources of these data see "Regional Economic Trends and Levels of Living," submitted at the Norman Waite Harris Foundation Institute of the University of Chicago in November 1954 (in press in the volume of proceedings). This paper, and an earlier one, "Underdeveloped Countries and the Pre-industrial Phases in the Advanced Countries: An Attempt at Comparison," prepared for the World Population Meetings in Rome held in September 1954 (in press) discuss issues raised in this section.

⁹ In one year since the second world war, the non-African group in Southern Rhodesia, which accounted for only 5 per cent of total population, received 57 per cent of total income; in Kenya, the minority of only 2.9 per cent of total population, received 51 per cent of total income; in Northern Rhodesia, the minority of only 1.4 per cent of total population, received 45 per cent of total income. See United Nations, *National Income and Its Distribution in Underdeveloped Countries*, Statistical Paper, Ser. E, no. 3, 1951, Table 12, p. 19.

with low average income, the secular level of income in the lower brackets could not be below a fairly sizable proportion of average income—otherwise, the groups could not survive. This means, to use a purely hypothetical figure, that the secular level of the share of the lowest decile could not fall far short of 6 or 7 per cent, *i.e.*, the lowest decile could not have a per capita income less than six- or seven-tenths of the countrywide average. In more advanced countries, with higher average per capita incomes, even the *secular* share of the lowest bracket could easily be a smaller fraction of the countrywide average, say as small as 2 or 3 per cent for the lowest decile, *i.e.*, from a fifth to a third of the countrywide average—without implying a materially impossible economic position for that group. To be sure, there is in all countries continuous pressure to raise the relative position of the bottom-income groups; but the fact remains that the lower limit of the proportional share in the secular income structure is higher when the real countrywide per capita income is low than when it is high.

If the long-term share of the lower-income groups is larger in the underdeveloped than in the average countries, income inequality in the former should be narrower, not wider as we have found. However, if the lower brackets receive larger shares, and at the same time the very top brackets also receive larger shares—which would mean that the intermediate income classes would not show as great a progression from the bottom—the net effect may well be wider inequality. To illustrate, let us compare the distributions for India and the United States. The first quintile in India receives 8 per cent of total income, more than the 6 per cent share of the first quintile in the United States. But the second quintile in India receives only 9 per cent, the third 11, and the fourth 16; whereas in the United States, the shares of these quintiles are 12, 16, and 22 respectively. This is a rough statistical reflection of a fairly common observation relating to income distributions in underdeveloped compared with developed countries. The former have no “middle” classes: there is a sharp contrast between the preponderant proportion of population whose average income is well below the generally low countrywide average, and a small top group with a very large relative income excess. The developed countries, on the other hand, are characterized by a much more gradual rise from low to high shares, with substantial groups receiving more than the high countrywide income average, and the top groups securing smaller shares than the comparable ordinal groups in underdeveloped countries.

It is, therefore, possible that even the distributions of secular income levels would be more unequal in underdeveloped than in developed countries—not in the sense that the shares of the lower brackets would be lower in the former than in the latter, but in the sense that the shares

of the very top groups would be higher and that those of the groups below the top would all be significantly lower than a low countrywide income average. This is even more likely to be true of the distribution of income net of direct taxes and inclusive of free government benefits. But whether a high probability weight can be attached to this conjecture is a matter for further study.

In the absence of evidence to the contrary, I assume that it is true: that the secular income structure is somewhat more unequal in underdeveloped countries than in the more advanced—particularly in those of Western and Northern Europe and their economically developed descendants in the New World (the United States, Canada, Australia, and New Zealand). This conclusion has a variety of important implications and leads to some pregnant questions, of which only a few can be stated here.

In the first place, the wider inequality in the secular income structure of underdeveloped countries is associated with a much lower level of average income per capita. Two corollaries follow—and they would follow even if the income inequalities were of the same relative range in the two groups of countries. First, the impact is far sharper in the underdeveloped countries, where the failure to reach an already low countrywide average spells much greater material and psychological misery than similar proportional deviations from the average in the richer, more advanced countries. Second, positive savings are obviously possible only at much higher relative income levels in the underdeveloped countries: if in the more advanced countries some savings are possible in the fourth quintile, in the underdeveloped countries savings could be realized only at the very peak of the income pyramid, say by the top 5 or 3 per cent. If so, the concentration of savings and of assets is even more pronounced than in the developed countries; and the effects of such concentration in the past may serve to explain the peculiar characteristics of the secular income structure in underdeveloped countries today.

The second implication is that this unequal income structure presumably coexisted with a low rate of growth of income per capita. The underdeveloped countries today have not always lagged behind the presently developed areas in level of economic performance; indeed, some of the former may have been the economic leaders of the world in the centuries preceding the last two. The countries of Latin America, Africa, and particularly those of Asia, are underdeveloped today because in the last two centuries, and even in recent decades, their rate of economic growth has been far lower than that in the Western World—and low indeed, if any growth there was, on a per capita basis. The underlying shifts in industrial structure, the opportunities for internal

mobility and for economic improvement, were far more limited than in the more rapidly growing countries now in the developed category. There was no hope, within the lifetime of a generation, of a significantly perceptible rise in the level of real income, or even that the next generation might fare much better. It was this hope that served as an important and realistic compensation for the wide inequality in income distribution that characterized the presently developed countries during the earlier phases of their growth.

The third implication follows from the preceding two. It is quite possible that income inequality has not narrowed in the underdeveloped countries within recent decades. There is no empirical evidence to check this conjectural implication, but it is suggested by the absence, in these areas, of the dynamic forces associated with rapid growth that in the developed countries checked the upward trend of the upper-income shares that was due to the cumulative effect of continuous concentration of past savings; and it is also indicated by the failure of the political and social systems of underdeveloped countries to initiate the governmental or political practices that effectively bolster the weak positions of the lower-income classes. Indeed, there is a possibility that inequality in the secular income structure of underdeveloped countries may have widened in recent decades—the only qualification being that where there has been a recent shift from colonial to independent status, a privileged, *nonnative* minority may have been eliminated. But the implication, in terms of the income distribution among the *native* population proper, still remains plausible.

The somber picture just presented may be an oversimplified one. But I believe that it is sufficiently realistic to lend weight to the questions it poses—questions as to the bearing of the recent levels and trends in income inequality, and the factors that determine them, upon the future prospect of underdeveloped countries within the orbit of the free world.

The questions are difficult, but they must be faced unless we are willing completely to disregard past experience or to extrapolate mechanically oversimplified impressions of past development. The first question is: Is the pattern of the older developed countries likely to be repeated in the sense that in the early phases of industrialization in the underdeveloped countries income inequalities will tend to widen before the leveling forces become strong enough first to stabilize and then reduce income inequalities? While the future cannot be an exact repetition of the past, there are already certain elements in the present conditions of underdeveloped societies, e.g., "swarming" of population due to sharp cuts in death rates unaccompanied by declines in birth rates—that threaten to widen inequality by depressing the relative position of lower-income groups even further. Furthermore, if and when industrialization

begins, the dislocating effects on these societies, in which there is often an old hardened crust of economic and social institutions, are likely to be quite sharp—so sharp as to destroy the positions of some of the lower groups more rapidly than opportunities elsewhere in the economy may be created for them.

The next question follows from an affirmative answer to the first. Can the political framework of the underdeveloped societies withstand the strain which further widening of income inequality is likely to generate? This query is pertinent if it is realized that the real per capita income level of many underdeveloped societies today is lower than the per capita income level of the presently developed societies before *their* initial phases of industrialization. And yet the stresses of the dislocations incident to early phases of industrialization in the developed countries were sufficiently acute to strain the political and social fabric of society, force major political reforms, and sometimes result in civil war.

The answer to the second question may be negative, even granted that industrialization may be accompanied by a rise in real per capita product. If, for many groups in society, the rise is even partly offset by a decline in their proportional share in total product; if, consequently, it is accompanied by widening of income inequality, the resulting pressures and conflicts may necessitate drastic changes in social and political organization. This gives rise to the next and crucial question: How can either the institutional and political framework of the underdeveloped societies or the processes of economic growth and industrialization be modified to favor a sustained rise to higher levels of economic performance and yet avoid the fatally simple remedy of an authoritarian regime that would use the population as cannon-fodder in the fight for economic achievement? How to minimize the cost of transition and avoid paying the heavy price—in internal tensions, in long-run inefficiency in providing means for satisfying wants of human beings as individuals—which the inflation of political power represented by authoritarian regimes requires?

Facing these acute problems, one is cognizant of the dangers of taking an extreme position. One extreme—particularly tempting to us—is to favor repetition of past patterns of the now developed countries, patterns that, under the markedly different conditions of the presently underdeveloped countries, are almost bound to put a strain on the existing social and economic institutions and eventuate in revolutionary explosions and authoritarian regimes. There is danger in simple analogies; in arguing that because an unequal income distribution in Western Europe in the past led to accumulation of savings and financing of basic capital formation, the preservation or accentuation of present income inequalities in the underdeveloped countries is necessary to secure the

same result. Even disregarding the implications for the lower-income groups, we may find that in at least some of these countries today the consumption propensities of upper-income groups are far higher and savings propensities far lower than were those of the more puritanical upper-income groups of the presently developed countries. Because they may have proved favorable in the past, it is dangerous to argue that completely free markets, lack of penalties implicit in progressive taxation, and the like are indispensable for the economic growth of the now underdeveloped countries. Under present conditions the results may be quite the opposite—withdrawal of accumulated assets to relatively “safe” channels, either by flight abroad or into real estate; and the inability of governments to serve as basic agents in the kind of capital formation that is indispensable to economic growth. It is dangerous to argue that, because in the past foreign investment provided capital resources to spark satisfactory economic growth in some of the smaller European countries or in Europe's descendants across the seas, similar effects can be expected today if only the underdeveloped countries can be convinced of the need of a “favorable climate.” Yet, it is equally dangerous to take the opposite position and claim that the present problems are entirely new and that we must devise solutions that are the product of imagination unrestrained by knowledge of the past, and therefore full of romantic violence. What we need, and I am afraid it is but a truism, is a clear perception of past trends and of conditions under which they occurred, as well as knowledge of the conditions that characterize the underdeveloped countries today. With this as a beginning, we can then attempt to translate the elements of a properly understood past into the conditions of an adequately understood present.

V. Concluding Remarks

In concluding this paper, I am acutely conscious of the meagerness of reliable information presented. The paper is perhaps 5 per cent empirical information and 95 per cent speculation, some of it possibly tainted by wishful thinking. The excuse for building an elaborate structure on such a shaky foundation is a deep interest in the subject and a wish to share it with members of the Association. The formal and no less genuine excuse is that the subject is central to much of economic analysis and thinking; that our knowledge of it is inadequate; that a more cogent view of the whole field may help channel our interests and work in intellectually profitable directions; that speculation is an effective way of presenting a broad view of the field; and that so long as it is recognized as a collection of hunches calling for further investigation rather than a set of fully tested conclusions, little harm and much good may result.

Let me add two final comments. The first bears upon the importance of additional knowledge and a better view of the secular structure of personal income distribution. Since this distribution is a focal point at which the functioning of the economic system impinges upon the human beings who are the living members of society and for whom and through whom the society operates, it is an important datum for understanding the reactions and behavior patterns of human beings as producers, consumers, and savers. It follows that better knowledge and comprehension of the subject are indispensable, not only in and of itself but also as a step in learning more about the functioning of society—in both the long and short run. Without better knowledge of the trends in secular income structure and of the factors that determine them, our understanding of the whole process of economic growth is limited; and any insight we may derive from observing changes in countrywide aggregates over time will be defective if these changes are not translated into movements of shares of the various income groups.

But more than that, such knowledge will contribute to a better evaluation of past and present theorizing on the subject of economic growth. It was pointed out in the opening lines of this paper that the field is distinguished by looseness of concepts, extreme scarcity of relevant data, and, particularly, pressures of strongly held opinions. The distribution of national product among the various groups is a subject of acute interest to many and is discussed at length in any half-articulate society. When empirical data are scanty, as they are in this field, the natural tendency in such discussion is to generalize from what little experience is available—most often the short stretch of historical experience within the horizon of the interested scholar, which is brought to bear upon the particular policy problems in the forefront. It has repeatedly been observed that the grand dynamic economics of the classical school of the late 18th and early 19th centuries was a generalization, the main empirical contents of which were the observed developments during half to three quarters of a century in England, the mother country of that school; and that it bore many of the limitations which the brevity and exceptional character of that period and that place naturally imposed upon the theoretical structure. It is also possible that much of Marxian economics may be an overgeneralization of imperfectly understood trends in England during the first half of the 19th century when income inequality may have widened; and that extrapolations of these trends (*e.g.*, increasing misery of the working classes, polarization of society, etc.) proved wrong because due regard was not given to the possible effects upon the economic and social structure of technological changes, extension of the economic system to much of the then unoccupied world, and the very structure of human wants. Wider empirical foundations,

observation of a greater variety of historical experience, and a recognition that any body of generalizations tends to reflect some limited stretch of historical experience must force us to evaluate any theory—past or present—in terms of its empirical contents and the consequent limits of its applicability—a precept which naturally should also be applied to the oversimplified generalizations contained in the present paper.

My final comment relates to the directions in which further exploration of the subject is likely to lead us. Even in this simple initial sketch, findings in the field of demography were used and references to political aspects of social life were made. Uncomfortable as are such ventures into unfamiliar and perhaps treacherous fields, they can not and should not be avoided. If we are to deal adequately with processes of economic growth, processes of long-term change in which the very technological, demographic, and social frameworks are also changing—and in ways that decidedly affect the operation of economic forces proper—it is inevitable that we venture into fields beyond those recognized in recent decades as the province of economics proper. For the study of the economic growth of nations, it is imperative that we become more familiar with findings in those related social disciplines that can help us understand population growth patterns, the nature and forces in technological change, the factors that determine the characteristics and trends in political institutions, and generally patterns of behavior of human beings—partly as a biological species, partly as social animals. Effective work in this field necessarily calls for a shift from market economics to political and social economy.

The American Economic Review

VOLUME XLVIII

JUNE 1958

NUMBER THREE

THE COST OF CAPITAL, CORPORATION FINANCE AND THE THEORY OF INVESTMENT

*By FRANCO MODIGLIANI AND MERTON H. MILLER**

What is the "cost of capital" to a firm in a world in which funds are used to acquire assets whose yields are uncertain; and in which capital can be obtained by many different media, ranging from pure debt instruments, representing money-fixed claims, to pure equity issues, giving holders only the right to a pro-rata share in the uncertain venture? This question has vexed at least three classes of economists: (1) the corporation finance specialist concerned with the techniques of financing firms so as to ensure their survival and growth; (2) the managerial economist concerned with capital budgeting; and (3) the economic theorist concerned with explaining investment behavior at both the micro and macro levels.¹

In much of his formal analysis, the economic theorist at least has tended to side-step the essence of this cost-of-capital problem by proceeding as though physical assets—like bonds—could be regarded as yielding known, sure streams. Given this assumption, the theorist has concluded that the cost of capital to the owners of a firm is simply the rate of interest on bonds; and has derived the familiar proposition that the firm, acting rationally, will tend to push investment to the point

* The authors are, respectively, professor and associate professor of economics in the Graduate School of Industrial Administration, Carnegie Institute of Technology. This article is a revised version of a paper delivered at the annual meeting of the Econometric Society, December 1956. The authors express thanks for the comments and suggestions made at that time by the discussants of the paper, Evsey Domar, Robert Eisner and John Lintner, and subsequently by James Duesenberry. They are also greatly indebted to many of their present and former colleagues and students at Carnegie Tech who served so often and with such remarkable patience as a critical forum for the ideas here presented.

¹ The literature bearing on the cost-of-capital problem is far too extensive for listing here. Numerous references to it will be found throughout the paper though we make no claim to completeness. One phase of the problem which we do not consider explicitly, but which has a considerable literature of its own is the relation between the cost of capital and public utility rates. For a recent summary of the "cost-of-capital theory" of rate regulation and a brief discussion of some of its implications, the reader may refer to H. M. Somers [20].

where the marginal yield on physical assets is equal to the market rate of interest.² This proposition can be shown to follow from either of two criteria of rational decision-making which are equivalent under certainty, namely (1) the maximization of profits and (2) the maximization of market value.

According to the first criterion, a physical asset is worth acquiring if it will increase the net profit of the owners of the firm. But net profit will increase only if the expected rate of return, or yield, of the asset exceeds the rate of interest. According to the second criterion, an asset is worth acquiring if it increases the value of the owners' equity, *i.e.*, if it adds more to the market value of the firm than the costs of acquisition. But what the asset adds is given by capitalizing the stream it generates at the market rate of interest, and this capitalized value will exceed its cost if and only if the yield of the asset exceeds the rate of interest. Note that, under either formulation, the cost of capital is equal to the rate of interest on bonds, regardless of whether the funds are acquired through debt instruments or through new issues of common stock. Indeed, in a world of sure returns, the distinction between debt and equity funds reduces largely to one of terminology.

It must be acknowledged that some attempt is usually made in this type of analysis to allow for the existence of uncertainty. This attempt typically takes the form of superimposing on the results of the certainty analysis the notion of a "risk discount" to be subtracted from the expected yield (or a "risk premium" to be added to the market rate of interest). Investment decisions are then supposed to be based on a comparison of this "risk adjusted" or "certainty equivalent" yield with the market rate of interest.³ No satisfactory explanation has yet been provided, however, as to what determines the size of the risk discount and how it varies in response to changes in other variables.

Considered as a convenient approximation, the model of the firm constructed via this certainty—or certainty-equivalent—approach has admittedly been useful in dealing with some of the grosser aspects of the processes of capital accumulation and economic fluctuations. Such a model underlies, for example, the familiar Keynesian aggregate investment function in which aggregate investment is written as a function of the rate of interest—the same riskless rate of interest which appears later in the system in the liquidity-preference equation. Yet few would maintain that this approximation is adequate. At the macroeconomic level there are ample grounds for doubting that the rate of interest has

² Or, more accurately, to the marginal cost of borrowed funds since it is customary, at least in advanced analysis, to draw the supply curve of borrowed funds to the firm as a rising one. For an advanced treatment of the certainty case, see F. and V. Lutz [13].

³ The classic examples of the certainty-equivalent approach are found in J. R. Hicks [8] and O. Lange [11].

as large and as direct an influence on the rate of investment as this analysis would lead us to believe. At the microeconomic level the certainty model has little descriptive value and provides no real guidance to the finance specialist or managerial economist whose main problems cannot be treated in a framework which deals so cavalierly with uncertainty and ignores all forms of financing other than debt issues.⁴

Only recently have economists begun to face up seriously to the problem of the cost of capital *cum* risk. In the process they have found their interests and endeavors merging with those of the finance specialist and the managerial economist who have lived with the problem longer and more intimately. In this joint search to establish the principles which govern rational investment and financial policy in a world of uncertainty two main lines of attack can be discerned. These lines represent, in effect, attempts to extrapolate to the world of uncertainty each of the two criteria—profit maximization and market value maximization—which were seen to have equivalent implications in the special case of certainty. With the recognition of uncertainty this equivalence vanishes. In fact, the profit maximization criterion is no longer even well defined. Under uncertainty there corresponds to each decision of the firm not a unique profit outcome, but a plurality of mutually exclusive outcomes which can at best be described by a subjective probability distribution. The profit outcome, in short, has become a random variable and as such its maximization no longer has an operational meaning. Nor can this difficulty generally be disposed of by using the mathematical expectation of profits as the variable to be maximized. For decisions which affect the expected value will also tend to affect the dispersion and other characteristics of the distribution of outcomes. In particular, the use of debt rather than equity funds to finance a given venture may well increase the expected return to the owners, but only at the cost of increased dispersion of the outcomes.

Under these conditions the profit outcomes of alternative investment and financing decisions can be compared and ranked only in terms of a *subjective* "utility function" of the owners which weighs the expected yield against other characteristics of the distribution. Accordingly, the extrapolation of the profit maximization criterion of the certainty model has tended to evolve into utility maximization, sometimes explicitly, more frequently in a qualitative and heuristic form.⁵

The utility approach undoubtedly represents an advance over the certainty or certainty-equivalent approach. It does at least permit us

⁴ Those who have taken a "case-method" course in finance in recent years will recall in this connection the famous Liquigas case of Hunt and Williams, [9, pp. 193-96] a case which is often used to introduce the student to the cost-of-capital problem and to poke a bit of fun at the economist's certainty-model.

⁵ For an attempt at a rigorous explicit development of this line of attack, see F. Modigliani and M. Zeman [14].

to explore (within limits) some of the implications of different financing arrangements, and it does give some meaning to the "cost" of different types of funds. However, because the cost of capital has become an essentially subjective concept, the utility approach has serious drawbacks for normative as well as analytical purposes. How, for example, is management to ascertain the risk preferences of its stockholders and to compromise among their tastes? And how can the economist build a meaningful investment function in the face of the fact that any given investment opportunity might or might not be worth exploiting depending on precisely who happen to be the owners of the firm at the moment?

Fortunately, these questions do not have to be answered; for the alternative approach, based on market value maximization, can provide the basis for an operational definition of the cost of capital and a workable theory of investment. Under this approach any investment project and its concomitant financing plan must pass only the following test: Will the project, as financed, raise the market value of the firm's shares? If so, it is worth undertaking; if not, its return is less than the marginal cost of capital to the firm. Note that such a test is entirely independent of the tastes of the current owners, since market prices will reflect not only their preferences but those of all potential owners as well. If any current stockholder disagrees with management and the market over the valuation of the project, he is free to sell out and reinvest elsewhere, but will still benefit from the capital appreciation resulting from management's decision.

The potential advantages of the market-value approach have long been appreciated; yet analytical results have been meager. What appears to be keeping this line of development from achieving its promise is largely the lack of an adequate theory of the effect of financial structure on market valuations, and of how these effects can be inferred from objective market data. It is with the development of such a theory and of its implications for the cost-of-capital problem that we shall be concerned in this paper.

Our procedure will be to develop in Section I the basic theory itself and to give some brief account of its empirical relevance. In Section II, we show how the theory can be used to answer the cost-of-capital question and how it permits us to develop a theory of investment of the firm under conditions of uncertainty. Throughout these sections the approach is essentially a partial-equilibrium one focusing on the firm and "industry." Accordingly, the "prices" of certain income streams will be treated as constant and given from outside the model, just as in the standard Marshallian analysis of the firm and industry the prices of all inputs and of all other products are taken as given. We have chosen to focus at this level rather than on the economy as a whole because it

is at the level of the firm and the industry that the interests of the various specialists concerned with the cost-of-capital problem come most closely together. Although the emphasis has thus been placed on partial-equilibrium analysis, the results obtained also provide the essential building blocks for a general equilibrium model which shows how those prices which are here taken as given, are themselves determined. For reasons of space, however, and because the material is of interest in its own right, the presentation of the general equilibrium model which rounds out the analysis must be deferred to a subsequent paper.

I. *The Valuation of Securities, Leverage, and the Cost of Capital*

A. *The Capitalization Rate for Uncertain Streams*

As a starting point, consider an economy in which all physical assets are owned by corporations. For the moment, assume that these corporations can finance their assets by issuing common stock only; the introduction of bond issues, or their equivalent, as a source of corporate funds is postponed until the next part of this section.

The physical assets held by each firm will yield to the owners of the firm—its stockholders—a stream of “profits” over time; but the elements of this series need not be constant and in any event are uncertain. This stream of income, and hence the stream accruing to any share of common stock, will be regarded as extending indefinitely into the future. We assume, however, that the mean value of the stream over time, or average profit per unit of time, is finite and represents a random variable subject to a (subjective) probability distribution. We shall refer to the average value over time of the stream accruing to a given share as the return of that share; and to the mathematical expectation of this average as the expected return of the share.⁶ Although individual investors may have different views as to the shape of the probability distri-

⁶ These propositions can be restated analytically as follows: The assets of the i th firm generate a stream:

$$X_i(1), X_i(2) \dots X_i(T)$$

whose elements are random variables subject to the joint probability distribution:

$$\chi_i[X_i(1), X_i(2) \dots X_i(t)].$$

The return to the i th firm is defined as:

$$X_i = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T X_i(t).$$

X_i is itself a random variable with a probability distribution $\Phi_i(X_i)$ whose form is determined uniquely by χ_i . The expected return \bar{X}_i is defined as $\bar{X}_i = E(X_i) = \int_{X_i} X_i \Phi_i(X_i) dX_i$. If N_i is the number of shares outstanding, the return of the i th share is $x_i = (1/N_i)X_i$ with probability distribution $\phi_i(x_i)dx_i = \Phi_i(Nx_i)d(Nx_i)$ and expected value $\bar{x}_i = (1/N_i)\bar{X}_i$.

bution of the return of any share, we shall assume for simplicity that they are at least in agreement as to the expected return.⁷

This way of characterizing uncertain streams merits brief comment. Notice first that the stream is a stream of profits, not dividends. As will become clear later, as long as management is presumed to be acting in the best interests of the stockholders, retained earnings can be regarded as equivalent to a fully subscribed, pre-emptive issue of common stock. Hence, for present purposes, the division of the stream between cash dividends and retained earnings in any period is a mere detail. Notice also that the uncertainty attaches to the mean value over time of the stream of profits and should not be confused with variability over time of the successive elements of the stream. That variability and uncertainty are two totally different concepts should be clear from the fact that the elements of a stream can be variable even though known with certainty. It can be shown, furthermore, that whether the elements of a stream are sure or uncertain, the effect of variability per se on the valuation of the stream is at best a second-order one which can safely be neglected for our purposes (and indeed most others too).⁸

The next assumption plays a strategic role in the rest of the analysis. We shall assume that firms can be divided into "equivalent return" classes such that the return on the shares issued by any firm in any given class is proportional to (and hence perfectly correlated with) the return on the shares issued by any other firm in the same class. This assumption implies that the various shares within the same class differ, at most, by a "scale factor." Accordingly, if we adjust for the difference in scale, by taking the *ratio* of the return to the expected return, the probability distribution of that ratio is identical for all shares in the class. It follows that all relevant properties of a share are uniquely characterized by specifying (1) the class to which it belongs and (2) its expected return.

The significance of this assumption is that it permits us to classify firms into groups within which the shares of different firms are "homogeneous," that is, perfect substitutes for one another. We have, thus, an analogue to the familiar concept of the industry in which it is the commodity produced by the firms that is taken as homogeneous. To complete this analogy with Marshallian price theory, we shall assume in the

⁷ To deal adequately with refinements such as differences among investors in estimates of expected returns would require extensive discussion of the theory of portfolio selection. Brief references to these and related topics will be made in the succeeding article on the general equilibrium model.

⁸ The reader may convince himself of this by asking how much he would be willing to rebate to his employer for the privilege of receiving his annual salary in equal monthly installments rather than in irregular amounts over the year. See also J. M. Keynes [10, esp. pp. 53-54].

analysis to follow that the shares concerned are traded in perfect markets under conditions of atomistic competition.⁹

From our definition of homogeneous classes of stock it follows that in equilibrium in a perfect capital market the price per dollar's worth of expected return must be the same for all shares of any given class. Or, equivalently, in any given class the price of every share must be proportional to its expected return. Let us denote this factor of proportionality for any class, say the k th class, by $1/\rho_k$. Then if p_j denotes the price and \bar{x}_j is the expected return per share of the j th firm in class k , we must have:

$$(1) \quad p_j = \frac{1}{\rho_k} \bar{x}_j;$$

or, equivalently,

$$(2) \quad \frac{\bar{x}_j}{p_j} = \rho_k \text{ a constant for all firms } j \text{ in class } k.$$

The constants ρ_k (one for each of the k classes) can be given several economic interpretations: (a) From (2) we see that each ρ_k is the expected rate of return of any share in class k . (b) From (1) $1/\rho_k$ is the price which an investor has to pay for a dollar's worth of expected return in the class k . (c) Again from (1), by analogy with the terminology for perpetual bonds, ρ_k can be regarded as the market rate of capitalization for the expected value of the uncertain streams of the kind generated by the k th class of firms.¹⁰

B. Debt Financing and Its Effects on Security Prices

Having developed an apparatus for dealing with uncertain streams we can now approach the heart of the cost-of-capital problem by dropping the assumption that firms cannot issue bonds. The introduction of debt-financing changes the market for shares in a very fundamental way. Because firms may have different proportions of debt in their capi-

⁹ Just what our classes of stocks contain and how the different classes can be identified by outside observers are empirical questions to which we shall return later. For the present, it is sufficient to observe: (1) Our concept of a class, while not identical to that of the industry is at least closely related to it. Certainly the basic characteristics of the probability distributions of the returns on assets will depend to a significant extent on the product sold and the technology used. (2) What are the appropriate class boundaries will depend on the particular problem being studied. An economist concerned with general tendencies in the market, for example, might well be prepared to work with far wider classes than would be appropriate for an investor planning his portfolio, or a firm planning its financial strategy.

¹⁰ We cannot, on the basis of the assumptions so far, make any statements about the relationship or spread between the various ρ 's or capitalization rates. Before we could do so we would have to make further specific assumptions about the way investors believe the probability distributions vary from class to class, as well as assumptions about investors' preferences as between the characteristics of different distributions.

tal structure, shares of different companies, even in the same class, can give rise to different probability distributions of returns. In the language of finance, the shares will be subject to different degrees of financial risk or "leverage" and hence they will no longer be perfect substitutes for one another.

To exhibit the mechanism determining the relative prices of shares under these conditions, we make the following two assumptions about the nature of bonds and the bond market, though they are actually stronger than is necessary and will be relaxed later: (1) All bonds (including any debts issued by households for the purpose of carrying shares) are assumed to yield a constant income per unit of time, and this income is regarded as certain by all traders regardless of the issuer. (2) Bonds, like stocks, are traded in a perfect market, where the term perfect is to be taken in its usual sense as implying that any two commodities which are perfect substitutes for each other must sell, in equilibrium, at the same price. It follows from assumption (1) that all bonds are in fact perfect substitutes up to a scale factor. It follows from assumption (2) that they must all sell at the same price per dollar's worth of return, or what amounts to the same thing must yield the same rate of return. This rate of return will be denoted by r and referred to as the rate of interest or, equivalently, as the capitalization rate for sure streams. We now can derive the following two basic propositions with respect to the valuation of securities in companies with different capital structures:

Proposition I. Consider any company j and let \bar{X}_j stand as before for the expected return on the assets owned by the company (that is, its expected profit before deduction of interest). Denote by D_j the market value of the debts of the company; by S_j the market value of its common shares; and by $V_j \equiv S_j + D_j$, the market value of all its securities or, as we shall say, the market value of the firm. Then, our Proposition I asserts that we must have in equilibrium:

$$(3) \quad V_j \equiv (S_j + D_j) = \bar{X}_j / \rho_k, \text{ for any firm } j \text{ in class } k.$$

That is, the *market value of any firm is independent of its capital structure and is given by capitalizing its expected return at the rate ρ_k appropriate to its class.*

This proposition can be stated in an equivalent way in terms of the firm's "average cost of capital," \bar{X}_j/V_j , which is the ratio of its expected return to the market value of all its securities. Our proposition then is:

$$(4) \quad \frac{\bar{X}_j}{(S_j + D_j)} \equiv \frac{\bar{X}_j}{V_j} = \rho_k, \text{ for any firm } j, \text{ in class } k.$$

That is, *the average cost of capital to any firm is completely independent of*

its capital structure and is equal to the capitalization rate of a pure equity stream of its class.

To establish Proposition I we will show that as long as the relations (3) or (4) do not hold between any pair of firms in a class, arbitrage will take place and restore the stated equalities. We use the term arbitrage advisedly. For if Proposition I did not hold, an investor could buy and sell stocks and bonds in such a way as to exchange one income stream for another stream, identical in all relevant respects but selling at a lower price. The exchange would therefore be advantageous to the investor quite independently of his attitudes toward risk.¹¹ As investors exploit these arbitrage opportunities, the value of the overpriced shares will fall and that of the underpriced shares will rise, thereby tending to eliminate the discrepancy between the market values of the firms.

By way of proof, consider two firms in the same class and assume for simplicity only, that the expected return, X , is the same for both firms. Let company 1 be financed entirely with common stock while company 2 has some debt in its capital structure. Suppose first the value of the levered firm, V_2 , to be larger than that of the unlevered one, V_1 . Consider an investor holding s_2 dollars' worth of the shares of company 2, representing a fraction α of the total outstanding stock, S_2 . The return from this portfolio, denoted by Y_2 , will be a fraction α of the income available for the stockholders of company 2, which is equal to the total return X_2 less the interest charge, rD_2 . Since under our assumption of homogeneity, the anticipated total return of company 2, X_2 , is, under all circumstances, the same as the anticipated total return to company 1, X_1 , we can hereafter replace X_2 and X_1 by a common symbol X . Hence, the return from the initial portfolio can be written as:

$$(5) \quad Y_2 = \alpha(X - rD_2).$$

Now suppose the investor sold his αS_2 worth of company 2 shares and acquired instead an amount $s_1 = \alpha(S_2 + D_2)$ of the shares of company 1. He could do so by utilizing the amount αS_2 realized from the sale of his initial holding and borrowing an additional amount αD_2 on his own credit, pledging his new holdings in company 1 as a collateral. He would thus secure for himself a fraction $s_1/S_1 = \alpha(S_2 + D_2)/S_1$ of the shares and earnings of company 1. Making proper allowance for the interest payments on his personal debt αD_2 , the return from the new portfolio, Y_1 , is given by:

¹¹ In the language of the theory of choice, the exchanges are movements from inefficient points in the interior to efficient points on the boundary of the investor's opportunity set; and not movements between efficient points along the boundary. Hence for this part of the analysis nothing is involved in the way of specific assumptions about investor attitudes or behavior other than that investors behave consistently and prefer more income to less income, *ceteris paribus*.

$$(6) \quad Y_1 = \frac{\alpha(S_2 + D_2)}{S_1} X - r\alpha D_2 = \alpha \frac{V_2}{V_1} X - r\alpha D_2.$$

Comparing (5) with (6) we see that as long as $V_2 > V_1$ we must have $Y_1 > Y_2$, so that it pays owners of company 2's shares to sell their holdings, thereby depressing S_2 and hence V_2 ; and to acquire shares of company 1, thereby raising S_1 and thus V_1 . We conclude therefore that levered companies cannot command a premium over unlevered companies because investors have the opportunity of putting the equivalent leverage into their portfolio directly by borrowing on personal account.

Consider now the other possibility, namely that the market value of the levered company V_2 is less than V_1 . Suppose an investor holds initially an amount s_1 of shares of company 1, representing a fraction α of the total outstanding stock, S_1 . His return from this holding is:

$$Y_1 = \frac{s_1}{S_1} X = \alpha X.$$

Suppose he were to exchange this initial holding for another portfolio, also worth s_1 , but consisting of s_2 dollars of stock of company 2 and of d dollars of bonds, where s_2 and d are given by:

$$(7) \quad s_2 = \frac{S_2}{V_2} s_1, \quad d = \frac{D_2}{V_2} s_1.$$

In other words the new portfolio is to consist of stock of company 2 and of bonds in the proportions S_2/V_2 and D_2/V_2 , respectively. The return from the stock in the new portfolio will be a fraction s_2/S_2 of the total return to stockholders of company 2, which is $(X - rD_2)$, and the return from the bonds will be rd . Making use of (7), the total return from the portfolio, Y_2 , can be expressed as follows:

$$Y_2 = \frac{s_2}{S_2} (X - rD_2) + rd = \frac{s_1}{V_2} (X - rD_2) + r \frac{D_2}{V_2} s_1 = \frac{s_1}{V_2} X = \alpha \frac{S_1}{V_2} X$$

(since $s_1 = \alpha S_1$). Comparing Y_2 with Y_1 we see that, if $V_2 < S_1 \equiv V_1$, then Y_2 will exceed Y_1 . Hence it pays the holders of company 1's shares to sell these holdings and replace them with a mixed portfolio containing an appropriate fraction of the shares of company 2.

The acquisition of a mixed portfolio of stock of a levered company j and of bonds in the proportion S_j/V_j and D_j/V_j respectively, may be regarded as an operation which "undoes" the leverage, giving access to an appropriate fraction of the unlevered return X_j . It is this possibility of undoing leverage which prevents the value of levered firms from being consistently less than those of unlevered firms, or more generally prevents the average cost of capital \bar{X}_j/V_j from being systematically higher for levered than for nonlevered companies in the same class.

Since we have already shown that arbitrage will also prevent V_2 from being larger than V_1 , we can conclude that in equilibrium we must have $V_2 = V_1$, as stated in Proposition I.

Proposition II. From Proposition I we can derive the following proposition concerning the rate of return on common stock in companies whose capital structure includes some debt: the expected rate of return or yield, i , on the stock of any company j belonging to the k th class is a linear function of leverage as follows:

$$(8) \quad i_j = \rho_k + (\rho_k - r) D_j / S_j.$$

That is, *the expected yield of a share of stock is equal to the appropriate capitalization rate ρ_k for a pure equity stream in the class, plus a premium related to financial risk equal to the debt-to-equity ratio times the spread between ρ_k and r .* Or equivalently, the market price of any share of stock is given by capitalizing its expected return at the continuously variable rate i , of (8).¹²

A number of writers have stated close equivalents of our Proposition I although by appealing to intuition rather than by attempting a proof and only to insist immediately that the results were not applicable to the actual capital markets.¹³ Proposition II, however, so far as we have been able to discover is new.¹⁴ To establish it we first note that, by definition, the expected rate of return, i , is given by:

$$(9) \quad i_j \equiv \frac{\bar{X}_j - r D_j}{S_j}.$$

From Proposition I, equation (3), we know that:

$$\bar{X}_j = \rho_k (S_j + D_j).$$

Substituting in (9) and simplifying, we obtain equation (8).

¹² To illustrate, suppose $\bar{X} = 1000$, $D = 4000$, $r = 5$ per cent and $\rho_k = 10$ per cent. These values imply that $V = 10,000$ and $S = 6000$ by virtue of Proposition I. The expected yield or rate of return per share is then:

$$i = \frac{1000 - 200}{6000} = .1 + (.1 - .05) \frac{4000}{6000} = 13\frac{1}{3} \text{ per cent.}$$

¹³ See, for example, J. B. Williams [21, esp. pp. 72-73]; David Durand [3]; and W. A. Morton [15]. None of these writers describe in any detail the mechanism which is supposed to keep the average cost of capital constant under changes in capital structure. They seem, however, to be visualizing the equilibrating mechanism in terms of switches by investors between stocks and bonds as the yields of each get out of line with their "riskiness." This is an argument quite different from the pure arbitrage mechanism underlying our proof, and the difference is crucial. Regarding Proposition I as resting on investors' attitudes toward risk leads inevitably to a misunderstanding of many factors influencing relative yields such as, for example, limitations on the portfolio composition of financial institutions. See below, esp. Section I.D.

¹⁴ Morton does make reference to a linear yield function but only "... for the sake of simplicity and because the particular function used makes no essential difference in my conclusions" [15, p. 443, note 2].

C. Some Qualifications and Extensions of the Basic Propositions

The methods and results developed so far can be extended in a number of useful directions, of which we shall consider here only three: (1) allowing for a corporate profits tax under which interest payments are deductible; (2) recognizing the existence of a multiplicity of bonds and interest rates; and (3) acknowledging the presence of market imperfections which might interfere with the process of arbitrage. The first two will be examined briefly in this section with some further attention given to the tax problem in Section II. Market imperfections will be discussed in Part D of this section in the course of a comparison of our results with those of received doctrines in the field of finance.

Effects of the Present Method of Taxing Corporations. The deduction of interest in computing taxable corporate profits will prevent the arbitrage process from making the value of all firms in a given class proportional to the expected returns generated by their physical assets. Instead, it can be shown (by the same type of proof used for the original version of Proposition I) that the market values of firms in each class must be proportional in equilibrium to their expected return net of taxes (that is, to the sum of the interest paid and expected net stockholder income). This means we must replace each \bar{X}_j in the original versions of Propositions I and II with a new variable \bar{X}_j^{τ} representing the total income net of taxes generated by the firm:

$$(10) \quad \bar{X}_j^{\tau} \equiv (\bar{X}_j - rD_j)(1 - \tau) + rD_j \equiv \bar{\pi}_j^{\tau} + rD_j,$$

where $\bar{\pi}_j^{\tau}$ represents the expected net income accruing to the common stockholders and τ stands for the average rate of corporate income tax.¹⁶

After making these substitutions, the propositions, when adjusted for taxes, continue to have the same form as their originals. That is, Proposition I becomes:

$$(11) \quad \frac{\bar{X}_j^{\tau}}{V_j} = \rho_k^{\tau}, \text{ for any firm in class } k,$$

and Proposition II becomes

$$(12) \quad i_j \equiv \frac{\bar{\pi}_j^{\tau}}{S_j} = \rho_j^{\tau} + (\rho_k^{\tau} - r)D_j/S_j$$

where ρ_k^{τ} is the capitalization rate for income net of taxes in class k .

Although the form of the propositions is unaffected, certain interpretations must be changed. In particular, the after-tax capitalization rate

¹⁶ For simplicity, we shall ignore throughout the tiny element of progression in our present corporate tax and treat τ as a constant independent of $(\bar{X}_j - rD_j)$.

ρ_k^r can no longer be identified with the "average cost of capital" which is $\rho_k = \bar{X}_j/V_j$. The difference between ρ_k^r and the "true" average cost of capital, as we shall see, is a matter of some relevance in connection with investment planning within the firm (Section II). For the description of market behavior, however, which is our immediate concern here, the distinction is not essential. To simplify presentation, therefore, and to preserve continuity with the terminology in the standard literature we shall continue in this section to refer to ρ_k^r as the average cost of capital, though strictly speaking this identification is correct only in the absence of taxes.

Effects of a Plurality of Bonds and Interest Rates. In existing capital markets we find not one, but a whole family of interest rates varying with maturity, with the technical provisions of the loan and, what is most relevant for present purposes, with the financial condition of the borrower.¹⁶ Economic theory and market experience both suggest that the yields demanded by lenders tend to increase with the debt-equity ratio of the borrowing firm (or individual). If so, and if we can assume as a first approximation that this yield curve, $r=r(D/S)$, whatever its precise form, is the same for all borrowers, then we can readily extend our propositions to the case of a rising supply curve for borrowed funds.¹⁷

Proposition I is actually unaffected in form and interpretation by the fact that the rate of interest may rise with leverage; while the average cost of *borrowed* funds will tend to increase as debt rises, the average cost of funds from *all* sources will still be independent of leverage (apart from the tax effect). This conclusion follows directly from the ability of those who engage in arbitrage to undo the leverage in any financial structure by acquiring an appropriately mixed portfolio of bonds and stocks. Because of this ability, the ratio of earnings (*before* interest charges) to market value—*i.e.*, the average cost of capital from all

¹⁶ We shall not consider here the extension of the analysis to encompass the time structure of interest rates. Although some of the problems posed by the time structure can be handled within our comparative statics framework, an adequate discussion would require a separate paper.

¹⁷ We can also develop a theory of bond valuation along lines essentially parallel to those followed for the case of shares. We conjecture that the curve of bond yields as a function of leverage will turn out to be a nonlinear one in contrast to the linear function of leverage developed for common shares. However, we would also expect that the rate of increase in the yield on new issues would not be substantial in practice. This relatively slow rise would reflect the fact that interest rate increases by themselves can never be completely satisfactory to creditors as compensation for their increased risk. Such increases may simply serve to raise r so high relative to ρ that they become self-defeating by giving rise to a situation in which even normal fluctuations in earnings may force the company into bankruptcy. The difficulty of borrowing more, therefore, tends to show up in the usual case not so much in higher rates as in the form of increasingly stringent restrictions imposed on the company's management and finances by the creditors; and ultimately in a complete inability to obtain new borrowed funds, at least from the institutional investors who normally set the standards in the market for bonds.

sources—must be the same for all firms in a given class.¹⁸ In other words, the increased cost of borrowed funds as leverage increases will tend to be offset by a corresponding reduction in the yield of common stock. This seemingly paradoxical result will be examined more closely below in connection with Proposition II.

A significant modification of Proposition I would be required only if the yield curve $r=r(D/S)$ were different for different borrowers, as might happen if creditors had marked preferences for the securities of a particular class of debtors. If, for example, corporations as a class were able to borrow at lower rates than individuals having equivalent personal leverage, then the average cost of capital to corporations might fall slightly, as leverage increased over some range, in reflection of this differential. In evaluating this possibility, however, remember that the relevant interest rate for our arbitrage operators is the rate on brokers' loans and, historically, that rate has not been noticeably higher than representative corporate rates.¹⁹ The operations of holding companies and investment trusts which can borrow on terms comparable to operating companies represent still another force which could be expected to wipe out any marked or prolonged advantages from holding levered stocks.²⁰

Although Proposition I remains unaffected as long as the yield curve is the same for all borrowers, the relation between common stock yields and leverage will no longer be the strictly linear one given by the original Proposition II. If r increases with leverage, the yield i will still tend to

¹⁸ One normally minor qualification might be noted. Once we relax the assumption that all bonds have certain yields, our arbitrage operator faces the danger of something comparable to "gambler's ruin." That is, there is always the possibility that an otherwise sound concern—one whose long-run expected income is greater than its interest liability—might be forced into liquidation as a result of a run of temporary losses. Since reorganization generally involves costs, and because the operation of the firm may be hampered during the period of reorganization with lasting unfavorable effects on earnings prospects, we might perhaps expect heavily levered companies to sell at a slight discount relative to less heavily indebted companies of the same class.

¹⁹ Under normal conditions, moreover, a substantial part of the arbitrage process could be expected to take the form, not of having the arbitrage operators go into debt on personal account to put the required leverage into their portfolios, but simply of having them reduce the amount of corporate bonds they already hold when they acquire underpriced unlevered stock. Margin requirements are also somewhat less of an obstacle to maintaining any desired degree of leverage in a portfolio than might be thought at first glance. Leverage could be largely restored in the face of higher margin requirements by switching to stocks having more leverage at the corporate level.

²⁰ An extreme form of inequality between borrowing and lending rates occurs, of course, in the case of preferred stocks, which can not be directly issued by individuals on personal account. Here again, however, we would expect that the operations of investment corporations plus the ability of arbitrage operators to sell off their holdings of preferred stocks would act to prevent the emergence of any substantial premiums (for this reason) on capital structures containing preferred stocks. Nor are preferred stocks so far removed from bonds as to make it impossible for arbitrage operators to approximate closely the risk and leverage of a corporate preferred stock by incurring a somewhat smaller debt on personal account.

rise as D/S increases, but at a decreasing rather than a constant rate. Beyond some high level of leverage, depending on the exact form of the interest function, the yield may even start to fall.²¹ The relation between i and D/S could conceivably take the form indicated by the curve MD

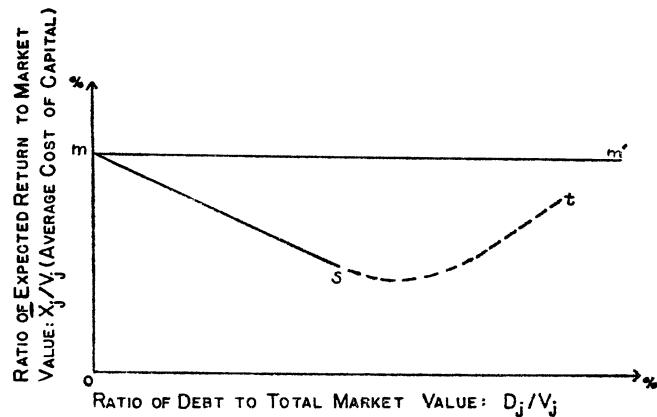


FIGURE 1

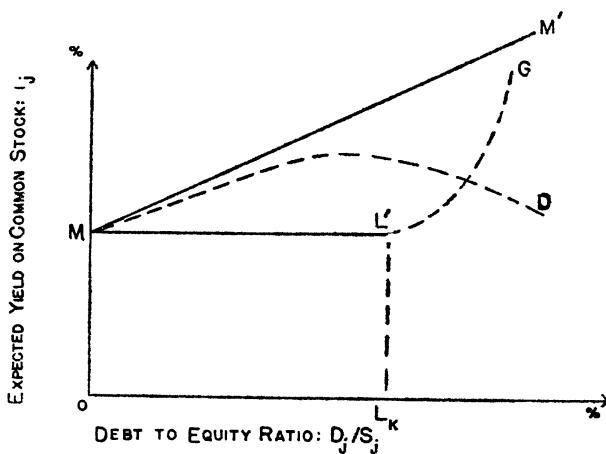


FIGURE 2

in Figure 2, although in practice the curvature would be much less pronounced. By contrast, with a constant rate of interest, the relation would be linear throughout as shown by line MM' , Figure 2.

The downward sloping part of the curve MD perhaps requires some

²¹ Since new lenders are unlikely to permit this much leverage (*cf.* note 17), this range of the curve is likely to be occupied by companies whose earnings prospects have fallen substantially since the time when their debts were issued.

comment since it may be hard to imagine why investors, other than those who like lotteries, would purchase stocks in this range. Remember, however, that the yield curve of Proposition II is a consequence of the more fundamental Proposition I. Should the demand by the risk-lovers prove insufficient to keep the market to the peculiar yield-curve MD , this demand would be reinforced by the action of arbitrage operators. The latter would find it profitable to own a pro-rata share of the firm as a whole by holding its stock *and* bonds, the lower yield of the shares being thus offset by the higher return on bonds.

D. The Relation of Propositions I and II to Current Doctrines

The propositions we have developed with respect to the valuation of firms and shares appear to be substantially at variance with current doctrines in the field of finance. The main differences between our view and the current view are summarized graphically in Figures 1 and 2. Our Proposition I [equation (4)] asserts that the average cost of capital, \bar{X}_j^r/V_j , is a constant for all firms j in class k , independently of their financial structure. This implies that, if we were to take a sample of firms in a given class, and if for each firm we were to plot the ratio of expected return to market value against some measure of leverage or financial structure, the points would tend to fall on a horizontal straight line with intercept ρ_k^r , like the solid line mm' in Figure 1.²² From Proposition I we derived Proposition II [equation (8)] which, taking the simplest version with r constant, asserts that, for all firms in a class, the relation between the yield on common stock and financial structure, measured by D_j/S_j , will approximate a straight line with slope $(\rho_k^r - r)$ and intercept ρ_k^r . This relationship is shown as the solid line MM' in Figure 2, to which reference has been made earlier.²³

By contrast, the conventional view among finance specialists appears to start from the proposition that, other things equal, the earnings-price ratio (or its reciprocal, the times-earnings multiplier) of a firm's common stock will normally be only slightly affected by "moderate" amounts of debt in the firm's capital structure.²⁴ Translated into our no-

²² In Figure 1 the measure of leverage used is D_j/V_j (the ratio of debt to market value) rather than D_j/S_j (the ratio of debt to equity), the concept used in the analytical development. The D_j/V_j measure is introduced at this point because it simplifies comparison and contrast of our view with the traditional position.

²³ The line MM' in Figure 2 has been drawn with a positive slope on the assumption that $\rho_k^r > r$, a condition which will normally obtain. Our Proposition II as given in equation (8) would continue to be valid, of course, even in the unlikely event that $\rho_k^r < r$, but the slope of MM' would be negative.

²⁴ See, e.g., Graham and Dodd [6, pp. 464-66]. Without doing violence to this position, we can bring out its implications more sharply by ignoring the qualification and treating the yield as a virtual constant over the relevant range. See in this connection the discussion in Durand [3, esp. pp. 225-37] of what he calls the "net income method" of valuation.

tation, it asserts that for any firm j in the class k ,

$$(13) \quad \frac{\bar{X}_{j^*} - rD_j}{S_j} \equiv \frac{\bar{\pi}_{j^*}}{S_j} = i_k^*, \text{ a constant for } \frac{D_j}{S_j} \leq L_k$$

or, equivalently,

$$(14) \quad S_j = \bar{\pi}_{j^*}/i_k^*.$$

Here i_k^* represents the capitalization rate or earnings-price ratio on the common stock and L_k denotes some amount of leverage regarded as the maximum "reasonable" amount for firms of the class k . This assumed relationship between yield and leverage is the horizontal solid line ML' of Figure 2. Beyond L' , the yield will presumably rise sharply as the market discounts "excessive" trading on the equity. This possibility of a rising range for high leverages is indicated by the broken-line segment $L'G$ in the figure.²⁵

If the value of shares were really given by (14) then the over-all market value of the firm must be:

$$(16) \quad V_j \equiv S_j + D_j = \frac{\bar{X}_{j^*} - rD_j}{i_k^*} + D_j = \frac{\bar{X}_{j^*}}{i_k^*} + \frac{(i_k^* - r)D_j}{i_k^*}.$$

That is, for any given level of expected total returns after taxes (\bar{X}_{j^*}) and assuming, as seems natural, that $i_k^* > r$, the value of the firm must tend to *rise* with debt;²⁶ whereas our Proposition I asserts that the value of the firm is completely independent of the capital structure. Another way of contrasting our position with the traditional one is in terms of the cost of capital. Solving (16) for \bar{X}_{j^*}/V_j yields:

$$(17) \quad \bar{X}_{j^*}/V_j = i_k^* - (i_k^* - r)D_j/V_j.$$

According to this equation, the average cost of capital is not independent of capital structure as we have argued, but should tend to *fall* with increasing leverage, at least within the relevant range of moderate debt ratios, as shown by the line ms in Figure 1. Or to put it in more familiar terms, debt-financing should be "cheaper" than equity-financing if not carried too far.

When we also allow for the possibility of a rising range of stock yields for large values of leverage, we obtain a U-shaped curve like mst in

²⁵ To make it easier to see some of the implications of this hypothesis as well as to prepare the ground for later statistical testing, it will be helpful to assume that the notion of a critical limit on leverage beyond which yields rise rapidly, can be epitomized by a quadratic relation of the form:

$$(15) \quad \bar{\pi}_{j^*}/S_j = i_k^* + \beta(D_j/S_j) + \alpha(D_j/S_j)^2, \quad \alpha > 0.$$

²⁶ For a typical discussion of how a promoter can, supposedly, increase the market value of a firm by recourse to debt issues, see W. J. Eiteman [4, esp. pp. 11-13].

Figure 1.²⁷ That a yield-curve for stocks of the form $ML'G$ in Figure 2 implies a U-shaped cost-of-capital curve has, of course, been recognized by many writers. A natural further step has been to suggest that the capital structure corresponding to the trough of the U is an "optimal capital structure" towards which management ought to strive in the best interests of the stockholders.²⁸ According to our model, by contrast, no such optimal structure exists—all structures being equivalent from the point of view of the cost of capital.

Although the falling, or at least U-shaped, cost-of-capital function is in one form or another the dominant view in the literature, the ultimate rationale of that view is by no means clear. The crucial element in the position—that the expected earnings-price ratio of the stock is largely unaffected by leverage up to some conventional limit—is rarely even regarded as something which requires explanation. It is usually simply taken for granted or it is merely asserted that this is the way the market behaves.²⁹ To the extent that the constant earnings-price ratio has a rationale at all we suspect that it reflects in most cases the feeling that moderate amounts of debt in "sound" corporations do not really add very much to the "riskiness" of the stock. Since the extra risk is slight, it seems natural to suppose that firms will not have to pay noticeably higher yields in order to induce investors to hold the stock.³⁰

A more sophisticated line of argument has been advanced by David Durand [3, pp. 231–33]. He suggests that because insurance companies and certain other important institutional investors are restricted to debt securities, nonfinancial corporations are able to borrow from them at interest rates which are lower than would be required to compensate

²⁷ The U-shaped nature of the cost-of-capital curve can be exhibited explicitly if the yield curve for shares as a function of leverage can be approximated by equation (15) of footnote 25. From that equation, multiplying both sides by S_i , we obtain: $\bar{\pi}_i^r = \bar{X}_i^r - rD_i = i_k^*S_i + \beta D_i + \alpha D_i^2 / S_i$; or, adding and subtracting $i_k^*D_k$ from the right-hand side and collecting terms,

$$(18) \quad \bar{X}_i^r = i_k^*(S_i + D_i) + (\beta + r - i_k^*)D_i + \alpha D_i^2 / S_i.$$

Dividing (18) by V_i gives an expression for the cost of capital:

$$(19) \quad \begin{aligned} \bar{X}_i^r / V_i &= i_k^* - (i_k^* - r - \beta)D_i / V_i + \alpha D_i^2 / S_i V_i = i_k^* - (i_k^* - r - \beta)D_i / V_i \\ &\quad + \alpha(D_i / V_i)^2 / (1 - D_i / V_i) \end{aligned}$$

which is clearly U-shaped since α is supposed to be positive.

²⁸ For a typical statement see S. M. Robbins [16, p. 307]. See also Graham and Dodd [6, pp. 468–74].

²⁹ See e.g., Graham and Dodd [6, p. 466].

³⁰ A typical statement is the following by Guthmann and Dougall [7, p. 245]: "Theoretically it might be argued that the increased hazard from using bonds and preferred stocks would counterbalance this additional income and so prevent the common stock from being more attractive than when it had a lower return but fewer prior obligations. In practice, the extra earnings from 'trading on the equity' are often regarded by investors as more than sufficient to serve as a 'premium for risk' when the proportions of the several securities are judiciously mixed."

creditors in a free market. Thus, while he would presumably agree with our conclusions that stockholders could not gain from leverage in an unconstrained market, he concludes that they can gain under present institutional arrangements. This gain would arise by virtue of the "safety superpremium" which lenders are willing to pay corporations for the privilege of lending.³¹

The defective link in both the traditional and the Durand version of the argument lies in the confusion between investors' subjective risk preferences and their objective market opportunities. Our Propositions I and II, as noted earlier, do not depend for their validity on any assumption about individual risk preferences. Nor do they involve any assertion as to what is an adequate compensation to investors for assuming a given degree of risk. They rely merely on the fact that a given commodity cannot consistently sell at more than one price in the market; or more precisely that the price of a commodity representing a "bundle" of two other commodities cannot be consistently different from the weighted average of the prices of the two components (the weights being equal to the proportion of the two commodities in the bundle).

An analogy may be helpful at this point. The relations between $1/\rho_k$, the price per dollar of an unlevered stream in class k ; $1/r$, the price per dollar of a sure stream, and $1/i_j$, the price per dollar of a levered stream j , in the k th class, are essentially the same as those between, respectively, the price of whole milk, the price of butter fat, and the price of milk which has been thinned out by skimming off some of the butter fat. Our Proposition I states that a firm cannot reduce the cost of capital—*i.e.*, increase the market value of the stream it generates—by securing part of its capital through the sale of bonds, even though debt money appears to be cheaper. This assertion is equivalent to the proposition that, under perfect markets, a dairy farmer cannot in general earn more for the milk he produces by skimming some of the butter fat and selling it separately, even though butter fat per unit weight, sells for more than whole milk. The advantage from skimming the milk rather than selling whole milk would be purely illusory; for what would be gained from selling the high-priced butter fat would be lost in selling the low-priced residue of thinned milk. Similarly our Proposition II—that the price per dollar of a levered stream falls as leverage increases—is an ex-

³¹ Like Durand, Morton [15] contends "that the actual market deviates from [Proposition I] by giving a changing over-all cost of money at different points of the [leverage] scale" (p. 443, note 2, inserts ours), but the basis for this contention is nowhere clearly stated. Judging by the great emphasis given to the lack of mobility of investment funds between stocks and bonds and to the psychological and institutional pressures toward debt portfolios (see pp. 444-51 and especially his discussion of the optimal capital structure on p. 453) he would seem to be taking a position very similar to that of Durand above.

act analogue of the statement that the price per gallon of thinned milk falls continuously as more butter fat is skimmed off.³²

It is clear that this last assertion is true as long as butter fat is worth more per unit weight than whole milk, and it holds even if, for many consumers, taking a little cream out of the milk (adding a little leverage to the stock) does not detract noticeably from the taste (does not add noticeably to the risk). Furthermore the argument remains valid even in the face of institutional limitations of the type envisaged by Durand. For suppose that a large fraction of the population habitually dines in restaurants which are required by law to serve only cream in lieu of milk (entrust their savings to institutional investors who can only buy bonds). To be sure the price of butter fat will then tend to be higher in relation to that of skinned milk than in the absence such restrictions (the rate of interest will tend to be lower), and this will benefit people who eat at home and who like skim milk (who manage their own portfolio and are able and willing to take risk). But it will still be the case that a farmer cannot gain by skimming some of the butter fat and selling it separately (firm cannot reduce the cost of capital by recourse to borrowed funds).³³

Our propositions can be regarded as the extension of the classical theory of markets to the particular case of the capital markets. Those who hold the current view—whether they realize it or not—must as-

³² Let M denote the quantity of whole milk, B/M the proportion of butter fat in the whole milk, and let p_M , p_B and p_α denote, respectively, the price per unit weight of whole milk, butter fat and thinned milk from which a fraction α of the butter fat has been skinned off. We then have the fundamental perfect market relation:

$$(a) \quad p_\alpha(M - \alpha B) + p_B \alpha B = p_M M, \quad 0 \leq \alpha \leq 1,$$

stating that total receipts will be the same amount $p_M M$, independently of the amount αB of butter fat that may have been sold separately. Since p_M corresponds to $1/\rho$, p_B to $1/r$, p_α to $1/i$, M to \bar{X} and αB to rD , (a) is equivalent to Proposition I, $S+D=\bar{X}/\rho$. From (a) we derive:

$$(b) \quad p_\alpha = p_M \frac{M}{M - \alpha B} - p_B \frac{\alpha B}{M - \alpha B}$$

which gives the price of thinned milk as an explicit function of the proportion of butter fat skinned off; the function decreasing as long as $p_B > p_M$. From (a) also follows:

$$(c) \quad 1/p_\alpha = 1/p_M + (1/p_M - 1/p_B) \frac{p_B \alpha B}{p_\alpha (M - \alpha B)}$$

which is the exact analogue of Proposition II, as given by (8).

³³ The reader who likes parables will find that the analogy with interrelated commodity markets can be pushed a good deal farther than we have done in the text. For instance, the effect of changes in the market rate of interest on the over-all cost of capital is the same as the effect of a change in the price of butter on the price of whole milk. Similarly, just as the relation between the prices of skim milk and butter fat influences the kind of cows that will be reared, so the relation between i and r influences the kind of ventures that will be undertaken. If people like butter we shall have Guernseys; if they are willing to pay a high price for safety, this will encourage ventures which promise smaller but less uncertain streams per dollar of physical assets.

sume not merely that there are lags and frictions in the equilibrating process—a feeling we certainly share,³⁴ claiming for our propositions only that they describe the central tendency around which observations will scatter—but also that there are large and *systematic* imperfections in the market which permanently bias the outcome. This is an assumption that economists, at any rate, will instinctively eye with some skepticism.

In any event, whether such prolonged, systematic departures from equilibrium really exist or whether our propositions are better descriptions of long-run market behavior can be settled only by empirical research. Before going on to the theory of investment it may be helpful, therefore, to look at the evidence.

E. Some Preliminary Evidence on the Basic Propositions

Unfortunately the evidence which has been assembled so far is amazingly skimpy. Indeed, we have been able to locate only two recent studies—and these of rather limited scope—which were designed to throw light on the issue. Pending the results of more comprehensive tests which we hope will soon be available, we shall review briefly such evidence as is provided by the two studies in question: (1) an analysis of the relation between security yields and financial structure for some 43 large electric utilities by F. B. Allen [1], and (2) a parallel (unpublished) study by Robert Smith [19], for 42 oil companies designed to test whether Allen's rather striking results would be found in an industry with very different characteristics.³⁵ The Allen study is based on average figures for the years 1947 and 1948, while the Smith study relates to the single year 1953.

The Effect of Leverage on the Cost of Capital. According to the received view, as shown in equation (17) the average cost of capital, \bar{X}^r/V , should decline linearly with leverage as measured by the ratio D/V , at least through most of the relevant range.³⁶ According to Proposition I, the average cost of capital within a given class k should tend to have the same value ρ_k^r independently of the degree of leverage. A simple test

³⁴ Several specific examples of the failure of the arbitrage mechanism can be found in Graham and Dodd [6, e.g., pp. 646–48]. The price discrepancy described on pp. 646–47 is particularly curious since it persists even today despite the fact that a whole generation of security analysts has been brought up on this book!

³⁵ We wish to express our thanks to both writers for making available to us some of their original worksheets. In addition to these recent studies there is a frequently cited (but apparently seldom read) study by the Federal Communications Commission in 1938 [22] which purports to show the existence of an optimal capital structure or range of structures (in the sense defined above) for public utilities in the 1930's. By current standards for statistical investigations, however, this study cannot be regarded as having any real evidential value for the problem at hand.

³⁶ We shall simplify our notation in this section by dropping the subscript j used to denote a particular firm wherever this will not lead to confusion.

of the merits of the two alternative hypotheses can thus be carried out by correlating \bar{X}'/V with D/V . If the traditional view is correct, the correlation should be significantly negative; if our view represents a better approximation to reality, then the correlation should not be significantly different from zero.

Both studies provide information about the average value of D —the market value of bonds and preferred stock—and of V —the market value of all securities.³⁷ From these data we can readily compute the ratio D/V and this ratio (expressed as a percentage) is represented by the symbol d in the regression equations below. The measurement of the variable \bar{X}'/V , however, presents serious difficulties. Strictly speaking, the numerator should measure the expected returns net of taxes, but this is a variable on which no direct information is available. As an approximation, we have followed both authors and used (1) the average value of actual net returns in 1947 and 1948 for Allen's utilities; and (2) actual net returns in 1953 for Smith's oil companies. Net return is defined in both cases as the sum of interest, preferred dividends and stockholders' income net of corporate income taxes. Although this approximation to expected returns is undoubtedly very crude, there is no reason to believe that it will systematically bias the test in so far as the sign of the regression coefficient is concerned. The roughness of the approximation, however, will tend to make for a wide scatter. Also contributing to the scatter is the crudeness of the industrial classification, since especially within the sample of oil companies, the assumption that all the firms belong to the same class in our sense, is at best only approximately valid.

Denoting by x our approximation to \bar{X}'/V (expressed, like d , as a percentage), the results of the tests are as follows:

$$\text{Electric Utilities } x = 5.3 + .006d \quad r = .12 \\ (\pm .008)$$

$$\text{Oil Companies } x = 8.5 + .006d \quad r = .04. \\ (\pm .024)$$

The data underlying these equations are also shown in scatter diagram form in Figures 3 and 4.

The results of these tests are clearly favorable to our hypothesis.

³⁷ Note that for purposes of this test preferred stocks, since they represent an *expected* fixed obligation, are properly classified with bonds even though the tax status of preferred dividends is different from that of interest payments and even though preferred dividends are really fixed only as to their maximum in any year. Some difficulty of classification does arise in the case of convertible preferred stocks (and convertible bonds) selling at a substantial premium, but fortunately very few such issues were involved for the companies included in the two studies. Smith included bank loans and certain other short-term obligations (at book values) in his data on oil company debts and this treatment is perhaps open to some question. However, the amounts involved were relatively small and check computations showed that their elimination would lead to only minor differences in the test results.

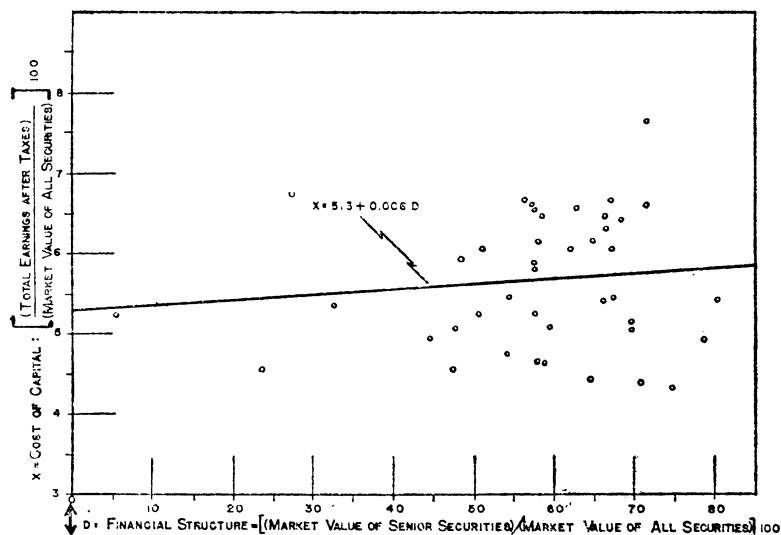


FIGURE 3. COST OF CAPITAL IN RELATION TO FINANCIAL STRUCTURE
FOR 43 ELECTRIC UTILITIES, 1947-48

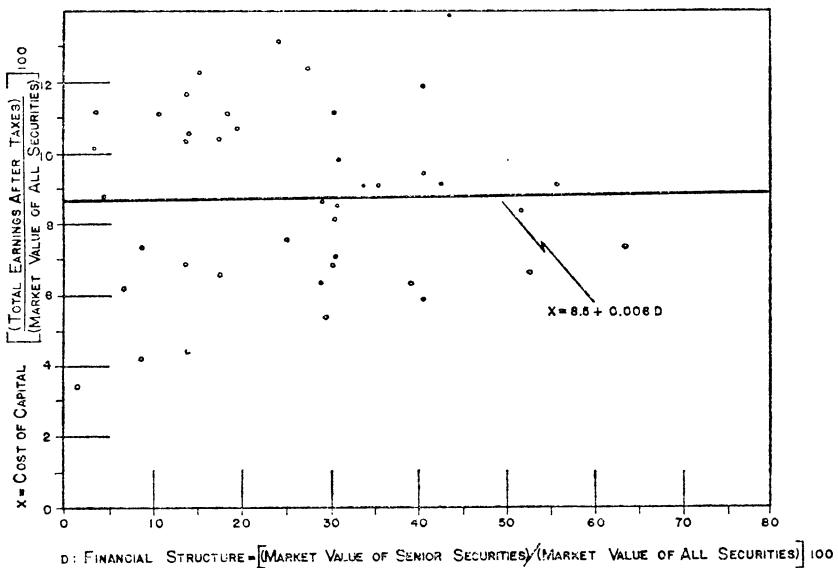


FIGURE 4. COST OF CAPITAL IN RELATION TO FINANCIAL STRUCTURE
FOR 42 OIL COMPANIES, 1953

Both correlation coefficients are very close to zero and not statistically significant. Furthermore, the implications of the traditional view fail to be supported even with respect to the sign of the correlation. The data in short provide no evidence of any tendency for the cost of capital to fall as the debt ratio increases.³⁸

It should also be apparent from the scatter diagrams that there is no hint of a curvilinear, U-shaped, relation of the kind which is widely believed to hold between the cost of capital and leverage. This graphical impression was confirmed by statistical tests which showed that for both industries the curvature was not significantly different from zero, its sign actually being opposite to that hypothesized.³⁹

Note also that according to our model, the constant terms of the regression equations are measures of ρ_k^r , the capitalization rates for unlevered streams and hence the average cost of capital in the classes in question. The estimates of 8.5 per cent for the oil companies as against 5.3 per cent for electric utilities appear to accord well with a priori expectations, both in absolute value and relative spread.

The Effect of Leverage on Common Stock Yields. According to our Proposition II—see equation 12 and Figure 2—the expected yield on common stock, $\bar{\pi}^r/S$, in any given class, should tend to increase with leverage as measured by the ratio D/S . The relation should tend to be linear and with positive slope through most of the relevant range (as in the curve MM' of Figure 2), though it might tend to flatten out if we move

³⁸ It may be argued that a test of the kind used is biased against the traditional view. The fact that both sides of the regression equation are divided by the variable V which may be subject to random variation might tend to impart a positive bias to the correlation. As a check on the results presented in the text, we have, therefore, carried out a supplementary test based on equation (16). This equation shows that, if the traditional view is correct, the market value of a company should, for given \bar{X}^r , increase with debt through most of the relevant range; according to our model the market value should be uncorrelated with D , given \bar{X}^r . Because of wide variations in the size of the firms included in our samples, all variables must be divided by a suitable scale factor in order to avoid spurious results in carrying out a test of equation (16). The factor we have used is the book value of the firm denoted by A . The hypothesis tested thus takes the specific form:

$$V/A = a + b(\bar{X}^r/A) + c(D/A)$$

and the numerator of the ratio X^r/A is again approximated by actual net returns. The partial correlation between V/A and D/A should now be positive according to the traditional view and zero according to our model. Although division by A should, if anything, bias the results in favor of the traditional hypothesis, the partial correlation turns out to be only .03 for the oil companies and $-.28$ for the electric utilities. Neither of these coefficients is significantly different from zero and the larger one even has the wrong sign.

³⁹ The tests consisted of fitting to the data the equation (19) of footnote 27. As shown there, it follows from the U-shaped hypothesis that the coefficient α of the variable $(D/V)^2/(1-D/V)$, denoted hereafter by d^* , should be significant and positive. The following regression equations and partials were obtained:

$$\text{Electric Utilities } x = 5.0 + .017d - .003d^*; r_{xd^*} = -.15$$

$$\text{Oil Companies } x = 8.0 + .05d - .03d^*; r_{xd^*} = -.14.$$

far enough to the right (as in the curve MD'), to the extent that high leverage tends to drive up the cost of senior capital. According to the conventional view, the yield curve as a function of leverage should be a horizontal straight line (like ML') through most of the relevant range; far enough to the right, the yield may tend to rise at an increasing rate. Here again, a straight-forward correlation—in this case between $\bar{\pi}'/S$ and D/S —can provide a test of the two positions. If our view is correct, the correlation should be significantly positive; if the traditional view is correct, the correlation should be negligible.

Subject to the same qualifications noted above in connection with \bar{X}' , we can approximate $\bar{\pi}'$ by actual stockholder net income.⁴⁰ Letting z denote in each case the approximation to $\bar{\pi}'/S$ (expressed as a percentage) and letting h denote the ratio D/S (also in percentage terms) the following results are obtained:

$$\text{Electric Utilities } z = 6.6 + .017h \quad r = .53 \\ (+.004)$$

$$\text{Oil Companies } z = 8.9 + .051h \quad r = .53. \\ (\pm .012)$$

These results are shown in scatter diagram form in Figures 5 and 6.

Here again the implications of our analysis seem to be borne out by the data. Both correlation coefficients are positive and highly significant when account is taken of the substantial sample size. Furthermore, the estimates of the coefficients of the equations seem to accord reasonably well with our hypothesis. According to equation (12) the constant term should be the value of $\rho_k r$ for the given class while the slope should be $(\rho_k r - r)$. From the test of Proposition I we have seen that for the oil companies the mean value of $\rho_k r$ could be estimated at around 8.7. Since the average yield of senior capital during the period covered was in the order of $3\frac{1}{2}$ per cent, we should expect a constant term of about 8.7 per cent and a slope of just over 5 per cent. These values closely approximate the regression estimates of 8.9 per cent and 5.1 per cent respectively. For the electric utilities, the yield of senior capital was also on the order of $3\frac{1}{2}$ per cent during the test years, but since the estimate of the mean value of $\rho_k r$ from the test of Proposition I was 5.6 per cent,

⁴⁰ As indicated earlier, Smith's data were for the single year 1953. Since the use of a single year's profits as a measure of expected profits might be open to objection we collected profit data for 1952 for the same companies and based the computation of $\bar{\pi}'/S$ on the average of the two years. The value of $\bar{\pi}'/S$ was obtained from the formula:

$$\left(\text{net earnings in 1952} \cdot \frac{\text{assets in '53}}{\text{assets in '52}} + \text{net earnings in '1953} \right) \frac{1}{2} \\ \div (\text{average market value of common stock in '53}).$$

The asset adjustment was introduced as rough allowance for the effects of possible growth in the size of the firm. It might be added that the correlation computed with $\bar{\pi}'/S$ based on net profits in 1953 alone was found to be only slightly smaller, namely .50.

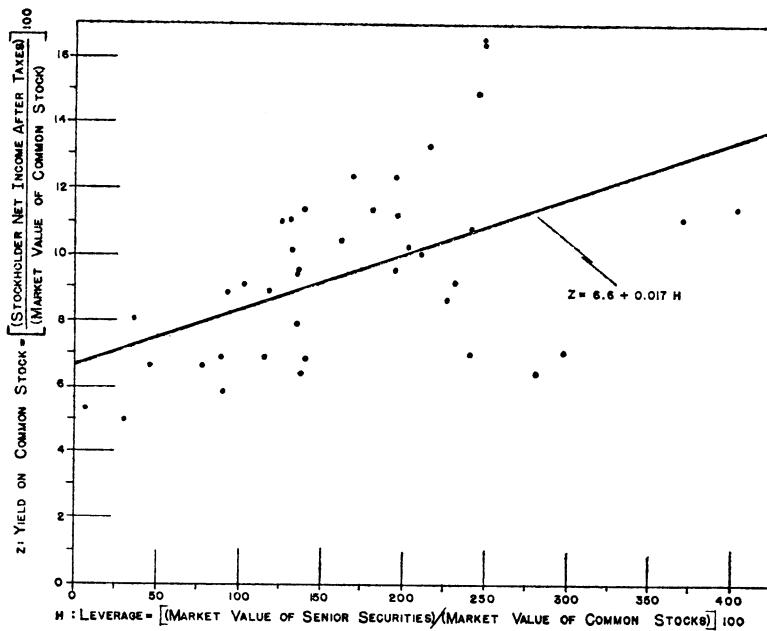


FIGURE 5. YIELD ON COMMON STOCK IN RELATION TO LEVERAGE FOR
43 ELECTRIC UTILITIES, 1947-48

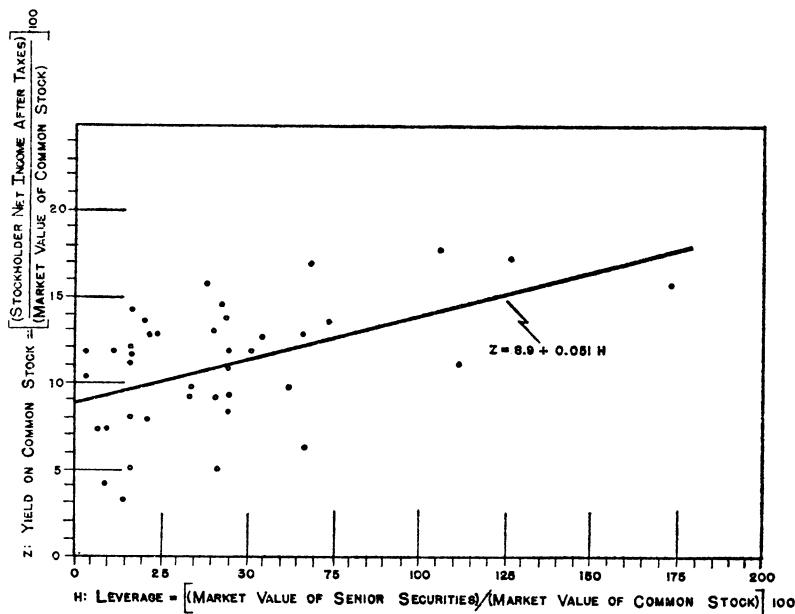


FIGURE 6. YIELD ON COMMON STOCK IN RELATION TO LEVERAGE FOR
42 OIL COMPANIES, 1952-53

the slope should be just above 2 per cent. The actual regression estimate for the slope of 1.7 per cent is thus somewhat low, but still within one standard error of its theoretical value. Because of this underestimate of the slope and because of the large mean value of leverage ($\bar{h}=160$ per cent) the regression estimate of the constant term, 6.6 per cent, is somewhat high, although not significantly different from the value of 5.6 per cent obtained in the test of Proposition I.

When we add a square term to the above equations to test for the presence and direction of curvature we obtain the following estimates:

$$\text{Electric Utilities } z = 4.6 + .004h - .007h^2$$

$$\text{Oil Companies } z = 8.5 + .072h - .016h^2.$$

For both cases the curvature is negative. In fact, for the electric utilities, where the observations cover a wider range of leverage ratios, the negative coefficient of the square term is actually significant at the 5 per cent level. Negative curvature, as we have seen, runs directly counter to the traditional hypothesis, whereas it can be readily accounted for by our model in terms of rising cost of borrowed funds.⁴¹

In summary, the empirical evidence we have reviewed seems to be broadly consistent with our model and largely inconsistent with traditional views. Needless to say much more extensive testing will be required before we can firmly conclude that our theory describes market behavior. Caution is indicated especially with regard to our test of Proposition II, partly because of possible statistical pitfalls⁴² and partly because not all the factors that might have a systematic effect on stock yields have been considered. In particular, no attempt was made to test the possible influence of the dividend pay-out ratio whose role has tended to receive a great deal of attention in current research and thinking. There are two reasons for this omission. First, our main objective has been to assess the *prima facie* tenability of *our* model, and in this model, based as it is on rational behavior by investors, dividends per se play no role. Second, in a world in which the policy of dividend stabilization is widespread, there is no simple way of disentangling the true effect of dividend payments on stock prices from their apparent effect,

⁴¹ That the yield of senior capital tended to rise for utilities as leverage increased is clearly shown in several of the scatter diagrams presented in the published version of Allen's study. This significant negative curvature between stock yields and leverage for utilities may be partly responsible for the fact, previously noted, that the constant in the linear regression is somewhat higher and the slope somewhat lower than implied by equation (12). Note also in connection with the estimate of ρ_k ⁷ that the introduction of the quadratic term reduces the constant considerably, pushing it in fact below the *a priori* expectation of 5.6, though the difference is again not statistically significant.

⁴² In our test, *e.g.*, the two variables z and h are both ratios with S appearing in the denominator, which may tend to impart a positive bias to the correlation (*cf.* note 38). Attempts were made to develop alternative tests, but although various possibilities were explored, we have so far been unable to find satisfactory alternatives.

the latter reflecting only the role of dividends as a proxy measure of long-term earning anticipations.⁴³ The difficulties just mentioned are further compounded by possible interrelations between dividend policy and leverage.⁴⁴

II. Implications of the Analysis for the Theory of Investment

A. Capital Structure and Investment Policy

On the basis of our propositions with respect to cost of capital and financial structure (and for the moment neglecting taxes), we can derive the following simple rule for optimal investment policy by the firm:

Proposition III. If a firm in class k is acting in the best interest of the stockholders at the time of the decision, it will exploit an investment opportunity if and only if the rate of return on the investment, say ρ^* , is as large as or larger than ρ_k . That is, *the cut-off point for investment in the firm will in all cases be ρ_k and will be completely unaffected by the type of security used to finance the investment.* Equivalently, we may say that regardless of the financing used, the marginal cost of capital to a firm is equal to the average cost of capital, which is in turn equal to the capitalization rate for an unlevered stream in the class to which the firm belongs.⁴⁵

To establish this result we will consider the three major financing alternatives open to the firm—bonds, retained earnings, and common stock issues—and show that in each case an investment is worth undertaking if, and only if, $\rho^* \geq \rho_k$.⁴⁶

Consider first the case of an investment financed by the sale of bonds. We know from Proposition I that the market value of the firm before the investment was undertaken was.⁴⁷

$$(20) \quad V_0 = \bar{X}_0 / \rho_k$$

⁴³ We suggest that failure to appreciate this difficulty is responsible for many fallacious, or at least unwarranted, conclusions about the role of dividends.

⁴⁴ In the sample of electric utilities, there is a substantial negative correlation between yields and pay-out ratios, but also between pay-out ratios and leverage, suggesting that either the association of yields and leverage or of yields and pay-out ratios may be (at least partly) spurious. These difficulties however do not arise in the case of the oil industry sample. A preliminary analysis indicates that there is here no significant relation between leverage and pay-out ratios and also no significant correlation (either gross or partial) between yields and pay-out ratios.

⁴⁵ The analysis developed in this paper is essentially a comparative-statics, not a dynamic analysis. This note of caution applies with special force to Proposition III. Such problems as those posed by expected changes in r and in ρ_k over time will not be treated here. Although they are in principle amenable to analysis within the general framework we have laid out, such an undertaking is sufficiently complex to deserve separate treatment. Cf. note 17.

⁴⁶ The extension of the proof to other types of financing, such as the sale of preferred stock or the issuance of stock rights is straightforward.

⁴⁷ Since no confusion is likely to arise, we have again, for simplicity, eliminated the subscripts identifying the firm in the equations to follow. Except for ρ_k , the subscripts now refer to time periods.

and that the value of the common stock was:

$$(21) \quad S_0 = V_0 - D_0.$$

If now the firm borrows I dollars to finance an investment yielding ρ^* its market value will become:

$$(22) \quad V_1 = \frac{\bar{X}_0 + \rho^* I}{\rho_k} = V_0 + \frac{\rho^* I}{\rho_k}$$

and the value of its common stock will be:

$$(23) \quad S_1 = V_1 - (D_0 + I) = V_0 + \frac{\rho^* I}{\rho_k} - D_0 - I$$

or using equation 21,

$$(24) \quad S_1 = S_0 + \frac{\rho^* I}{\rho_k} - I.$$

Hence $S_1 \geq S_0$ as $\rho^* \geq \rho_k$.⁴⁸

To illustrate, suppose the capitalization rate for uncertain streams in the k th class is 10 per cent and the rate of interest is 4 per cent. Then if a given company had an expected income of 1,000 and if it were financed entirely by common stock we know from Proposition I that the market value of its stock would be 10,000. Assume now that the managers of the firm discover an investment opportunity which will require an outlay of 100 and which is expected to yield 8 per cent. At first sight this might appear to be a profitable opportunity since the expected return is double the interest cost. If, however, the management borrows the necessary 100 at 4 per cent, the total expected income of the company rises to 1,008 and the market value of the firm to 10,080. But the firm now will have 100 of bonds in its capital structure so that, paradoxically, the market value of the stock must actually be reduced from 10,000 to 9,980 as a consequence of this apparently profitable investment. Or, to put it another way, the gains from being able to tap cheap, borrowed funds are more than offset for the stockholders by the market's discounting of the stock for the added leverage assumed.

Consider next the case of retained earnings. Suppose that in the course of its operations the firm acquired I dollars of cash (without impairing

⁴⁸ In the case of bond-financing the rate of interest on bonds does not enter explicitly into the decision (assuming the firm borrows at the market rate of interest). This is true, moreover, given the conditions outlined in Section I.C, even though interest rates may be an increasing function of debt outstanding. To the extent that the firm borrowed at a rate other than the market rate the two I 's in equation (24) would no longer be identical and an additional gain or loss, as the case might be, would accrue to the shareholders. It might also be noted in passing that permitting the two I 's in (24) to take on different values provides a simple method for introducing underwriting expenses into the analysis.

the earning power of its assets). If the cash is distributed as a dividend to the stockholders their wealth W_0 , after the distribution will be:

$$(25) \quad W_0 = S_0 + I = \frac{\bar{X}_0}{\rho_k} - D_0 + I$$

where \bar{X}_0 represents the expected return from the assets exclusive of the amount I in question. If however the funds are retained by the company and used to finance new assets whose expected rate of return is ρ^* , then the stockholders' wealth would become:

$$(26) \quad W_1 = S_1 = \frac{\bar{X}_0 + \rho^* I}{\rho_k} - D_0 = S_0 + \frac{\rho^* I}{\rho_k}.$$

Clearly $W_1 \geq W_0$ as $\rho^* \geq \rho_k$ so that an investment financed by retained earnings raises the net worth of the owners if and only if $\rho^* > \rho_k$.⁴⁹

Consider finally, the case of common-stock financing. Let P_0 denote the current market price per share of stock and assume, for simplicity, that this price reflects currently expected earnings only, that is, it does not reflect any future increase in earnings as a result of the investment under consideration.⁵⁰ Then if N is the original number of shares, the price per share is:

$$(27) \quad P_0 = S_0/N$$

and the number of new shares, M , needed to finance an investment of I dollars is given by:

$$(28) \quad M = \frac{I}{P_0}.$$

As a result of the investment the market value of the stock becomes:

$$S_1 = \frac{\bar{X}_0 + \rho^* I}{\rho_k} - D_0 = S_0 + \frac{\rho^* I}{\rho_k} = NP_0 + \frac{\rho^* I}{\rho_k}$$

and the price per share:

$$(29) \quad P_1 = \frac{S_1}{N + M} = \frac{1}{N + M} \left[NP_0 + \frac{\rho^* I}{\rho_k} \right].$$

⁴⁹ The conclusion that ρ_k is the cut-off point for investments financed from internal funds applies not only to undistributed net profits, but to depreciation allowances (and even to the funds represented by the current sale value of any asset or collection of assets). Since the owners can earn ρ_k by investing funds elsewhere in the class, partial or total liquidating distributions should be made whenever the firm cannot achieve a marginal internal rate of return equal to ρ_k .

⁵⁰ If we assumed that the market price of the stock did reflect the expected higher future earnings (as would be the case if our original set of assumptions above were strictly followed) the analysis would differ slightly in detail, but not in essentials. The cut-off point for new investment would still be ρ_k , but where $\rho^* > \rho_k$ the gain to the original owners would be larger than if the stock price were based on the pre-investment expectations only.

Since by equation (28), $I = MP_0$, we can add MP_0 and subtract I from the quantity in bracket, obtaining:

$$(30) \quad P_1 = \frac{1}{N + M} \left[(N + M)P_0 + \frac{\rho^* - \rho_k}{\rho_k} I \right]$$

$$= P_0 + \frac{1}{N + M} \frac{\rho^* - \rho_k}{\rho_k} I > P_0 \text{ if,}$$

and only if, $\rho^* > \rho_k$.

Thus an investment financed by common stock is advantageous to the current stockholders if and only if its yield exceeds the capitalization rate ρ_k .

Once again a numerical example may help to illustrate the result and make it clear why the relevant cut-off rate is ρ_k and not the current yield on common stock, i . Suppose that ρ_k is 10 per cent, r is 4 per cent, that the original expected income of our company is 1,000 and that management has the opportunity of investing 100 having an expected yield of 12 per cent. If the original capital structure is 50 per cent debt and 50 per cent equity, and 1,000 shares of stock are initially outstanding, then, by Proposition I, the market value of the common stock must be 5,000 or 5 per share. Furthermore, since the interest bill is $.04 \times 5,000 = 200$, the yield on common stock is $800/5,000 = 16$ per cent. It may then appear that financing the additional investment of 100 by issuing 20 shares to outsiders at 5 per share would dilute the equity of the original owners since the 100 promises to yield 12 per cent whereas the common stock is currently yielding 16 per cent. Actually, however, the income of the company would rise to 1,012; the value of the firm to 10,120; and the value of the common stock to 5,120. Since there are now 1,020 shares, each would be worth 5.02 and the wealth of the original stockholders would thus have been increased. What has happened is that the dilution in expected earnings per share (from .80 to .796) has been more than offset, in its effect upon the market price of the shares, by the decrease in leverage.

Our conclusion is, once again, at variance with conventional views,⁵¹ so much so as to be easily misinterpreted. Read hastily, Proposition III seems to imply that the capital structure of a firm is a matter of indifference; and that, consequently, one of the core problems of corporate finance—the problem of the optimal capital structure for a firm—is no problem at all. It may be helpful, therefore, to clear up such possible misunderstandings.

⁵¹ In the matter of investment policy under uncertainty there is no single position which represents "accepted" doctrine. For a sample of current formulations, all very different from ours, see Joel Dean [2, esp. Ch. 3], M. Gordon and E. Shapiro [5], and Harry Roberts [17].

B. Proposition III and Financial Planning by Firms

Misinterpretation of the scope of Proposition III can be avoided by remembering that this Proposition tells us only that the type of instrument used to finance an investment is irrelevant to the question of whether or not the investment is worth while. This does not mean that the owners (or the managers) have no grounds whatever for preferring one financing plan to another; or that there are no other policy or technical issues in finance at the level of the firm.

That grounds for preferring one type of financial structure to another will still exist within the framework of our model can readily be seen for the case of common-stock financing. In general, except for something like a widely publicized oil-strike, we would expect the market to place very heavy weight on current and recent past earnings in forming expectations as to future returns. Hence, if the owners of a firm discovered a major investment opportunity which they felt would yield much more than ρ_k , they might well prefer not to finance it via common stock at the then ruling price, because this price may fail to capitalize the new venture. A better course would be a pre-emptive issue of stock (and in this connection it should be remembered that stockholders are free to borrow and buy). Another possibility would be to finance the project initially with debt. Once the project had reflected itself in increased actual earnings, the debt could be retired either with an equity issue at much better prices or through retained earnings. Still another possibility along the same lines might be to combine the two steps by means of a convertible debenture or preferred stock, perhaps with a progressively declining conversion rate. Even such a double-stage financing plan may possibly be regarded as yielding too large a share to outsiders since the new stockholders are, in effect, being given an interest in any similar opportunities the firm may discover in the future. If there is a reasonable prospect that even larger opportunities may arise in the near future and if there is some danger that borrowing now would preclude more borrowing later, the owners might find their interests best protected by splitting off the current opportunity into a separate subsidiary with independent financing. Clearly the problems involved in making the crucial estimates and in planning the optimal financial strategy are by no means trivial, even though they should have no bearing on the basic decision to invest (as long as $\rho^* \geq \rho_k$).⁵²

Another reason why the alternatives in financial plans may not be a matter of indifference arises from the fact that managers are concerned

⁵² Nor can we rule out the possibility that the existing owners, if unable to use a financing plan which protects their interest, may actually prefer to pass up an otherwise profitable venture rather than give outsiders an "excessive" share of the business. It is presumably in situations of this kind that we could justifiably speak of a shortage of "equity capital," though this kind of market imperfection is likely to be of significance only for small or new firms.

with more than simply furthering the interest of the owners. Such other objectives of the management—which need not be necessarily in conflict with those of the owners—are much more likely to be served by some types of financing arrangements than others. In many forms of borrowing agreements, for example, creditors are able to stipulate terms which the current management may regard as infringing on its prerogatives or restricting its freedom to maneuver. The creditors might even be able to insist on having a direct voice in the formation of policy.⁵³ To the extent, therefore, that financial policies have these implications for the management of the firm, something like the utility approach described in the introductory section becomes relevant to financial (as opposed to investment) decision-making. It is, however, the utility functions of the managers *per se* and not of the owners that are now involved.⁵⁴

In summary, many of the specific considerations which bulk so large in traditional discussions of corporate finance can readily be superimposed on our simple framework without forcing any drastic (and certainly no systematic) alteration of the conclusion which is our principal concern, namely that for investment decisions, the marginal cost of capital is ρ_k .

C. The Effect of the Corporate Income Tax on Investment Decisions

In Section I it was shown that when an unintegrated corporate income tax is introduced, the original version of our Proposition I,

$$\bar{X}/V = \rho_k = \text{a constant}$$

must be rewritten as:

$$(11) \quad \frac{(\bar{X} - rD)(1 - \tau) + rD}{V} = \frac{\bar{X}'}{V} = \rho_k' = \text{a constant.}$$

Throughout Section I we found it convenient to refer to \bar{X}'/V as the cost of capital. The appropriate measure of the cost of capital relevant

⁵³ Similar considerations are involved in the matter of dividend policy. Even though the stockholders may be indifferent as to payout policy as long as investment policy is optimal, the management need not be so. Retained earnings involve far fewer threats to control than any of the alternative sources of funds and, of course, involve no underwriting expense or risk. But against these advantages management must balance the fact that sharp changes in dividend rates, which heavy reliance on retained earnings might imply, may give the impression that a firm's finances are being poorly managed, with consequent threats to the control and professional standing of the management.

⁵⁴ In principle, at least, this introduction of management's risk preferences with respect to financing methods would do much to reconcile the apparent conflict between Proposition III and such empirical findings as those of Modigliani and Zeman [14] on the close relation between interest rates and the ratio of new debt to new equity issues; or of John Lintner [12] on the considerable stability in target and actual dividend-payout ratios.

to investment decisions, however, is the ratio of the expected return *before* taxes to the market value, *i.e.*, \bar{X}/V . From (11) above we find:

$$(31) \quad \frac{\bar{X}}{V} = \frac{\rho_k^r - \tau r(D/V)}{1 - \tau} = \frac{\rho_k^r}{1 - \tau} \left[1 - \frac{\tau r D}{\rho_k^r V} \right],$$

which shows that the cost of capital now depends on the debt ratio, decreasing, as D/V rises, at the constant rate $\tau r/(1-\tau)$.⁵⁵ Thus, with a corporate income tax under which interest is a deductible expense, gains can accrue to stockholders from having debt in the capital structure, even when capital markets are perfect. The gains however are small, as can be seen from (31), and as will be shown more explicitly below.

From (31) we can develop the tax-adjusted counterpart of Proposition III by interpreting the term D/V in that equation as the proportion of debt used in any additional financing of V dollars. For example, in the case where the financing is entirely by new common stock, $D=0$ and the required rate of return ρ_k^S on a venture so financed becomes:

$$(32) \quad \rho_k^S = \frac{\rho_k^r}{1 - \tau} .$$

For the other extreme of pure debt financing $D=V$ and the required rate of return, ρ_k^D , becomes:

$$(33) \quad \rho_k^D = \frac{\rho_k^r}{1 - \tau} \left[1 - \tau \frac{r}{\rho_k^r} \right] = \rho_k^S \left[1 - \tau \frac{r}{\rho_k^r} \right] = \rho_k^S - \frac{\tau}{1 - \tau} r .^{56}$$

For investments financed out of retained earnings, the problem of defining the required rate of return is more difficult since it involves a comparison of the tax consequences to the individual stockholder of receiving a dividend versus having a capital gain. Depending on the time of realization, a capital gain produced by retained earnings may be taxed either at ordinary income tax rates, 50 per cent of these rates, 25 per

⁵⁵ Equation (31) is amenable, in principle, to statistical tests similar to those described in Section I.E. However we have not made any systematic attempt to carry out such tests so far, because neither the Allen nor the Smith study provides the required information. Actually, Smith's data included a very crude estimate of tax liability, and, using this estimate, we did in fact obtain a negative relation between \bar{X}/V and D/V . However, the correlation (-.28) turned out to be significant only at about the 10 per cent level. While this result is not conclusive, it should be remembered that, according to our theory, the slope of the regression equation should be in any event quite small. In fact, with a value of r in the order of .5, and values of ρ_k^r and r in the order of 8.5 and 3.5 per cent respectively (*cf.* Section I.E) an increase in D/V from 0 to 60 per cent (which is, approximately, the range of variation of this variable in the sample) should tend to reduce the average cost of capital only from about 17 to about 15 per cent.

⁵⁶ This conclusion does not extend to preferred stocks even though they have been classed with debt issues previously. Since preferred dividends except for a portion of those of public utilities are not in general deductible from the corporate tax, the cut-off point for new financing via preferred stock is exactly the same as that for common stock.

cent, or zero, if held till death. The rate on any dividends received in the event of a distribution will also be a variable depending on the amount of other income received by the stockholder, and with the added complications introduced by the current dividend-credit provisions. If we assume that the managers proceed on the basis of reasonable estimates as to the average values of the relevant tax rates for the owners, then the required return for retained earnings ρ_k^R can be shown to be:

$$(34) \quad \rho_k^R = \rho_k^r \frac{1}{1 - \tau} \frac{1 - \tau_d}{1 - \tau_g} = \frac{1 - \tau_d}{1 - \tau_g} \rho_k^r$$

where τ_d is the assumed rate of personal income tax on dividends and τ_g is the assumed rate of tax on capital gains.

A numerical illustration may perhaps be helpful in clarifying the relationship between these required rates of return. If we take the following round numbers as representative order-of-magnitude values under present conditions: an after-tax capitalization rate ρ_k^r of 10 per cent, a rate of interest on bonds of 4 per cent, a corporate tax rate of 50 per cent, a marginal personal income tax rate on dividends of 40 per cent (corresponding to an income of about \$25,000 on a joint return), and a capital gains rate of 20 per cent (one-half the marginal rate on dividends), then the required rates of return would be: (1) 20 per cent for investments financed entirely by issuance of new common shares; (2) 16 per cent for investments financed entirely by new debt; and (3) 15 per cent for investments financed wholly from internal funds.

These results would seem to have considerable significance for current discussions of the effect of the corporate income tax on financial policy and on investment. Although we cannot explore the implications of the results in any detail here, we should at least like to call attention to the remarkably small difference between the "cost" of equity funds and debt funds. With the numerical values assumed, equity money turned out to be only 25 per cent more expensive than debt money, rather than something on the order of 5 times as expensive as is commonly supposed to be the case.⁵⁷ The reason for the wide difference is that the traditional

⁵⁷ See e.g., D. T. Smith [18]. It should also be pointed out that our tax system acts in other ways to reduce the gains from debt financing. Heavy reliance on debt in the capital structure, for example, commits a company to paying out a substantial proportion of its income in the form of interest payments taxable to the owners under the personal income tax. A debt-free company, by contrast, can reinvest in the business all of its (smaller) net income and to this extent subject the owners only to the low capital gains rate (or possibly no tax at all by virtue of the loophole at death). Thus, we should expect a high degree of leverage to be of value to the owners, even in the case of closely held corporations, primarily in cases where their firm was not expected to have much need for additional funds to expand assets and earnings in the future. To the extent that opportunities for growth were available, as they presumably would be for most successful corporations, the interest of the stockholders would tend to be better served by a structure which permitted maximum use of retained earnings.

view starts from the position that debt funds are several times cheaper than equity funds even in the absence of taxes, with taxes serving simply to magnify the cost ratio in proportion to the corporate rate. By contrast, in our model in which the repercussions of debt financing on the value of shares are taken into account, the *only* difference in cost is that due to the tax effect, and its magnitude is simply the tax on the "grossed up" interest payment. Not only is this magnitude likely to be small but our analysis yields the further paradoxical implication that the stockholders' gain from, and hence incentive to use, debt financing is actually smaller the lower the rate of interest. In the extreme case where the firm could borrow for practically nothing, the advantage of debt financing would also be practically nothing.

III. Conclusion

With the development of Proposition III the main objectives we outlined in our introductory discussion have been reached. We have in our Propositions I and II at least the foundations of a theory of the valuation of firms and shares in a world of uncertainty. We have shown, moreover, how this theory can lead to an operational definition of the cost of capital and how that concept can be used in turn as a basis for rational investment decision-making within the firm. Needless to say, however, much remains to be done before the cost of capital can be put away on the shelf among the solved problems. Our approach has been that of static, partial equilibrium analysis. It has assumed among other things a state of atomistic competition in the capital markets and an ease of access to those markets which only a relatively small (though important) group of firms even come close to possessing. These and other drastic simplifications have been necessary in order to come to grips with the problem at all. Having served their purpose they can now be relaxed in the direction of greater realism and relevance, a task in which we hope others interested in this area will wish to share.

REFERENCES

1. F. B. ALLEN, "Does Going into Debt Lower the 'Cost of Capital'?", *Analysts Jour.*, Aug. 1954, 10, 57-61.
2. J. DEAN, *Capital Budgeting*. New York 1951.
3. D. DURAND, "Costs of Debt and Equity Funds for Business: Trends and Problems of Measurement" in Nat. Bur. Econ. Research, *Conference on Research in Business Finance*. New York 1952, pp. 215-47.
4. W. J. EITEMAN, "Financial Aspects of Promotion," in *Essays on Business Finance* by M. W. Waterford and W. J. Eiteman. Ann Arbor, Mich. 1952, pp. 1-17.
5. M. J. GORDON and E. SHAPIRO, "Capital Equipment Analysis: The Required Rate of Profit," *Manag. Sci.*, Oct. 1956, 3, 102-10.

6. B. GRAHAM and L. DODD, *Security Analysis*, 3rd ed. New York 1951.
7. G. GUTHMANN and H. E. DOUGALL, *Corporate Financial Policy*, 3rd ed. New York 1955.
8. J. R. HICKS, *Value and Capital*, 2nd ed. Oxford 1946.
9. P. HUNT and M. WILLIAMS, *Case Problems in Finance*, rev. ed. Homewood, Ill. 1954.
10. J. M. KEYNES, *The General Theory of Employment, Interest and Money*. New York 1936.
11. O. LANGE, *Price Flexibility and Employment*. Bloomington, Ind. 1944.
12. J. LINTNER, "Distribution of Incomes of Corporations among Dividends, Retained Earnings and Taxes," *Am. Econ. Rev.*, May 1956, 46, 97-113.
13. F. LUTZ and V. LUTZ, *The Theory of Investment of the Firm*. Princeton 1951.
14. F. MODIGLIANI and M. ZEMAN, "The Effect of the Availability of Funds, and the Terms Thereof, on Business Investment" in Nat. Bur. Econ. Research, *Conference on Research in Business Finance*. New York 1952, pp. 263-309.
15. W. A. MORTON, "The Structure of the Capital Market and the Price of Money," *Am. Econ. Rev.*, May 1954, 44, 440-54.
16. S. M. ROBBINS, *Managing Securities*. Boston 1954.
17. H. V. ROBERTS, "Current Problems in the Economics of Capital Budgeting," *Jour. Bus.*, 1957, 30 (1), 12-16.
18. D. T. SMITH, *Effects of Taxation on Corporate Financial Policy*. Boston 1952.
19. R. SMITH, "Cost of Capital in the Oil Industry," (hectograph). Pittsburgh: Carnegie Inst. Tech. 1955.
20. H. M. SOMERS, "'Cost of Money' as the Determinant of Public Utility Rates," *Buffalo Law Rev.*, Spring 1955, 4, 1-28.
21. J. B. WILLIAMS, *The Theory of Investment Value*. Cambridge, Mass. 1938.
22. U. S. Federal Communications Commission, *The Problem of the "Rate of Return" in Public Utility Regulation*. Washington 1938.

10. *United States v. E. I. duPont de Nemours and Co.*, 351 U.S. 377 (1956).
11. *United States v. E. I. duPont de Nemours and Co.*, 353 U.S. 586 (1957).
12. U. S. Senate Report No. 1775, 81st Cong., 2d Sess., 1950. *Amending Act to Supplement Existing Laws Against Unlawful Restraints and Monopolies.*

A Theory of Optimum Currency Areas

It is patently obvious that periodic balance-of-payments crises will remain an integral feature of the international economic system as long as fixed exchange rates and rigid wage and price levels prevent the terms of trade from fulfilling a natural role in the adjustment process. It is, however, far easier to pose the problem and to criticize the alternatives than it is to offer constructive and feasible suggestions for the elimination of what has become an international disequilibrium system.¹ The present paper, unfortunately, illustrates that proposition by cautioning against the practicability, in certain cases, of the most plausible alternative: a system of national currencies connected by flexible exchange rates.

A system of flexible exchange rates is usually presented, by its proponents,² as a device whereby depreciation can take the place of unemployment when the external balance is in deficit, and appreciation can replace inflation when it is in surplus. But the question then arises whether all existing national currencies should be flexible. Should the Ghanaian pound be freed to fluctuate against all currencies or ought the present sterling-area currencies remain pegged to the pound sterling? Or, supposing that the Common Market countries proceed with their plans for economic union, should these countries allow each national currency to fluctuate, or would a single currency area be preferable?

The problem can be posed in a general and more revealing way by defining a currency area as a domain within which exchange rates are fixed and asking: What is the appropriate domain of a currency area? It might seem at first that the question is purely academic since it hardly appears within the realm of political feasibility that national currencies would ever be abandoned in favor of any other arrangement. To this, three answers can be given: (1) Certain parts of the world are undergoing processes of economic integration and disintegration, new experiments are being made, and a conception of what constitutes an optimum currency area can clarify the meaning of these experiments. (2) Those countries, like Canada, which have experimented with flexible exchange rates are likely to face particular problems which the theory of *optimum* currency areas can elucidate if the national currency area does not coincide with the optimum currency area. (3) The idea can be used to illustrate certain functions of currencies which have been inadequately treated in the economic literature and which are sometimes neglected in the consideration of problems of economic policy.

¹ I have analyzed this system in some detail in [7].

² See, for example [1] [3] [5].

I. *Currency Areas and Common Currencies*

A single currency implies a single central bank (with note-issuing powers) and therefore a potentially elastic supply of interregional means of payments. But in a currency area comprising more than one currency the supply of international means of payment is conditional upon the cooperation of many central banks; no central bank can expand its own liabilities much faster than other central banks without losing reserves and impairing convertibility.³ This means that there will be a major difference between adjustment within a currency area which has a single currency and a currency area involving more than one currency; in other words there will be a difference between interregional adjustment and international adjustment even though exchange rates, in the latter case, are fixed.

To illustrate this difference consider a simple model of two entities (regions or countries), initially in full employment and balance-of-payments equilibrium, and see what happens when this equilibrium is disturbed by a shift of demand from the goods of entity B to the goods of entity A. Assume that money wages and prices cannot be reduced in the short run without causing unemployment, and that monetary authorities act to prevent inflation.

Suppose first that the entities are countries with national currencies. The shift of demand from B to A causes unemployment in B and inflationary pressure in A.⁴ To the extent that prices are allowed to rise in A the change in the terms of trade will relieve B of some of the burden of adjustment. But if A tightens credit restrictions to prevent prices from rising all the burden of adjustment is thrust onto country B; what is needed is a reduction in B's real income and if this cannot be effected by a change in the terms of trade—because B cannot lower, and A will not raise, prices—it must be accomplished by a decline in B's output and employment. The policy of surplus countries in restraining prices therefore imparts a recessive tendency to the world economy on fixed exchange rates or (more generally) to a currency area with many separate currencies.⁵

Contrast this situation with that where the entities are regions within a closed economy lubricated by a common currency; and suppose now that the national government pursues a full-employment policy. The shift of demand from B to A causes unemployment in region B and inflationary pressure in region A, and a surplus in A's balance of payments.⁶ To correct the unemployment in B the monetary authorities increase the money supply. The monetary expansion, however, aggravates inflationary pressure in region A: indeed, the

³ More exactly, the rates at which central banks can expand monetary liabilities depend on income elasticities of demand and output elasticities of supply.

⁴ For present purposes inflation is defined as a rise in the prices of home-produced goods.

⁵ The tendency of surplus countries to control (what is, from a national point of view) inflation can be amply documented from United States and French policy in the 1920's and West Germany policy today. But it is unfortunate that a simple change in world relative prices is interpreted, in the surplus countries, as inflation.

⁶ Instructive examples of balance-of-payments problems between different regions of the United States can be found in [2, Ch. 14] For purposes of this paper regions are defined as areas within which there is factor mobility, but between which there is factor immobility.

principal way in which the monetary policy is effective in correcting full employment in the deficit region is by raising prices in the surplus region, turning the terms of trade against B. Full employment thus imparts an inflationary bias to the multiregional economy or (more generally) to a currency area with common currency.

In a currency area comprising different countries with national currencies the pace of employment in deficit countries is set by the willingness of surplus countries to inflate. But in a currency area comprising many regions and a single currency, the pace of inflation is set by the willingness of central authorities to allow unemployment in deficit regions.

The two systems could be brought closer together by an institutional change: unemployment could be avoided in the world economy if central banks agreed that the burden of international adjustment should fall on surplus countries, which would then inflate until unemployment in deficit countries is eliminated; or a world central bank could be established with power to create an international means of payment. But a currency area of either type cannot prevent both unemployment and inflation among its members. The fault lies not with the type of currency area, but with the domain of the currency area. The optimum currency area is not the world.

II. National Currencies and Flexible Exchange Rates

The existence of more than one currency area in the world implies (by definition) variable exchange rates. In the international trade example, if demand shifts from the products of country B to the products of country A, a depreciation by country B or an appreciation by country A would correct the external imbalance and also relieve unemployment in country B and restrain inflation in country A. This is the most favorable case for flexible rates based on national currencies.

Other examples, however, might be equally relevant. Suppose that the world consists of two countries, Canada and the United States, each of which has separate currencies. Also assume that the continent is divided into two regions which do not correspond to national boundaries—the East, which produces goods like cars, and the West, which produces goods like lumber products. To test the flexible-exchange-rate-argument in this example assume that the United States dollar fluctuates relative to the Canadian dollar, and that an increase in productivity (say) in the automobile industry causes an excess demand for lumber products and an excess supply of cars.

The immediate impact of the shift in demand is to cause unemployment in the East and inflationary pressure in the West, and a flow of bank reserves from the East to the West because of the former's regional balance-of-payments deficit. To relieve the unemployment in the East the central banks in both countries would have to expand the national money supplies, or to prevent inflation in the West, contract the national money supplies. (Meanwhile the Canada-United States exchange rate would move to preserve equilibrium in the national balances.) Thus, unemployment can be prevented in both countries, but only at the expense of inflation; or, inflation can be re-

strained in both countries but at the expense of unemployment; or, finally, the burden of adjustment can be shared between East and West with some unemployment in the East and some inflation in the West. But both unemployment and inflation cannot be escaped. The flexible exchange rate system does not serve to correct the balance-of-payments situation between the two regions (which is the essential problem) although it will do so between the two countries; it is therefore not necessarily preferable to a common currency or national currencies connected by fixed exchange rates.

III. *Regional Currency Areas and Flexible Exchange Rates*

The preceding example does not destroy the argument for flexible exchange rates, but it might severely impair the relevance of the argument if it is applied to national currencies. The logic of the argument can in fact be rescued if national currencies are abandoned in favor of regional currencies.

To see this suppose that the "world" reorganizes currencies so that Eastern and Western dollars replace Canadian and United States dollars. Now if the exchange rate between the East and the West were pegged, a dilemma would arise similar to that discussed in the first section. But if the East-West exchange rate were flexible, then an excess demand for lumber products need cause neither inflation nor unemployment in either region. The Western dollar appreciates relative to the Eastern dollar thus assuring balance-of-payments equilibrium, while the Eastern and Western central banks adopt monetary policies to ensure constancy of effective demand in terms of the regional currencies, and therefore stable prices and employment.

The same argument could be approached from another direction. A system of flexible exchange rates was originally propounded as an alternative to the gold-standard mechanism which many economists blamed for the world-wide spread of depression after 1929. But if the arguments against the gold standard were correct, then why should a similar argument not apply against a common currency system in a multiregional country? Under the gold standard depression in one country would be transmitted, through the foreign-trade multiplier, to foreign countries. Similarly, under a common currency, depression in one region would be transmitted to other regions for precisely the same reasons. If the gold standard imposed a harsh discipline on the national economy and induced the transmission of economic fluctuations, then a common currency would be guilty of the same charges; interregional balance-of-payments problems are invisible, so to speak, precisely because there is no escape from the self-adjusting effects of interregional money flows. (It is true, of course, that interregional liquidity can always be supplied by the national central bank, whereas the gold standard and even the gold-exchange standard were hampered, on occasion, by periodic scarcities of internationally liquid assets; but the basic argument against the gold standard was essentially distinct from the liquidity problem.)

Today, if the case for flexible exchange rates is a strong one, it is, in logic, a case for flexible exchange rates based on *regional* currencies, not on national currencies. The optimum currency area is the region.

IV. A Practical Application

The theory of international trade was developed on the Ricardian assumption that factors of production are mobile internally but immobile internationally. Williams, Ohlin, Iversen and others, however, protested that this assumption was invalid and showed how its relaxation would affect the real theory of trade. I have tried to show that its relaxation has important consequences also for the monetary theory of trade and especially the theory of flexible exchange rates. The argument for flexible exchange rates based on national currencies is only as valid as the Ricardian assumption about factor mobility. If factor mobility is high internally and low internationally a system of flexible exchange rates based on national currencies might work effectively enough. But if regions cut across national boundaries or if countries are multiregional then the argument for flexible exchange rates is only valid if currencies are reorganized on a regional basis.

In the real world, of course, currencies are mainly an expression of national sovereignty, so that actual currency reorganization would be feasible only if it were accompanied by profound political changes. The concept of an optimum currency area therefore has direct practical applicability only in areas where political organization is in a state of flux, such as in ex-colonial areas and in Western Europe.

In Western Europe the creation of the Common Market is regarded by many as an important step toward eventual political union, and the subject of a common currency for the six countries has been much discussed. One can cite the well-known position of J. E. Meade [4, pp. 385-86], who argues that the conditions for a common currency in Western Europe do not exist, and that, especially because of the lack of labor mobility, a system of flexible exchange rates would be more effective in promoting balance-of-payments equilibrium and internal stability; and the apparently opposite view of Tibor Scitovsky [9, Ch. 2]⁷ who favors a common currency because he believes that it would induce a greater degree of capital mobility, but further adds that steps must be taken to make labor more mobile and to facilitate supranational employment policies. In terms of the language of this paper Meade favors national currency areas while Scitovsky gives qualified approval to the idea of a single currency area in Western Europe.

In spite of the apparent contradiction between these two views, the concept of optimum currency areas helps us to see that the conflict reduces to an empirical rather than a theoretical question. In both cases it is implied that an essential ingredient of a common currency, or a single currency area, is a high degree of factor mobility; but Meade believes that the necessary factor mobility does not exist, while Scitovsky argues that labor mobility must be improved and that the creation of a common currency would itself stimulate capital mobility. In other words neither writer disputes that the optimum currency area is the region—defined in terms of internal factor mobility and external factor immobility—but there is an implicit difference in views on the

⁷ These statements of course cannot do full justice to the arguments of Meade and Scitovsky.

precise degree of factor mobility required to delineate a region. The question thus reduces to whether or not Western Europe can be considered a single region, and this is essentially an empirical problem.

V. Upper Limits on the Number of Currencies and Currency Areas

A dilemma now arises: Factor mobility (and hence the delineation of regions) is most usefully considered a relative rather than an absolute concept, with both geographical and industrial dimensions, and it is likely to change over time with alterations in political and economic conditions. If, then, the goals of internal stability are to be rigidly pursued, it follows that the greater is the number of separate currency areas in the world, the more successfully will these goals be attained (assuming, as always, that the basic argument for flexible exchange rates *per se* is valid). But this seems to imply that regions ought to be defined so narrowly as to count every minor pocket of unemployment arising from labor immobility as a separate region, each of which should apparently have a separate currency!

Such an arrangement hardly appeals to common sense. The suggestion reflects the fact that we have, thus far, considered the reasons for keeping currency areas small, not the reasons for maintaining or increasing their size. In other words we have discussed only the stabilization argument, to which end it is preferable to have many currency areas, and not the increasing costs which are likely to be associated with the maintenance of many currency areas.

It will be recalled that the older economists of the nineteenth century were internationalists and generally favored a world currency. Thus, John Stuart Mill wrote [6, p. 176]:

. . . So much of barbarism, however, still remains in the transactions of most civilised nations, that almost all independent countries choose to assert their nationality by having, to their own inconvenience and that of their neighbours, a peculiar currency of their own.

Mill, like Bagehot and others, was concerned with the costs of valuation and money-changing, not stabilization policy, and it is readily seen that these costs tend to increase with the number of currencies. Any given money *qua* numeraire or unit of account fulfills this function less adequately if the prices of foreign goods are expressed in terms of foreign currency and must then be translated into domestic currency prices. Similarly, money in its role of medium of exchange is less useful if there are many currencies; although the costs of currency conversion are always present, they loom exceptionally large under inconvertibility or flexible exchange rates. (Indeed, in a hypothetical world in which the number of currencies equaled the number of commodities, the usefulness of money in its roles of unit of account and medium of exchange would disappear, and trade might just as well be conducted in terms of pure barter.) Money is a convenience and this restricts the optimum number of currencies. In terms of this argument alone the optimum currency area is the world, regardless of the number of regions of which it is composed.

There are two other factors which would inhibit the creation of an arbitrarily large number of currency areas. In the first place markets for foreign

exchange must not be so thin that any single speculator (perhaps excepting central banks) can affect the market price; otherwise the speculation argument against flexible exchange rates would assume weighty dimensions. The other argument limiting "Balkanization" concerns the very pillar on which the flexible exchange-rate argument rests. The thesis of those who favor flexible exchange rates is that the community in question is not willing to accept variations in its real income through adjustments in its money wage rate or price level, but that it is willing to accept virtually the same changes in its real income through variations in the rate of exchange. In other words it is assumed that unions bargain for a money rather than a real wage, and adjust their wage demands to changes in the cost of living, if at all, only if the cost-of-living index excludes imports. Now as the currency area grows smaller and the proportion of imports in total consumption grows, this assumption becomes increasingly unlikely. It may not be implausible to suppose that there is some degree of money illusion in the bargaining process between unions and management (or frictions and lags having the same effects), but it is unrealistic to assume the extreme degree of money illusion that would have to exist in small currency areas. Since the necessary degree of money illusion becomes greater the smaller are currency areas, it is plausible to conclude that this also imposes an upper limit on the number of currency areas.

VI. Concluding Argument

The subject of flexible exchange rates can logically be separated into two distinct questions. The first is whether a system of flexible exchange rates can work effectively and efficiently in the modern world economy. For this to be possible it must be demonstrated that: (1) an international price system based on flexible exchange rates is dynamically stable after taking speculative demands into account; (2) the exchange rate changes necessary to eliminate normal disturbances to dynamic equilibrium are not so large as to cause violent and reversible shifts between export and import-competing industries (this is not ruled out by stability); (3) the risks created by variable exchange rates can be covered at reasonable costs in the forward markets; (4) central banks will refrain from monopolistic speculation; (5) monetary discipline will be maintained by the unfavorable political consequences of continuing depreciation, as it is to some extent maintained today by threats to the levels of foreign exchange reserves; (6) reasonable protection of debtors and creditors can be assured to maintain an increasing flow of long-term capital movements; and (7) wages and profits are not tied to a price index in which import goods are heavily weighted. I have not explicitly discussed these issues in my paper.

The second question concerns how the world should be divided into currency areas. I have argued that the stabilization argument for flexible exchange rates is valid only if it is based on regional currency areas. If the world can be divided into regions within each of which there is factor mobility and between which there is factor immobility, then each of these regions should have a separate currency which fluctuates relative to all other currencies. This carries the argument for flexible exchange rates to its logical conclusion.

But a region is an economic unit while a currency domain is partly an ex-

pression of national sovereignty. Except in areas where national sovereignty is being given up it is not feasible to suggest that currencies should be reorganized; the validity of the argument for flexible exchange rates therefore hinges on the closeness with which nations correspond to regions. The argument works best if each nation (and currency) has internal factor mobility and external factor immobility. But if labor and capital are insufficiently mobile within a country then flexibility of the external price of the national currency cannot be expected to perform the stabilization function attributed to it, and one could expect varying rates of unemployment or inflation in the different regions. Similarly, if factors are mobile across national boundaries then a flexible exchange system becomes unnecessary, and may even be positively harmful, as I have suggested elsewhere.⁸

Canada provides the only modern example where an advanced country has experimented with flexible exchange rates. According to my argument the experiment should be largely unsuccessful as far as stabilization is concerned. Because of the factor immobility between regions an increase in foreign demand for the products of one of the regions would cause an appreciation of the exchange rate and therefore increased unemployment in the remaining regions, a process which could be corrected by a monetary policy which aggravated inflationary pressures in the first region; every change in demand for the products in one region is likely to induce opposite changes in other regions which can not be entirely modified by national stabilization policies. Similarly the high degree of external capital mobility is likely to interfere with stabilization policy for completely different reasons: to achieve internal stability the central bank can alter credit conditions but it is the change in the exchange rate rather than the alteration in the interest rate which produces the stabilizing effect; this indirectness conduces to a cyclical approach to equilibrium. Although an explicit empirical study would be necessary to verify that the Canadian experiment has not fulfilled the claims made for flexible exchange rates, the *prima facie* evidence indicates that it has not. It must be emphasized, though, that a failure of the Canadian experiment would cast doubt only on the effectiveness of a flexible exchange system in a multiregional country, not on a flexible exchange system in a unitary country.⁹

ROBERT A. MUNDELL*

* In my paper, "The Monetary Dynamics of International Adjustment Under Fixed and Flexible Exchange Rates," [8], I advanced the argument that stabilization policy would be more difficult under fixed exchange rates if short-term capital were immobile than if it were mobile, and more difficult under flexible exchange rates if capital were mobile than if it were immobile. Although the method of analysis was fundamentally different the conclusions support the hypothesis of this paper that the fixed-exchange-rate system is better within areas where factors are mobile and the flexible-exchange-rate system is better for areas between which factors are immobile. The argument of my other paper imposes an additional argument against increasing the number of currencies.

* Other economists have advanced arguments in favor of Balkanization of multiregional countries (see for example, A. D. Scott [10]); and the argument for regional currency areas adds to the list; but, as Scott is careful to emphasize, no country can make such decisions on purely economic grounds.

* The author is an economist in the Special Research Section of the International Monetary Fund.

REFERENCES

1. MILTON FRIEDMAN, "The Case for Flexible Exchange Rates," *Essays in Positive Economics*. Chicago 1953.
2. S. E. HARRIS, *Interregional and International Economics*. New York 1957.
3. F. L. LUTZ, "The Case for Flexible Exchange Rates," *Banca Naz. del Lavoro*, Dec. 1954.
4. J. E. MEADE, "The Balance of Payments Problems of a Free Trade Area," *Econ. Jour.*, Sept. 1957, 67, 379-96.
5. ——, "The Case for Variable Exchange Rates," *Three Banks Rev.*, Sept. 1955.
6. J. S. MILL, *Principles of Political Economy*, Vol. II. New York 1894.
7. R. A. MUNDELL, "The International Disequilibrium System," *Kyklos*, 1961 (2), 14, 153-72.
8. ——, "The Monetary Dynamics of International Adjustment under Fixed and Flexible Exchange Rates," *Quart. Jour. Econ.*, May 1960, 74, 227-57.
9. TIBOR SCITOVSKY, *Economic Theory and Western European Integration*. Stanford 1958.
10. A. D. SCOTT, "A Note on Grants in Federal Countries," *Economica*, Nov. 1950, 17 (N.S.), 416-22.

Institutional Affiliation of the Contributors to Three Professional Journals

Two analyses of the institutional affiliations of authors of economics papers have appeared in past issues of the *American Economic Review*. One made some years ago referred to the origins of the participants (excluding the discussants) in the programs of the American Economic Association meetings for the five-year period 1950-54 [2]. A more recent one analyzed the affiliations of the contributors to the AER (apart from the Proceedings) for the decade 1950-59 [1]. Some interesting observations can be made by comparing the two studies. The omission of some schools in both cases is rather surprising. Fusfeld's list of the institutions who contributed two or more papers to the American Economic Association meetings over the five-year period did not include schools like Princeton University or Johns Hopkins University. Cleary and Edwards' list of institutions who aggregated 100 or more pages in the AER over the ten-year period did not mention schools like Columbia University or Duke University. With regard to the ranking of the institutions according to the volume of their aggregate contribution, the results are equally striking. In the case of the Proceedings, Harvard University, the University of California, the University of Chicago, and Columbia University top the list. In the case of the regular quarterly issues of the AER, the University of California, the Massachusetts Institute of Technology, Stanford University, and the University of Chicago ranked as the most important originators. From the four institutions that did not repeat in the top four, Columbia was conspicuously absent from the Cleary and Edwards' list while Harvard University

THE AMERICAN ECONOMIC REVIEW

VOLUME LIII

DECEMBER 1963

NUMBER 5

UNCERTAINTY AND THE WELFARE ECONOMICS OF MEDICAL CARE

By KENNETH J. ARROW*

I. *Introduction: Scope and Method*

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 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 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 flows of services that would be

* The author is professor of economics at Stanford University. He wishes to express his thanks for useful comments to F. Bator, R. Dorfman, V. Fuchs, Dr. 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.

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 the going prices were such that the amounts of services which were available equalled 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 (due to V. Pareto): There is no other allocation of resources to services which will make all participants in the market better off.

Both the conditions of this optimality theorem and the definition of optimality 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 proposition, but it is a very weak one. From this it follows that it is not desirable to put up with a non-optimal 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 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

¹ This point has been stressed by I. M. D. Little [19, pp. 71-74]. For the concept of a "persuasive definition," see C. L. Stevenson [27, pp. 210-17].

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 *ex post* judge the distribution of satisfactions 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

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 [18]. The best summary of the various ways in which the theorems can fail to hold is probably Bator's [6].

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. Recently, M. Friedman [15] 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 paper, 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, insure that competitive equilibrium is necessarily optimal; the third insures 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, i.e., imply that there will be some set of prices which will clear all markets.⁵

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 not only risks 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.

⁴ There are further minor conditions, for which see Koopmans [18, pp. 50-55].

⁵ For a more precise statement of the existence conditions, see Koopmans [18, pp. 56-60] or Debreu [12, Ch. 5].

Either system would lead to an optimal state, though 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, though it is obviously of importance for policy; for the present 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 in Section IV.C 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," i.e., desired protection against many risks, are

⁶The theory, in variant forms, seems to have been first worked out by Allais [2], Arrow [5], and Baudier [7]. For further generalization, see Debreu [11] and [12, Ch. 7].

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 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.

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.

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 [21] and Arrow [4].

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 an optimal state in the sense of Pareto. 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 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 due to 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 towards optimality thus puts obstacles in its own path.

The doctrine that society will seek to achieve optimality by non-market 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 unpre-

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 [23, pp. 581-82] (I am indebted for this reference to V. Fuchs) and [1, pp. 71-75].

¹⁰ For an explicit statement of this view, see Baumol [8]. But I believe this position is implicit in most discussions of the functions of government.

dictable: 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. Section II is a catalogue of stylized generalizations about the medical-care market which differentiate it from the usual commodity markets. In Section III the behavior of the market is compared with that of the competitive model which disregards the fact of uncertainty. In Section IV, the medical-care market is compared, both as to behavior and as to preconditions, with the ideal competitive market that takes account of uncertainty; an attempt will be made to demonstrate that the characteristics outlined in Section II 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.

II. A Survey of the Special Characteristics of the Medical-Care Market¹²

This section will list selectively some characteristics of medical care which distinguish it from the usual commodity of economics textbooks. The list is not exhaustive, and it is not claimed 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.

A. The Nature of Demand

The most obvious distinguishing characteristics of an individual's demand for medical services is that it is not steady in origin as, 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

¹¹Since writing the above, I find that Buchanan and Tullock [10, Ch. 13] have argued that all redistribution can be interpreted as "income insurance."

¹²For an illuminating survey to which I am much indebted, see S. Mushkin [20].

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, where 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.

B. *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 men 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 relation.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 businessman.¹⁶ (1) Advertising and overt price competition are virtually eliminated among physicians. (2) Advice given by physicians as to further treatment by himself or others is supposed to be completely

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, though the precise social structure is different for reasons that are not hard to analyze.

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 [22, p. 463]. The whole of [22, Ch. 10] is a most illuminating analysis of the social role of medical practice; though Parsons' 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.

divorced from self-interest. (3) 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 as absolute in fact as it is in theory, we can hardly suppose that it 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 (4) The physician is relied on as an expert in certifying to 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 do 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

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 René 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 [13, 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.

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, e.g., 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 [14, Ch. 9].

²⁰ About 3 per cent of beds were in proprietary hospitals in 1958, against 30 per cent in voluntary nonprofit, and the remainder in federal, state, and local hospitals; see [26, 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)

Conformity to collectivity-oriented behavior is especially important since it is a commonplace that the physician-patient relation 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 [14, pp. 194-96]. That purely psychic interactions 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 [25].

C. 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 is its incidence. In 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, in general, not true; the uncertainty due to 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 so 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 relation 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

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.

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 the subjective belief of both parties, as manifested in their market behavior, that is relevant.

have as good or nearly as good an understanding of the utility of the product as the producer.

D. *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 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, which, however, 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 nonprofit-oriented agencies.

Since this last phenomenon goes well beyond the purely medical aspect, we 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 ex-

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

pect 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 subsidy element in medical education has become significant. Previously, many medical schools paid their way or even yielded a profit.

Another interesting feature of limitation on entry to subsidized education is the extent of individual preferences concerning the social welfare, as manifested by contributions to private universities. But whether support is public or private, the important point is that both the quality and the quantity of the supply of medical care are being strongly influenced by social nonmarket forces.^{24, 25}

One striking consequence of the control of quality is the restriction on the range offered. If many qualities of a commodity are possible, it would usually happen in a competitive market that many qualities will be offered on the market, at suitably varying prices, to appeal to different tastes and incomes. Both the licensing laws and the standards of medical-school training have limited the possibilities of alternative qualities of medical care. The declining ratio of physicians to total employees in the medical-care industry shows that substitution of less trained personnel, technicians, and the like, is not prevented completely, but the central role of the highly trained physician is not affected at all.²⁶

E. Pricing Practices

The unusual pricing practices and attitudes of the medical profession are well known: extensive price discrimination by income (with an extreme of zero prices for sufficiently indigent patients) and, formerly, a strong insistence on fee for services as against such alternatives as prepayment.

²⁴ Strictly speaking, there are four variables in the market for physicians: price, quality of entering students, quality of education, and quantity. The basic market forces, demand for medical services and supply of entering students, determine two relations among the four variables. Hence, if the nonmarket forces determine the last two, market forces will determine price and quality of entrants.

²⁵ The supply of Ph.D.'s is similarly governed, but there are other conditions in the market which are much different, especially on the demand side.

²⁶ Today only the Soviet Union offers an alternative lower level of medical personnel, the feldshers, who practice primarily in the rural districts (the institution dates back to the 18th century). According to Field [14, pp. 98-100, 132-33], there is clear evidence of strain in the relations between physicians and feldshers, but it is not certain that the feldshers will gradually disappear as physicians grow in numbers.

The opposition to prepayment is closely related to an even stronger opposition to closed-panel practice (contractual arrangements which bind the patient to a particular group of physicians). Again these attitudes seem to differentiate professions from business. Prepayment and closed-panel plans are virtually nonexistent in the legal profession. In ordinary business, on the other hand, there exists a wide variety of exclusive service contracts involving sharing of risks; it is assumed that competition will select those which satisfy needs best.²⁷

The problems of implicit and explicit price-fixing should also be mentioned. Price competition is frowned on. Arrangements of this type are not uncommon in service industries, and they have not been subjected to antitrust action. How important this is is hard to assess. It has been pointed out many times that the apparent rigidity of so-called administered prices considerably understates the actual flexibility. Here, too, if physicians find themselves with unoccupied time, rates are likely to go down, openly or covertly; if there is insufficient time for the demand, rates will surely rise. The "ethics" of price competition may decrease the flexibility of price responses, but probably that is all.

III. Comparisons with the Competitive Model under Certainty

A. Nonmarketable Commodities

As already noted, the diffusion of communicable diseases provides an obvious example of nonmarket interactions. But from a theoretical viewpoint, the issues are well understood, and there is little point in expanding on this theme. (This should not be interpreted as minimizing the contribution of public health to welfare; there is every reason to suppose that it is considerably more important than all other aspects of medical care.)

Beyond this special area there is a more general interdependence, the concern of individuals for the health of others. The economic manifestations of this taste are to be found in individual donations to hospitals and to medical education, as well as in the widely accepted responsibilities of government in this area. The taste for improving the health of others appears to be stronger than for improving other aspects of their welfare.²⁸

In interdependencies generated by concern for the welfare of others there is always a theoretical case for collective action if each participant derives satisfaction from the contributions of all.

²⁷ The law does impose some limits on risk-shifting in contracts, for example, its general refusal to honor exculpatory clauses.

²⁸ There may be an identification problem in this observation. If the failure of the market system is, or appears to be, greater in medical care than in, say, food an individual otherwise equally concerned about the two aspects of others' welfare may prefer to help in the first.

B. *Increasing Returns*

Problems associated with increasing returns play some role in allocation of resources in the medical field, particularly in areas of low density or low income. Hospitals show increasing returns up to a point; specialists and some medical equipment constitute significant indivisibilities. In many parts of the world the individual physician may be a large unit relative to demand. In such cases it can be socially desirable to subsidize the appropriate medical-care unit. The appropriate mode of analysis is much the same as for water-resource projects. Increasing returns are hardly apt to be a significant problem in general practice in large cities in the United States, and improved transportation to some extent reduces their importance elsewhere.

C. *Entry*

The most striking departure from competitive behavior is restriction on entry to the field, as discussed in II.D above. Friedman and Kuznets, in a detailed examination of the pre-World War II data, have argued that the higher income of physicians could be attributed to this restriction.²⁹

There is some evidence that the demand for admission to medical school has dropped (as indicated by the number of applicants per place and the quality of those admitted), so that the number of medical-school places is not as significant a barrier to entry as in the early 1950's [28, pp. 14-15]. But it certainly has operated over the past and it is still operating to a considerable extent today. It has, of course, constituted a direct and unsubtle restriction on the supply of medical care.

There are several considerations that must be added to help evaluate the importance of entry restrictions: (1) Additional entrants would be, in general, of lower quality; hence, the addition to the supply of medical care, properly adjusted for quality, is less than purely quantitative calculations would show.³⁰ (2) To achieve genuinely competitive conditions, it would be necessary not only to remove numerical restrictions on entry but also to remove the subsidy in medical education. Like any other producer, the physician should bear all the costs of production,

²⁹ See [16, pp. 118-37]. The calculations involve many assumptions and must be regarded as tenuous; see the comments by C. Reinold Noyes in [16, pp. 407-10].

³⁰ It might be argued that the existence of racial discrimination in entrance has meant that some of the rejected applicants are superior to some accepted. However, there is no necessary connection between an increase in the number of entrants and a reduction in racial discrimination; so long as there is excess demand for entry, discrimination can continue unabated and new entrants will be inferior to those previously accepted.

including, in this case, education.³¹ It is not so clear that this change would not keep even unrestricted entry down below the present level. (3) To some extent, the effect of making tuition carry the full cost of education will be to create too few entrants, rather than too many. Given the imperfections of the capital market, loans for this purpose to those who do not have the cash are difficult to obtain. The lender really has no security. The obvious answer is some form of insured loans, as has frequently been argued; not too much ingenuity would be needed to create a credit system for medical (and other branches of higher) education. Under these conditions the cost would still constitute a deterrent, but one to be compared with the high future incomes to be obtained.

If entry were governed by ideal competitive conditions, it may be that the quantity on balance would be increased, though this conclusion is not obvious. The average quality would probably fall, even under an ideal credit system, since subsidy plus selected entry draw some highly qualified individuals who would otherwise get into other fields. The decline in quality is not an over-all social loss, since it is accompanied by increase in quality in other fields of endeavor; indeed, if demands accurately reflected utilities, there would be a net social gain through a switch to competitive entry.³²

There is a second aspect of entry in which the contrast with competitive behavior is, in many respects, even sharper. It is the exclusion of many imperfect substitutes for physicians. The licensing laws, though they do not effectively limit the number of physicians, do exclude all others from engaging in any one of the activities known as medical practice. As a result, costly physician time may be employed at specific tasks for which only a small fraction of their training is needed, and which could be performed by others less well trained and therefore less expensive. One might expect immunization centers, privately operated, but not necessarily requiring the services of doctors.

In the competitive model without uncertainty, consumers are presumed to be able to distinguish qualities of the commodities they buy. Under this hypothesis, licensing would be, at best, superfluous and exclude those from whom consumers would not buy anyway; but it might exclude too many.

D. *Pricing*

The pricing practices of the medical industry (see II.E above) de-

³¹ One problem here is that the tax laws do not permit depreciation of professional education, so that there is a discrimination against this form of investment.

³² To anticipate later discussion, this condition is not necessarily fulfilled. When it comes to quality choices, the market may be inaccurate.

part sharply from the competitive norm. As Kessel [17] has pointed out with great vigor, not only is price discrimination incompatible with the competitive model, but its preservation in the face of the large number of physicians is equivalent to a collective monopoly. In the past, the opposition to prepayment plans has taken distinctly coercive forms, certainly transcending market pressures, to say the least.

Kessel has argued that price discrimination is designed to maximize profits along the classic lines of discriminating monopoly and that organized medical opposition to prepayment was motivated by the desire to protect these profits. In principle, prepayment schemes are compatible with discrimination, but in practice they do not usually discriminate. I do not believe the evidence that the actual scale of discrimination is profit-maximizing is convincing. In particular, note that for any monopoly, discriminating or otherwise, the elasticity of demand in each market at the point of maximum profits is greater than one. But it is almost surely true for medical care that the price elasticity of demand for all income levels is less than one. That price discrimination by income is not completely profit-maximizing is obvious in the extreme case of charity; Kessel argues that this represents an appeasement of public opinion. But this already shows the incompleteness of the model and suggests the relevance and importance of social and ethical factors.

Certainly one important part of the opposition to prepayment was its close relation to closed-panel plans. Prepayment is a form of insurance, and naturally the individual physician did not wish to assume the risks. Pooling was intrinsically involved, and this strongly motivates, as we shall discuss further in Section IV below, control over prices and benefits. The simplest administrative form is the closed panel; physicians involved are, in effect, the insuring agent. From this point of view, Blue Cross solved the prepayment problem by universalizing the closed panel.

The case that price discrimination by income is a form of profit maximization which was zealously defended by opposition to fees for service seems far from proven. But it remains true that this price discrimination, for whatever cause, is a source of nonoptimality. Hypothetically, it means everyone would be better off if prices were made equal for all, and the rich compensated the poor for the changes in the relative positions. The importance of this welfare loss depends on the actual amount of discrimination and on the elasticities of demand for medical services by the different income groups. If the discussion is simplified by considering only two income levels, rich and poor, and if the elasticity of demand by either one is zero, then no reallocation of medical services will take place and the initial situation is optimal. The

only effect of a change in price will be the redistribution of income as between the medical profession and the group with the zero elasticity of demand. With low elasticities of demand, the gain will be small. To illustrate, suppose the price of medical care to the rich is double that to the poor, the medical expenditures by the rich are 20 per cent of those by the poor, and the elasticity of demand for both classes is .5; then the net social gain due to the abolition of discrimination is slightly over 1 per cent of previous medical expenditures.³³

The issues involved in the opposition to prepayment, the other major anomaly in medical pricing, are not meaningful in the world of certainty and will be discussed below.

IV. Comparison with the Ideal Competitive Model under Uncertainty

A. Introduction

In this section we will compare the operations of the actual medical-care market with those of an ideal system in which not only the usual commodities and services but also insurance policies against all conceivable risks are available.³⁴ Departures consist for the most part of

³³ It is assumed that there are two classes, rich and poor; the price of medical services to the rich is twice that to the poor, medical expenditures by the rich are 20 per cent of those by the poor, and the elasticity of demand for medical services is .5 for both classes. Let us choose our quantity and monetary units so that the quantity of medical services consumed by the poor and the price they pay are both 1. Then the rich purchase .1 units of medical services at a price of 2. Given the assumption about the elasticities of demand, the demand function of the rich is $D_R(p) = .14 p^{-0.5}$ and that of the poor is $D_P(p) = p^{-0.5}$. The supply of medical services is assumed fixed and therefore must equal 1.1. If price discrimination were abolished, the equilibrium price, \bar{p} , must satisfy the relation,

$$D_R(\bar{p}) + D_P(\bar{p}) = 1.1,$$

and therefore $\bar{p} = 1.07$. The quantities of medical care purchased by the rich and poor, respectively, would be $D_R(\bar{p}) = .135$ and $D_P(\bar{p}) = .965$.

The inverse demand functions, the price to be paid corresponding to any given quantity are $d_R(q) = .02/q^0.5$, and $d_P(q) = 1/q^0.5$. Therefore, the consumers' surplus to the rich generated by the change is:

$$(1) \quad \int_{.1}^{.135} (.02/q^0.5)dq - \bar{p}(1.135 - .1),$$

and similarly the loss in consumers' surplus by the poor is:

$$(2) \quad \int_{.965}^1 (1/q^0.5)dq - \bar{p}(1 - .965)$$

If (2) is subtracted from (1), the second terms cancel, and the aggregate increase in consumers' surplus is .0156, or a little over 1 per cent of the initial expenditures.

³⁴ A striking illustration of the desire for security in medical care is provided by the expressed preferences of émigrés from the Soviet Union as between Soviet medical practice and German or American practice; see Field [14, Ch. 12]. Those in Germany preferred the German system to the Soviet, but those in the United States preferred (in a ratio of 3 to 1) the Soviet system. The reasons given boil down to the certainty of medical care, independent of income or health fluctuations.

insurance policies that might conceivably be written, but are in fact not. Whether these potential commodities are nonmarketable, or, merely because of some imperfection in the market, are not actually marketed, is a somewhat fine point.

To recall what has already been said in Section I, there are two kinds of risks involved in medical care: the risk of becoming ill, and the risk of total or incomplete or delayed recovery. The loss due to illness is only partially the cost of medical care. It also consists of discomfort and loss of productive time during illness, and, in more serious cases, death or prolonged deprivation of normal function. From the point of view of the welfare economics of uncertainty, both losses are risks against which individuals would like to insure. The nonexistence of suitable insurance policies for either risk implies a loss of welfare.

B. *The Theory of Ideal Insurance*

In this section, the basic principles of an optimal regime for risk-bearing will be presented. For illustration, reference will usually be made to the case of insurance against cost in medical care. The principles are equally applicable to any of the risks. There is no single source to which the reader can be easily referred, though I think the principles are at least reasonably well understood.

As a basis for the analysis, the assumption is made that each individual acts so as to maximize the expected value of a utility function. If we think of utility as attached to income, then the costs of medical care act as a random deduction from this income, and it is the expected value of the utility of income after medical costs that we are concerned with. (Income after medical costs is the ability to spend money on other objects which give satisfaction. We presuppose that illness is not a source of satisfaction in itself; to the extent that it is a source of dissatisfaction, the illness should enter into the utility function as a separate variable.) The expected-utility hypothesis, due originally to Daniel Bernoulli (1738), is plausible and is the most analytically manageable of all hypotheses that have been proposed to explain behavior under uncertainty. In any case, the results to follow probably would not be significantly affected by moving to another mode of analysis.

It is further assumed that individuals are normally risk-aversers. In utility terms, this means that they have a diminishing marginal utility of income. This assumption may reasonably be taken to hold for most of the significant affairs of life for a majority of people, but the presence of gambling provides some difficulty in the full application of this view. It follows from the assumption of risk aversion that if an individual is given a choice between a probability distribution of income, with a given mean m , and the certainty of the income m , he would prefer

the latter. Suppose, therefore, an agency, a large insurance company plan, or the government, stands ready to offer insurance against medical costs on an actuarially fair basis; that is, if the costs of medical care are a random variable with mean m , the company will charge a premium m , and agree to indemnify the individual for all medical costs. Under these circumstances, the individual will certainly prefer to take out a policy and will have a welfare gain thereby.

Will this be a social gain? Obviously yes, if the insurance agent is suffering no social loss. Under the assumption that medical risks on different individuals are basically independent, the pooling of them reduces the risk involved to the insurer to relatively small proportions. In the limit, the welfare loss, even assuming risk aversion on the part of the insurer, would vanish and there is a net social gain which may be of quite substantial magnitude. In fact, of course, the pooling of risks does not go to the limit; there is only a finite number of them and there may be some interdependence among the risks due to epidemics and the like. But then a premium, perhaps slightly above the actuarial level, would be sufficient to offset this welfare loss. From the point of view of the individual, since he has a strict preference for the actuarially fair policy over assuming the risks himself, he will still have a preference for an actuarially unfair policy, provided, of course, that it is not too unfair.

In addition to a residual degree of risk aversion by insurers, there are other reasons for the loading of the premium (i.e., an excess of premium over the actuarial value). Insurance involves administrative costs. Also, because of the irregularity of payments there is likely to be a cost of capital tied up. Suppose, to take a simple case, the insurance company is not willing to sell any insurance policy that a consumer wants but will charge a fixed-percentage loading above the actuarial value for its premium. Then it can be shown that the most preferred policy from the point of view of an individual is a coverage with a deductible amount; that is, the insurance policy provides 100 per cent coverage for all medical costs in excess of some fixed-dollar limit. If, however, the insurance company has some degree of risk aversion, its loading may also depend on the degree of uncertainty of the risk. In that case, the Pareto optimal policy will involve some element of co-insurance, i.e., the coverage for costs over the minimum limit will be some fraction less than 100 per cent (for proofs of these statements, see Appendix).

These results can also be applied to the hypothetical concept of insurance against failure to recover from illness. For simplicity, let us assume that the cost of failure to recover is regarded purely as a money cost, either simply productive opportunities foregone or, more gener-

ally, the money equivalent of all dissatisfactions. Suppose further that, given that a person is ill, the expected value of medical care is greater than its cost; that is, the expected money value attributable to recovery with medical help is greater than resources devoted to medical help. However, the recovery, though on the average beneficial, is uncertain; in the absence of insurance a risk-avertor may well prefer not to take a chance on further impoverishment by buying medical care. A suitable insurance policy would, however, mean that he paid nothing if he doesn't benefit; since the expected value is greater than the cost, there would be a net social gain.³⁵

C. Problems of Insurance

1. *The moral hazard.* The welfare case for insurance policies of all sorts is overwhelming. It follows that the government should undertake insurance in those cases where this market, for whatever reason, has failed to emerge. Nevertheless, there are a number of significant practical limitations on the use of insurance. It is important to understand them, though I do not believe that they alter the case for the creation of a much wider class of insurance policies than now exists.

One of the limits which has been much stressed in insurance literature is the effect of insurance on incentives. What is desired in the case of insurance is that the event against which insurance is taken be out of the control of the individual. Unfortunately, in real life this separation can never be made perfectly. The outbreak of fire in one's house or business may be largely uncontrollable by the individual, but the probability of fire is somewhat influenced by carelessness, and of course arson is a possibility, if an extreme one. Similarly, in medical policies the cost of medical care is not completely determined by the illness suffered by the individual but depends on the choice of a doctor and his willingness to use medical services. It is frequently observed that widespread medical insurance increases the demand for medical care. Coinsurance provisions have been introduced into many major medical policies to meet this contingency as well as the risk aversion of the insurance companies.

To some extent the professional relationship between physician and patient limits the normal hazard in various forms of medical insurance. By certifying to the necessity of given treatment or the lack thereof, the physician acts as a controlling agent on behalf of the insurance companies. Needless to say, it is a far from perfect check; the physicians themselves are not under any control and it may be convenient for them or pleasing to their patients to prescribe more expensive medi-

³⁵ It is a popular belief that the Chinese, at one time, paid their physicians when well but not when sick.

cation, private nurses, more frequent treatments, and other marginal variations of care. It is probably true that hospitalization and surgery are more under the casual inspection of others than is general practice and therefore less subject to moral hazard; this may be one reason why insurance policies in those fields have been more widespread.

2. *Alternative methods of insurance payment.* It is interesting that no less than three different methods of coverage of the costs of medical care have arisen: prepayment, indemnities according to a fixed schedule, and insurance against costs, whatever they may be. In prepayment plans, insurance in effect is paid in kind—that is, directly in medical services. The other two forms both involve cash payments to the beneficiary, but in the one case the amounts to be paid involving a medical contingency are fixed in advance, while in the other the insurance carrier pays all the costs, whatever they may be, subject, of course, to provisions like deductibles and coinsurance.

In hypothetically perfect markets these three forms of insurance would be equivalent. The indemnities stipulated would, in fact, equal the market price of the services, so that value to the insured would be the same if he were to be paid the fixed sum or the market price or were given the services free. In fact, of course, insurance against full costs and prepayment plans both offer insurance against uncertainty as to the price of medical services, in addition to uncertainty about their needs. Further, by their mode of compensation to the physician, prepayment plans are inevitably bound up with closed panels so that the freedom of choice of the physician by the patient is less than it would be under a scheme more strictly confined to the provision of insurance. These remarks are tentative, and the question of coexistence of the different schemes should be a fruitful subject for investigation.

3. *Third-party control over payments.* The moral hazard in physicians' control noted in paragraph 1 above shows itself in those insurance schemes where the physician has the greatest control, namely, major medical insurance. Here there has been a marked rise in expenditures over time. In prepayment plans, where the insurance and medical service are supplied by the same group, the incentive to keep medical costs to a minimum is strongest. In plans of the Blue Cross group, there has developed a conflict of interest between the insurance carrier and the medical-service supplier, in this case particularly the hospital.

The need for third-party control is reinforced by another aspect of the moral hazard. Insurance removes the incentive on the part of individuals, patients, and physicians to shop around for better prices for hospitalization and surgical care. The market forces, therefore, tend to be replaced by direct institutional control.

4. *Administrative costs.* The pure theory of insurance sketched in Section B above omits one very important consideration: the costs of operating an insurance company. There are several types of operating costs, but one of the most important categories includes commissions and acquisition costs, selling costs in usual economic terminology. Not only does this mean that insurance policies must be sold for considerably more than their actuarial value, but it also means there is a great differential among different types of insurance. It is very striking to observe that among health insurance policies of insurance companies in 1958, expenses of one sort or another constitute 51.6 per cent of total premium income for individual policies, and only 9.5 per cent for group policies [26, Table 14-1, p. 272]. This striking differential would seem to imply enormous economies of scale in the provision of insurance, quite apart from the coverage of the risks themselves. Obviously, this provides a very strong argument for widespread plans, including, in particular, compulsory ones.

5. *Predictability and insurance.* Clearly, from the risk-aversion point of view, insurance is more valuable, the greater the uncertainty in the risk being insured against. This is usually used as an argument for putting greater emphasis on insurance against hospitalization and surgery than other forms of medical care. The empirical assumption has been challenged by O. W. Anderson and others [3, pp. 53-54], who asserted that out-of-hospital expenses were equally as unpredictable as in-hospital costs. What was in fact shown was that the probability of costs exceeding \$200 is about the same for the two categories, but this is not, of course, a correct measure of predictability, and a quick glance at the supporting evidence shows that in relation to the average cost the variability is much lower for ordinary medical expenses. Thus, for the city of Birmingham, the mean expenditure on surgery was \$7, as opposed to \$20 for other medical expenses, but of those who paid something for surgery the average bill was \$99, as against \$36 for those with some ordinary medical cost. Eighty-two per cent of those interviewed had no surgery, and only 20 per cent had no ordinary medical expenses [3, Tables A-13, A-18, and A-19 on pp. 72, 77, and 79, respectively].

The issue of predictability also has bearing on the merits of insurance against chronic illness or maternity. On a lifetime insurance basis, insurance against chronic illness makes sense, since this is both highly unpredictable and highly significant in costs. Among people who already have chronic illness, or symptoms which reliably indicate it, insurance in the strict sense is probably pointless.

6. *Pooling of unequal risks.* Hypothetically, insurance requires for its full social benefit a maximum possible discrimination of risks. Those

in groups of higher incidences of illness should pay higher premiums. In fact, however, there is a tendency to equalize, rather than to differentiate, premiums, especially in the Blue Cross and similar widespread schemes. This constitutes, in effect, a redistribution of income from those with a low propensity to illness to those with a high propensity. The equalization, of course, could not in fact be carried through if the market were genuinely competitive. Under those circumstances, insurance plans could arise which charged lower premiums to preferred risks and draw them off, leaving the plan which does not discriminate among risks with only an adverse selection of them.

As we have already seen in the case of income redistribution, some of this may be thought of as insurance with a longer time perspective. If a plan guarantees to everybody a premium that corresponds to total experience but not to experience as it might be segregated by smaller subgroups, everybody is, in effect, insured against a change in his basic state of health which would lead to a reclassification. This corresponds precisely to the use of a level premium in life insurance instead of a premium varying by age, as would be the case for term insurance.

7. *Gaps and coverage.* We may briefly note that, at any rate to date, insurances against the cost of medical care are far from universal. Certain groups—the unemployed, the institutionalized, and the aged—are almost completely uncovered. Of total expenditures, between one-fifth and one-fourth are covered by insurance. It should be noted, however, that over half of all hospital expenses and about 35 per cent of the medical payments of those with bills of \$1,000 a year and over, are included [26, p. 376]. Thus, the coverage on the more variable parts of medical expenditure is somewhat better than the over-all figures would indicate, but it must be assumed that the insurance mechanism is still very far from achieving the full coverage of which it is capable.

D. *Uncertainty of Effects of Treatment*

1. There are really two major aspects of uncertainty for an individual already suffering from an illness. He is uncertain about the effectiveness of medical treatment, and his uncertainty may be quite different from that of his physician, based on the presumably quite different medical knowledges.

2. *Ideal insurance.* This will necessarily involve insurance against a failure to benefit from medical care, whether through recovery, relief of pain, or arrest of further deterioration. One form would be a system in which the payment to the physician is made in accordance with the degree of benefit. Since this would involve transferring the risks from the patient to the physician, who might certainly have an aversion to bearing them, there is room for insurance carriers to pool the risks,

either by contract with physicians or by contract with the potential patients. Under ideal insurance, medical care will always be undertaken in any case in which the expected utility, taking account of the probabilities, exceeds the expected medical cost. This prescription would lead to an economic optimum. If we think of the failure to recover mainly in terms of lost working time, then this policy would, in fact, maximize economic welfare as ordinarily measured.

3. *The concepts of trust and delegation.* In the absence of ideal insurance, there arise institutions which offer some sort of substitute guarantees. Under ideal insurance the patient would actually have no concern with the informational inequality between himself and the physician, since he would only be paying by results anyway, and his utility position would in fact be thoroughly guaranteed. In its absence he wants to have some guarantee that at least the physician is using his knowledge to the best advantage. This leads to the setting up of a relationship of trust and confidence, one which the physician has a social obligation to live up to. Since the patient does not, at least in his belief, know as much as the physician, he cannot completely enforce standards of care. In part, he replaces direct observation by generalized belief in the ability of the physician.³⁶ To put it another way, the social obligation for best practice is part of the commodity the physician sells, even though it is a part that is not subject to thorough inspection by the buyer.

One consequence of such trust relations is that the physician cannot act, or at least appear to act, as if he is maximizing his income at every moment of time. As a signal to the buyer of his intentions to act as thoroughly in the buyer's behalf as possible, the physician avoids the obvious stigmata of profit-maximizing. Purely arms-length bargaining behavior would be incompatible, not logically, but surely psychologically, with the trust relations. From these special relations come the various forms of ethical behavior discussed above, and so also, I suggest, the relative unimportance of profit-making in hospitals. The very word, "profit," is a signal that denies the trust relations.

Price discrimination and its extreme, free treatment for the indigent, also follow. If the obligation of the physician is understood to be first of all to the welfare of the patient, then in particular it takes precedence over financial difficulties.

As a second consequence of informational inequality between physician and patient and the lack of insurance of a suitable type, the patient must delegate to the physician much of his freedom of choice.

³⁶ Francis Bator points out to me that some protection can be achieved, at a price, by securing additional opinions.

He does not have the knowledge to make decisions on treatment, referral, or hospitalization. To justify this delegation, the physician finds himself somewhat limited, just as any agent would in similar circumstances. The safest course to take to avoid not being a true agent is to give the socially prescribed "best" treatment of the day. Compromise in quality, even for the purpose of saving the patient money, is to risk an imputation of failure to live up to the social bond.

The special trust relation of physicians (and allied occupations, such as priests) extends to third parties so that the certifications of physicians as to illness and injury are accepted as especially reliable (see Section II.B above). The social value to all concerned of such presumptively reliable sources of information is obvious.

Notice the general principle here. Because there are barriers to the information flow and because there is no market in which the risks involved can be insured, coordination of purchase and sales must take place through convergent expectations, but these are greatly assisted by having clear and prominent signals, and these, in turn, force patterns of behavior which are not in themselves logical necessities for optimality.³⁷

4. Licensing and educational standards. Delegation and trust are the social institutions designed to obviate the problem of informational inequality. The general uncertainty about the prospects of medical treatment is socially handled by rigid entry requirements. These are designed to reduce the uncertainty in the mind of the consumer as to the quality of product insofar as this is possible.³⁸ I think this explanation, which is perhaps the naive one, is much more tenable than any idea of a monopoly seeking to increase incomes. No doubt restriction on entry is desirable from the point of view of the existing physicians, but the public pressure needed to achieve the restriction must come from deeper causes.

The social demand for guaranteed quality can be met in more than one way, however. At least three attitudes can be taken by the state or other social institutions toward entry into an occupation or toward the production of commodities in general; examples of all three types exist. (1) The occupation can be licensed, nonqualified entrants being simply excluded. The licensing may be more complex than it is in medicine; individuals could be licensed for some, but not all, medical activities, for example. Indeed, the present all-or-none approach could

³⁷ The situation is very reminiscent of the crucial role of the focal point in Schelling's theory of tacit games, in which two parties have to find a common course of action without being able to communicate; see [24, esp. pp. 225 ff.].

³⁸ How well they achieve this end is another matter. R. Kessel points out to me that they merely guarantee training, not continued good performance as medical technology changes.

be criticized as being insufficient with regard to complicated specialist treatment, as well as excessive with regard to minor medical skills. Graded licensing may, however, be much harder to enforce. Controls could be exercised analogous to those for foods; they can be excluded as being dangerous, or they can be permitted for animals but not for humans. (2) The state or other agency can certify or label, without compulsory exclusion. The category of Certified Psychologist is now under active discussion; canned goods are graded. Certification can be done by nongovernmental agencies, as in the medical-board examinations for specialists. (3) Nothing at all may be done; consumers make their own choices.

The choice among these alternatives in any given case depends on the degree of difficulty consumers have in making the choice unaided, and on the consequences of errors of judgment. It is the general social consensus, clearly, that the *laissez-faire* solution for medicine is intolerable. The certification proposal never seems to have been discussed seriously. It is beyond the scope of this paper to discuss these proposals in detail. I wish simply to point out that they should be judged in terms of the ability to relieve the uncertainty of the patient in regard to the quality of the commodity he is purchasing, and that entry restrictions are the consequences of an apparent inability to devise a system in which the risks of gaps in medical knowledge and skill are borne primarily by the patient, not the physician.

Postscript

I wish to repeat here what has been suggested above in several places: that the failure of the market to insure against uncertainties has created many social institutions in which the usual assumptions of the market are to some extent contradicted. The medical profession is only one example, though in many respects an extreme one. All professions share some of the same properties. The economic importance of personal and especially family relationships, though declining, is by no means trivial in the most advanced economies; it is based on non-market relations that create guarantees of behavior which would otherwise be afflicted with excessive uncertainty. Many other examples can be given. The logic and limitations of ideal competitive behavior under uncertainty force us to recognize the incomplete description of reality supplied by the impersonal price system.

REFERENCES

1. A. A. ALCHIAN, K. J. ARROW, AND W. M. CAPRON, *An Economic Analysis of the Market for Scientists and Engineers*, RAND RM-2190-RC. Santa Monica 1958.

2. M. ALLAIS, "Généralisation des théories de l'équilibre économique général et du rendement social au cas du risque," in Centre National de la Recherche Scientifique, *Econometrie*, Paris 1953, pp. 1-20.
3. O. W. ANDERSON AND STAFF OF THE NATIONAL OPINION RESEARCH CENTER, *Voluntary Health Insurance in Two Cities*. Cambridge, Mass. 1957.
4. K. J. ARROW, "Economic Welfare and the Allocation of Resources for Invention," in Nat. Bur. Econ. Research, *The Role and Direction of Inventive Activity: Economic and Social Factors*, Princeton 1962, pp. 609-25.
5. ———, "Les rôle des valeurs boursières pour la répartition la meilleure des risques," in Centre National de la Recherche Scientifique, *Econometrie*, Paris 1953, pp. 41-46.
6. F. M. BATOR, "The Anatomy of Market Failure," *Quart. Jour. Econ.*, Aug. 1958, 72, 351-79.
7. E. BAUDIER, "L'introduction du temps dans la théorie de l'équilibre général," *Les Cahiers Economiques*, Dec. 1959, 9-16.
8. W. J. BAUMOL, *Welfare Economics and the Theory of the State*. Cambridge, Mass. 1952.
9. K. BORCH, "The Safety Loading of Reinsurance Premiums," *Skandinavisk Aktuariehdskrift*, 1960, pp. 163-84.
10. J. M. BUCHANAN AND G. TULLOCK, *The Calculus of Consent*. Ann Arbor 1962.
11. G. DEBREU, "Une économique de l'incertain," *Economie Appliquée*, 1960, 13, 111-16.
12. ———, *Theory of Values*. New York 1959.
13. R. DUBOS, "Medical Utopias," *Daedalus*, 1959, 88, 410-24.
14. M. G. FIELD, *Doctor and Patient in Soviet Russia*. Cambridge, Mass. 1957.
15. MILTON FRIEDMAN, "The Methodology of Positive Economics," in *Essays in Positive Economics*, Chicago 1953, pp. 3-43.
16. ——— AND S. S. KUZNETS, *Income from Independent Professional Practice*. Nat. Bur. Econ. Research, New York 1945.
17. R. A. KESSEL, "Price Discrimination in Medicine," *Jour. Law and Econ.*, 1958, 1, 20-53.
18. T. C. KOOPMANS, "Allocation of Resources and the Price System," in *Three Essays on the State of Economic Science*, New York 1957, pp. 1-120.
19. I. M. D. LITTLE, *A Critique of Welfare Economics*. Oxford 1950.
20. SELMA MUSHKIN, "Towards a Definition of Health Economics," *Public Health Reports*, 1958, 73, 785-93.
21. R. R. NELSON, "The Simple Economics of Basic Scientific Research," *Jour. Pol. Econ.*, June 1959, 67, 297-306.
22. T. PARSONS, *The Social System*. Glencoe 1951.
23. M. J. PECK AND F. M. SCHERER, *The Weapons Acquisition Process: An Economic Analysis*. Div. of Research, Graduate School of Business, Harvard University, Boston 1962.

24. T. C. SCHELLING, *The Strategy of Conflict*. Cambridge, Mass. 1960.
25. A. K. SHAPIRO, "A Contribution to a History of the Placebo Effect," *Behavioral Science*, 1960, 5, 109-35.
26. H. M. SOMERS AND A. R. SOMERS, *Doctors, Patients, and Health Insurance*. The Brookings Institution, Washington 1961.
27. C. L. STEVENSON, *Ethics and Language*. New Haven 1945.
28. U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE, *Physicians for a Growing America*, Public Health Service Publication No. 709, Oct. 1959.

APPENDIX

On Optimal Insurance Policies

The two propositions about the nature of optimal insurance policies asserted in Section IV.B above will be proved here.

Proposition 1. If an insurance company is willing to offer any insurance policy against loss desired by the buyer at a premium which depends only on the policy's actuarial value, then the policy chosen by a risk-averting buyer will take the form of 100 per cent coverage above a deductible minimum.

Note: The premium will, in general, exceed the actuarial value; it is only required that two policies with the same actuarial value will be offered by the company for the same premium.

Proof: Let W be the initial wealth of the individual, X his loss, a random variable, $I(X)$ the amount of insurance paid if loss X occurs, P the premium, and $Y(X)$ the wealth of the individual after paying the premium, incurring the loss, and receiving the insurance benefit.

$$(1) \quad Y(X) = W - P - X + I(X).$$

The individual values alternative policies by the expected utility of his final wealth position, $Y(X)$. Let $U(y)$ be the utility of final wealth, y ; then his aim is to maximize,

$$(2) \quad E\{U[Y(X)]\},$$

where the symbol, E , denotes mathematical expectation.

An insurance payment is necessarily nonnegative, so the insurance policy must satisfy the condition,

$$(3) \quad I(X) \geq 0 \quad \text{for all } X.$$

If a policy is optimal, it must in particular be better in the sense of the criterion (2), than any other policy with the same actuarial expectation, $E[I(X)]$. Consider a policy that pays some positive amount of insurance at one level of loss, say X_1 , but which permits the final wealth at some other loss level, say X_2 , to be lower than that corresponding to X_1 . Then, it is intuitively obvious that a risk-avertor would prefer an alternative policy with the same actuarial value which would offer slightly less protection for losses in the neighborhood of X_1 and slightly higher protection for those in the neighborhood of X_2 , since risk aversion implies that the marginal utility

of $Y(X)$ is greater when $Y(X)$ is smaller: hence, the original policy cannot be optimal.

To prove this formally, let $I_1(X)$ be the original policy, with $I_1(X) > 0$ and $Y_1(X_1) > Y_2(X_2)$, where $Y_1(X)$ is defined in terms of $I_1(X)$ by (I). Choose δ sufficiently small so that,

$$(4) \quad I_1(X) > 0 \quad \text{for } X_1 \leq X \leq X_1 + \delta,$$

$$(5) \quad Y_1(X') < Y_1(X) \quad \text{for } X_2 \leq X' \leq X_2 + \delta, \quad X_1 \leq X \leq X_1 + \delta.$$

(This choice of δ is possible if the functions $I_1(X)$, $Y_1(X)$ are continuous; this can be proved to be true for the optimal policy, and therefore we need only consider this case.)

Let π_1 be the probability that the loss, X , lies in the interval $\langle X_1, X_1 + \delta \rangle$, π_2 the probability that X lies in the interval $\langle X_2, X_2 + \delta \rangle$. From (4) and (5) we can choose $\epsilon > 0$ and sufficiently small so that,

$$(6) \quad I_1(X) - \pi_2\epsilon \geq 0 \quad \text{for } X_1 \leq X \leq X_1 + \delta,$$

$$(7) \quad Y_1(X') + \pi_1\epsilon < Y_1(X) - \pi_2\epsilon$$

$$\text{for } X_2 \leq X' \leq X_2 + \delta, \quad X_1 \leq X \leq X_1 + \delta.$$

Now define a new insurance policy, $I_2(X)$, which is the same as $I_1(X)$ except that it is smaller by $\pi_2\epsilon$ in the interval from X_1 to $X_1 + \delta$ and larger by $\pi_1\epsilon$ in the interval from X_2 to $X_2 + \delta$. From (6), $I_2(X) \geq 0$ everywhere, so that (3) is satisfied. We will show that $E[I_1(X)] = E[I_2(X)]$ and that $I_2(X)$ yields the higher expected utility, so that $I_1(X)$ is not optimal.

Note that $I_2(X) - I_1(X)$ equals $-\pi_2\epsilon$ for $X_1 \leq X \leq X_1 + \delta$, $\pi_1\epsilon$ for $X_2 \leq X \leq X_2 + \delta$, and 0 elsewhere. Let $\phi(X)$ be the density of the random variable X . Then,

$$\begin{aligned} E[I_2(X) - I_1(X)] &= \int_{X_1}^{X_1+\delta} [I_2(X) - I_1(X)]\phi(X)dX \\ &\quad + \int_{X_2}^{X_2+\delta} [I_2(X) - I_1(X)]dX \\ &= (-\pi_2\epsilon) \int_{X_1}^{X_1+\delta} \phi(X)dX + (\pi_1\epsilon) \int_{X_2}^{X_2+\delta} \phi(X)dX \\ &= -(\pi_2\epsilon)\pi_1 + (\pi_1\epsilon)\pi_2 = 0, \end{aligned}$$

so that the two policies have the same actuarial value and, by assumption, the same premium.

Define $Y_2(X)$ in terms of $I_2(X)$ by (1). Then $Y_2(X) - Y_1(X) = I_2(X) - I_1(X)$. From (7),

$$(8) \quad Y_1(X') < Y_2(X') < Y_2(X) < Y_1(X)$$

$$\text{for } X_2 \leq X' \leq X_2 + \delta, \quad X_1 \leq X \leq X_1 + \delta.$$

Since $Y_1(X) - Y_2(X) = 0$ outside the intervals $\langle X_1, X_1 + \delta \rangle$, $\langle X_2, X_2 + \delta \rangle$, we

can write,

$$(9) \quad E\{U[Y_2(X)] - U[Y_1(X)]\} = \int_{X_1}^{X_1+\delta} \{U[Y_2(X)] - U[Y_1(X)]\} \phi(X) dX \\ + \int_{X_2}^{X_2+\delta} \{U[Y_2(X)] - U[Y_1(X)]\} \phi(X) dX.$$

By the Mean Value Theorem, for any given value of X ,

$$(10) \quad U[Y_2(X)] - U[Y_1(X)] = U'[Y(X)][Y_2(X) - Y_1(X)] \\ = U'[Y(X)][I_2(X) - I_1(X)],$$

where $Y(X)$ lies between $Y_1(X)$ and $Y_2(X)$. From (8),

$$Y(X') < Y(X) \text{ for } X_2 \leq X' \leq X_2 + \delta, \quad X_1 \leq X \leq X_1 + \delta,$$

and, since $U'(y)$ is a diminishing function of y for a risk-avertor,

$$U'[Y(X')] > U'[Y(X)]$$

or, equivalently, for some number u ,

$$(11) \quad \begin{aligned} U'[Y(X')] &> u \text{ for } X_2 \leq X' \leq X_2 + \delta, \\ U'[Y(X)] &< u \text{ for } X_1 \leq X \leq X_1 + \delta. \end{aligned}$$

Now substitute (10) into (9),

$$E\{U[Y_2(X)] - U[Y_1(X)]\} = -\pi_2 \epsilon \int_{X_1}^{X_1+\delta} U'[Y(X)] \phi(X) dX \\ + \pi_1 \epsilon \int_{X_2}^{X_2+\delta} U'[Y(X)] \phi(X) dX.$$

From (11), it follows that,

$$E\{U[Y_2(X)] - U[Y_1(X)]\} > -\pi_2 \epsilon u \pi_1 + \pi_1 \epsilon u \pi_2 = 0,$$

so that the second policy is preferred.

It has thus been shown that a policy cannot be optimal if, for some X_1 and X_2 , $I(X_1) > 0$, $Y(X_1) > Y(X_2)$. This may be put in a different form: Let Y_{\min} be the minimum value taken on by $Y(X)$ under the optimal policy; then we must have $I(X) = 0$ if $Y(X) > Y_{\min}$. In other words, a minimum final wealth level is set; if the loss would not bring wealth below this level, no benefit is paid, but if it would, then the benefit is sufficient to bring up the final wealth position to the stipulated minimum. This is, of course, precisely a description of 100 per cent coverage for loss above a deductible.

We turn to the second proposition. It is now supposed that the insurance company, as well as the insured, is a risk-avertor; however, there are no administrative or other costs to be covered beyond protection against loss.

Proposition 2. If the insured and the insurer are both risk-averters and there are no costs other than coverage of losses, then any nontrivial Pareto-

optimal policy, $I(X)$, as a function of the loss, X , must have the property, $0 < dI/dX < 1$.

That is, any increment in loss will be partly but not wholly compensated by the insurance company; this type of provision is known as coinsurance. Proposition 2 is due to Borch [9, Sec. 2]; we give here a somewhat simpler proof.

Proof: Let $U(y)$ be the utility function of the insured, $V(z)$ that of the insurer. Let W_0 and W_1 be the initial wealths of the two, respectively. In this case, we let $I(X)$ be the insurance benefits less the premium; for the present purpose, this is the only significant magnitude (since the premium is independent of X , this definition does not change the value of dI/dX). The final wealth positions of the insured and insurer are:

$$(12) \quad \begin{aligned} Y(X) &= W_0 - X + I(X), \\ Z(X) &= W_1 - I(X), \end{aligned}$$

respectively. Any given insurance policy then defines expected utilities, $u = E\{U[Y(X)]\}$ and $v = E\{V[Z(X)]\}$, for the insured and insurer, respectively. If we plot all points (u, v) obtained by considering all possible insurance policies, the resulting expected-utility-possibility set has a boundary that is convex to the northeast. To see this, let $I_1(X)$ and $I_2(X)$ be any two policies, and let (u_1, v_1) and (u_2, v_2) be the corresponding points in the two-dimensional expected-utility-possibility set. Let a third insurance policy, $I(X)$, be defined as the average of the two given ones,

$$I(X) = (\frac{1}{2})I_1(X) + (\frac{1}{2})I_2(X),$$

for each X . Then, if $Y(X)$, $Y_1(X)$, and $Y_2(X)$ are the final wealth positions of the insured, and $Z(X)$, $Z_1(X)$, and $Z_2(X)$ those of the insurer for each of the three policies, $I(X)$, $I_1(X)$, and $I_2(X)$, respectively,

$$\begin{aligned} Y(X) &= (\frac{1}{2})Y_1(X) + (\frac{1}{2})Y_2(X), \\ Z(X) &= (\frac{1}{2})Z_1(X) + (\frac{1}{2})Z_2(X), \end{aligned}$$

and, because both parties have diminishing marginal utility,

$$\begin{aligned} U[Y(X)] &\geq (\frac{1}{2})U[Y_1(X)] + (\frac{1}{2})U[Y_2(X)], \\ V[Z(X)] &\geq (\frac{1}{2})V[Z_1(X)] + (\frac{1}{2})V[Z_2(X)]. \end{aligned}$$

Since these statements hold for all X , they also hold when expectations are taken. Hence, there is a point (u, v) in the expected-utility-possibility set for which $u \geq (\frac{1}{2})u_1 + (\frac{1}{2})u_2$, $v \geq (\frac{1}{2})v_1 + (\frac{1}{2})v_2$. Since this statement holds for every pair of points (u_1, v_1) and (u_2, v_2) in the expected-utility-possibility set, and in particular for pairs of points on the northeast boundary, it follows that the boundary must be convex to the northeast.

From this, in turn, it follows that any given Pareto-optimal point (i.e., any point on the northeast boundary) can be obtained by maximizing a linear function, $\alpha u + \beta v$, with suitably chosen α and β nonnegative and at least one positive, over the expected-utility-possibility set. In other words, a Pareto-optimal insurance policy, $I(X)$, is one which maximizes,

$$\alpha E\{U[Y(X)]\} + \beta E\{V[Z(X)]\} = E\{\alpha U[Y(X)] + \beta V[Z(X)]\},$$

for some $\alpha \geq 0$, $\beta \geq 0$, $\alpha > 0$ or $\beta > 0$. To maximize this expectation, it is obviously sufficient to maximize:

$$(13) \quad \alpha U[Y(X)] + \beta V[Z(X)],$$

with respect to $I(X)$, for each X . Since, for given X , it follows from (12) that,

$$dY(X)/dI(X) = 1, \quad dZ(X)/dI(X) = -1,$$

it follows by differentiation of (13) that $I(X)$ is the solution of the equation,

$$(14) \quad \alpha U'[Y(X)] - \beta V'[Z(X)] = 0.$$

The cases $\alpha=0$ or $\beta=0$ lead to obvious trivialities (one party simply hands over all his wealth to the other), so we assume $\alpha>0$, $\beta>0$. Now differentiate (14) with respect to X and use the relations, derived from (12),

$$dY/dX = (dI/dX) - 1, \quad dZ/dX = -(dI/dX).$$

$$\alpha U''[Y(X)][(dI/dX) - 1] + \beta V''[Z(X)](dI/dX) = 0,$$

or

$$dI/dX = \alpha U''[Y(X)]/\{\alpha U''[Y(X)] + \beta V''[Z(X)]\}.$$

Since $U''[Y(X)]<0$, $V''[Z(X)]<0$ by the hypothesis that both parties are risk-aversers, Proposition 2 follows.

NATIONAL DEBT IN A NEOCLASSICAL GROWTH MODEL

*By PETER A. DIAMOND**

This paper contains a model designed to serve two purposes, to examine long-run competitive equilibrium in a growth model and then to explore the effects on this equilibrium of government debt. Samuelson [8] has examined the determination of interest rates in a single-commodity world without durable goods. In such an economy, interest rates are determined by consumption loans between individuals of different ages. By introducing production employing a durable capital good into this model, one can examine the case where individuals provide for their retirement years by lending to entrepreneurs. After describing alternative long-run equilibria available to a centrally planned economy, the competitive solution is described. In this economy, which has an infinitely long life, it is seen that, despite the absence of all the usual sources of inefficiency, the competitive solution can be inefficient.

Modigliani [4] has explored the effects of the existence of government debt in an aggregate growth model. By introducing a government which issues debt and levies taxes to finance interest payments into the model described in the first part, it is possible to re-examine his conclusions in a model where consumption decisions are made individually, where taxes to finance the debt are included in the analysis, and where the changes in output arising from changes in the capital stock are explicitly acknowledged. It is seen that in the "normal" case external debt reduces the utility of an individual living in long-run equilibrium. Surprisingly, internal debt is seen to cause an even larger decline in this utility level.

External debt has two effects in the long run, both arising from the taxes needed to finance the interest payments. The taxes directly reduce available lifetime consumption of the individual taxpayer. Further, by reducing his disposable income, taxes reduce his savings and thus the capital stock. Internal debt has both of these effects as well as a further reduction in the capital stock arising from the substitution of government debt for physical capital in individual portfolios.

* The author is assistant professor of economics at the University of California, Berkeley. He wishes to thank his colleagues Bernard Saffran and Sidney G. Winter, Jr. for many helpful discussions on this subject. Errors and interpretations are solely his own.

1. Technology

The economy being considered here is assumed to have an infinite future. Its unchanging technology is assumed to be representable by a constant returns to scale aggregate production function, $F(K, L)$.¹ Since the economy exists in discrete time, the capital argument of the production function is the saving of the previous period plus the capital stock employed in the previous period. (It is assumed that there is no depreciation and that, since capital and output are the same commodity, one can consume one's capital.)

Individuals in this economy live for two periods, working in the first while being retired in the second. Each person has an ordinal utility function $U(e^1, e^2)$ based on his consumption in the two years of his life.² Denoting the number of persons born at the start of the t th period by L_t , labor force, growth satisfies:

$$L_t = L_0(1 + n)^t.$$

2. Centrally Planned Economy

It is simplest to examine the production possibilities of this economy by examining the alternatives available to a central planning authority. With the capital stock in period t (which was determined in period $t-1$) and the labor force in this period (which is exogenous), output will satisfy $Y_t = F(K_t, L_t)$. At the end of the production process (and before the start of consumption in this period) the central authorities have command over the capital stock and the newly produced output, $K_t + Y_t$. This must be divided between the capital stock which will be available for production in the next period, K_{t+1} , and aggregate consumption in this period, C_t . This consumption must be further divided between members of the younger generation, E_t^1 , and those of the older generation, E_t^2 . Assuming that all members of the same generation consume the same amount, we have:³

$$E_t^1 = e_t^1 L_t, \quad E_t^2 = e_t^2 L_{t-1}.$$

The division of the resources on hand between the alternative uses can be stated algebraically:

$$(1) \quad Y_t + K_t = K_{t+1} + C_t = K_{t+1} + e_t^1 L_t + e_t^2 L_{t-1},$$

¹ It is assumed that F is twice differentiable and exhibits positive marginal products and a diminishing marginal rate of substitution everywhere.

² The assumption of the absence of all bequests is important for intertemporal allocation conclusions. A relationship between changes in the size of the national debt and changes in bequests can alter the effects to be described.

³ Note that a person born in period t would consume e_t^1 and e_{t+1}^2 in his two years of life.

or more conventionally,

$$(2) \quad Y_t - (K_{t+1} - K_t) = C_t = e_t^1 L_t + e_t^2 L_{t-1}.$$

Assuming that the central authorities decide to preserve a constant capital-labor ratio, $k_t = K_t/L_t$, and thus $K_{t+1} = (1+n)K_t$, aggregate consumption will satisfy:

$$(3) \quad Y_t - nK_t = C_t = e_t^1 L_t + e_t^2 L_{t-1}.$$

Denoting the output-labor ratio by $y_t = Y_t/L_t$, this can be rewritten as:

$$(4) \quad y_t - nk_t = C_t/L_t = e_t^1 + e_t^2/(1+n).$$

Maintenance of a constant capital-labor ratio implies, of course, a constant output per worker over time. Thus, this equation describes the consumption possibilities in each year of any period during which the capital-labor ratio remains constant. In particular, if a given capital-labor ratio is held constant for all time, the economy is on what has become known as a Golden Age Path.

3. Neoclassical Stationary States or Golden Age Paths

A Golden Age Path for an economy is an expansion path on which the capital-labor ratio (and thus the capital-output ratio and marginal product of capital) is kept constant. From equation (4) we see that the central-planning authorities can maintain any capital-labor ratio for which the output-capital ratio is not smaller than n (which is equivalent to the condition that the savings rate not exceed one). From equation (4), again, one can derive the amount of consumption that is possible in each period and thus calculate the Golden Age Path for which this is maximized. Similarly we can examine the alternative divisions of this consumption between individuals of different generations. Assuming that all individuals have the same lifetime consumption pattern, the problem of selecting the optimal Golden Age Path, the Golden Age Path on which each individual would have the highest utility level, subject to the constraint that all individuals have the same level, can be written:

$$(5) \quad \text{Maximize } U(e^1, e^2) \quad \text{subject to } e^1 + e^2/(1+n) = y - nk.$$

Thus the solution of the problem of selecting an optimal Golden Age Path treats the allocation of consumption over the lifetime of an individual in a similar fashion to the allocation of consumption, in a single year, between individuals of different ages. The selection between Golden Age Paths is, as is seen from (5), a selection which ignores initial conditions, and thus not a selection available to an economy, which

must weigh the advantages of a given long-run equilibrium against the costs of achieving it.

4. *The Golden Rule Path*

This maximization problem decomposes naturally into two separate problems, that of selecting the optimal capital-labor ratio, and thus the height of the consumption constraint, $y-nk$; and that of dividing this amount of consumption between the different individuals. The maximizing capital-labor ratio is seen from (5) to satisfy the condition that the marginal product of capital equal the rate of growth, $F_K=n$. This is the standard result on the nature of the Golden Rule Path, see, e.g., Phelps [6]. Note that the optimality of this capital-labor ratio is independent of the exact division of consumption (and selecting the optimal division is independent of the capital-labor ratio chosen). If the central planners choose a higher capital-labor ratio, they would be selecting an inefficient solution (in the standard sense including the problem of initial conditions, not just as a comparison of Golden Age Paths) in that they could discard capital, lowering the capital-labor ratio to the Golden Rule level, and preserve this capital-labor ratio forever, permitting a higher level of consumption in each period forever.⁴

Utility maximizing consumption allocation clearly requires that

$$\frac{\partial U}{\partial e^1} = (1 + n) \frac{\partial U}{\partial e^2}.$$

This is the allocation that would occur if consumption decisions were individually made employing a rate of interest for consumer decisions equal to the rate of growth. In examining the division of consumption when the capital-labor ratio is held constant, we are equivalently examining a model in which there is only one factor of production, labor. Thus it is not surprising that the optimal allocation is the same as that found by Samuelson [8], which he called the biological optimum. Thus the optimal rate of interest is determined by the rate of population growth (which may or may not equal the marginal product of capital). This paradox arises from the comparison of stationary states. The shifting of one unit of consumption by an individual from his first to his second year is equivalent to removing one unit of consumption from each of the living members of the younger generation and giving this total to the contemporary older generation, of whom there are n per cent fewer members.

⁴ Dynamic inefficiencies of this sort in models both with and without technical change are examined by Phelps [7].

5. Competitive Framework

To the technological possibilities which have been described above, it is necessary to replace the central planning framework by a market process for the determination of the saving rate in each period. The annual savings behavior of the economy will determine the long-run equilibrium to which the economy converges. In particular, we will be interested in comparing alternative Golden Age Paths to which the economy converges with different quantites of government debt outstanding. Thus, only the long-run implications of national debt will be examined, thereby avoiding the problem of selecting a social welfare function for the evaluation of different individual utility levels (at the cost of failing to explore the total effects).

By following the life history of a single individual, born, say, in period t , it is possible to trace out the market relations. This individual works in period t , for which he receives a wage, w_t , which equals the marginal product of labor, $F_L(K_t, L_t)$. This wage he allocates between current and future consumption so as to maximize his utility function, given the rate of interest existing on one-period loans from period t to period $t+1$, r_{t+1} . Thus, the members of the younger generation make up the supply side of the capital market.

This individual will thus consume, in period t , the difference between his wage and the quantity he lends in the capital market, $e^1_t = w_t - s_t$. In period $t+1$, he will consume his savings plus the accrued interest, $e^2_{t+1} = (1+r_{t+1})s_t$.

Capital demanders are entrepreneurs who wish to employ capital for production in period $t+1$. Thus the equilibrium interest rate will equal the marginal product of capital, $r_{t+1} = F_K(K_{t+1}, L_{t+1})$.

6. Factor-Price Frontier

The existence of the constant returns to scale production function, $F(K, L)$, which can be written as $Lf(k)$, implies a relationship between the marginal products of labor and capital which will be denoted by $w = \phi(r)$.⁵ From the definitions $r = f'(k)$ and $w = f(k) - kf'(k)$ we see that:

$$(6) \quad \frac{dw}{dr} = \phi'(r) = -k \quad \text{and} \quad \frac{d^2w}{dr^2} = \phi''(r) = \frac{-1}{f''(k)}.$$

7. Utility Maximization

Utility maximization,⁶ given a wage level and a market interest rate,

⁵ For a description of the factor-price frontier see Samuelson [9].

⁶ It is assumed that the utility function has the following properties: no satiation, a diminishing marginal rate of substitution everywhere, and a shape which guarantees that consump-

implies that consumption will be allocated so that:

$$\frac{\partial U}{\partial e^1} = (1 + r) \frac{\partial U}{\partial e^2}.$$

Therefore, the quantity saved can be expressed as a function of the relevant wage and interest level, $s_t = s(w_t, r_{t+1})$. It will be assumed that s is a differentiable function. From the assumption of normality, we have $0 < \partial s / \partial w < 1$. However $\partial s / \partial r$ may be positive or negative.

In addition to writing individual savings as a function of the wage and interest rates, it is possible to express the utility function in terms of these variables. From this derived form of the utility function one has:⁷

$$(7) \quad \frac{\partial U}{\partial w} = \frac{\partial U}{\partial e^1}, \quad \frac{\partial U}{\partial r} = \frac{s}{(1 + r)} \frac{\partial U}{\partial e^1}.$$

8. Capital Market

From the discussion above, we know that we can write the supply schedule of capital, which is the sum of the individual savings functions, as:

$$(8) \quad S_t = s_t L_t = L_t s(w_t, r_{t+1}).$$

The demand curve for capital, which relates the capital stock in period $t+1$ to the interest rate, is merely the marginal product of capital as a function of the capital-labor ratio:

$$(9) \quad r_{t+1} = f'(K_{t+1}/L_{t+1}).$$

tion in each period is a normal good, i.e.,

$$0 < \frac{\partial s}{\partial w} < 1.$$

⁷ Using the optimality condition, we have:

$$\frac{\partial U}{\partial w} = \frac{\partial U}{\partial e^1} \frac{\partial e^1}{\partial w} + \frac{\partial U}{\partial e^2} \frac{\partial e^2}{\partial w} = \frac{\partial U}{\partial e^1} \left[\frac{\partial e^1}{\partial w} + \left(\frac{1}{1+r} \right) \frac{\partial e^2}{\partial w} \right].$$

From the net worth constraint $e^1 + e^2/(1+r) = w$, we have

$$\frac{\partial e^1}{\partial w} + \frac{\partial e^2}{\partial w} / (1+r) = 1,$$

which, upon substitution, yields equation (7). Similarly, the net worth constraint implies that

$$\frac{\partial e^1}{\partial r} + \frac{\partial e^2}{\partial r} / (1+r) - \frac{e^2}{(1+r)^2} = 0.$$

Thus

$$\frac{\partial U}{\partial r} = \frac{\partial U}{\partial e^1} \left[\frac{\partial e^1}{\partial r} + \left(\frac{1}{1+r} \right) \frac{\partial e^2}{\partial r} \right] = \frac{\partial U}{\partial e^1} \frac{e^2}{(1+r)^2} = \frac{s}{(1+r)} \frac{\partial U}{\partial e^1}.$$

Combining the demand and supply curves, equating S_t and K_{t+1} , we have the equilibrium condition in the capital market, which relates the interest rate to the wage rate of the previous period:

$$(10) \quad r_{t+1} = f'(S_t/L_{t+1}) = f'(s(w_t, r_{t+1})/(1+n)).$$

From the assumptions made above, we know that the demand curve is downward-sloping, while the supply curve may have a positive or negative slope. This suggests that there are two cases which need to be treated separately as the demand or supply curve is more steeply negatively sloped.⁸ This is shown diagrammatically in Diagram 1.

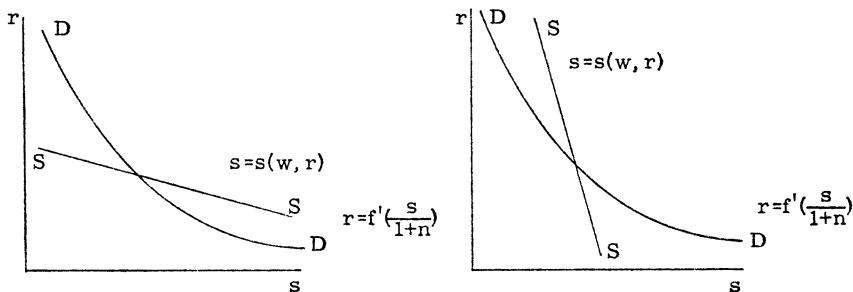


Diagram 1

The necessity of distinguishing the two cases is made clear by examining the relation between the equilibrium interest rate and the wage of the previous period. A higher wage in period t implies a greater quantity of saving at any interest rate, or a rightward shift of the saving curve in Diagram 1. However, whether this results in a higher or lower equilibrium level of saving depends on the relative slopes of the demand and supply curves. Geometrically, assuming $w' > w$, we have Diagram 2.

In the diagram on the right, which represents the "normal" case in the capital market, a higher income level results in a higher equilibrium level of saving. In the diagram on the left, where the elasticity of saving with respect to the interest rate is large and negative, a rise in the level of income results in a fall in the equilibrium level of saving. This somewhat perverse case leads to a reversal of the conclusions on the effect of debt (since taxes which reduce disposable income increase saving). Rather than complicate the text, we discuss this case in Appendix A.

By altering the wage levels in period t , we could trace out the equilibrium interest rates which would occur in period $t+1$. This relation

⁸ A requirement of Walrasian stability in the capital market would permit an elimination of the case where the supply curve is steeper than the demand curve. Marshallian stability would not, of course, permit this elimination. In the absence of a dynamic theory of the capital market, it seems best to consider both cases.

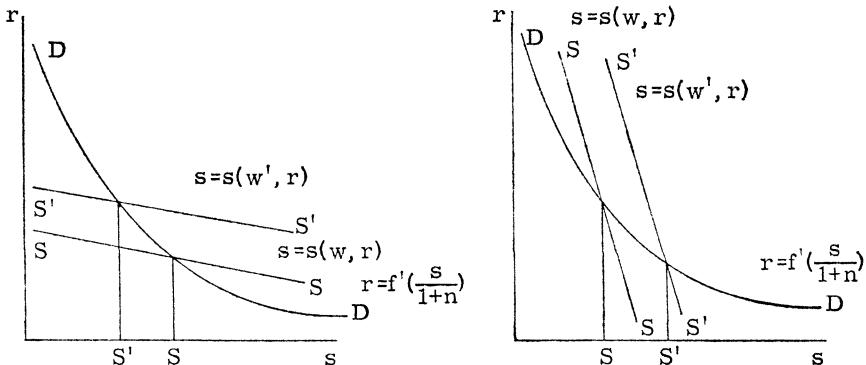


Diagram 2

will be denoted by $r_{t+1} = \psi(w_t)$. It will be assumed that ψ is differentiable. From the assumption on the relative slopes of the demand and supply curves for capital, we know that an increase in wages implies an increase in saving and thus a decrease in the interest rate. Taking the derivative of r with respect to w , we can express this as:

$$(11) \quad \frac{dr_{t+1}}{dw_t} = \psi' = \frac{f'' \frac{\partial s}{\partial w}}{1 + n - f'' \frac{\partial s}{\partial r}} < 0.$$

9. Competitive Solution

The history of this economy can be traced in Diagram 3 containing the ψ function (relating r_{t+1} to w_t) and the ϕ function (relating w_t and r_t).

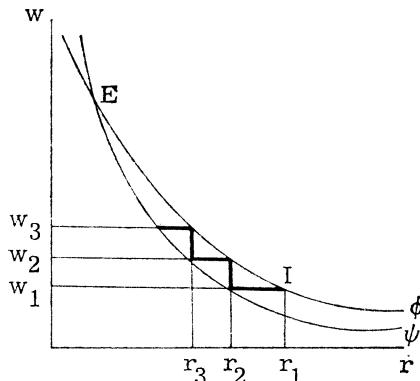


Diagram 3

Given a wage and interest pair in period one, (w_1, r_1) , which is denoted by I in Diagram 3, the interest rate in the second period can be read from the ψ curve, given the wage in period one. With this interest rate in period two, the factor-price frontier, ϕ , gives the value of the wage in period two. The entire time path of the economy can be traced out in this diagram in similar fashion.

As portrayed in Diagram 3, and as will be assumed throughout this paper, the economy has a single, stable equilibrium point. In order to derive this stability condition (which will be used to derive the direction of changes in equilibrium values when debt is introduced), one first expresses r_{t+1} as a function of r_t : $r_{t+1} = \psi(\phi(r_t))$. Taking the derivative of this, and recalling equation (11) which implies that this derivative is positive, we can express the necessary condition for stability as:

$$(12) \quad 0 < \frac{dr_{t+1}}{dr_t} = \psi' \phi' = \frac{-k f'' \frac{\partial s}{\partial w}}{1 + n - f'' \frac{\partial s}{\partial r}} \leq 1.$$

As is shown by the example in the next section, the competitive solution need not occur at an interest rate exceeding the Golden Rule level. Thus the competitive solution may be dynamically inefficient⁹ since there exists a time after which the capital-labor ratio will exceed the Golden Rule level by a nonvanishing amount.¹⁰

10. An Example

As an example, consider an economy with Cobb-Douglas production and utility functions. The utility function can be expressed as:

$$U(e^1, e^2) = \beta \log e^1 + (1 - \beta) \log e^2.$$

The saving function derived from this is independent of r :

$$s = (1 - \beta)w.$$

Thus ψ can be written:

$$r_{t+1} = f' \left(\frac{(1 - \beta)w_t}{(1 + n)} \right).$$

With production satisfying:

$$y = Ak^\alpha,$$

⁹ The possibility of an inefficient solution in an economy with infinitely many decision makers has been discussed by Koopmans [2].

¹⁰ That this implies dynamic inefficiency is proved by Phelps [7].

ψ becomes:

$$r_{t+1} = \alpha A \left(\frac{(1 - \beta)w_t}{(1 + n)} \right)^{\alpha-1},$$

while ϕ can be written:

$$w_t = (1 - \alpha)\alpha^{\alpha/1-\alpha} A^{1/1-\alpha} r_t^{\alpha/\alpha-1}.$$

Combining these we have:

$$r_{t+1} = \left(\frac{\alpha(1 + n)}{(1 - \beta)(1 - \alpha)} \right)^{1-\alpha} r_t.$$

The long-run equilibrium thus satisfies:

$$r^E = \lim_{t \rightarrow \infty} r_t = \frac{\alpha(1 + n)}{(1 - \alpha)(1 - \beta)}.$$

Except if

$$n = \frac{\alpha}{(1 - \alpha)(1 - \beta) - \alpha},$$

this does not coincide with the Golden Rule. With a positive rate of growth of labor, different economies with different values of α or β can clearly have interest rates either larger or smaller than n .

11. Framework of Analysis

In examining the long-run effects of national debt, there are two approaches that might be taken, corresponding to the two concepts of incidence, balanced-budget incidence and differential incidence.¹¹ With balanced-budget incidence, the effects of a combination of changed expenditures and changed financing are examined, weighing the relative benefits and costs. Differential incidence refers to a comparison of alternative methods of financing a given expenditure level.¹²

In this model, there are two forms which government expenditures could take, a current consumption item (which might best be viewed as lump-sum gifts to part of the populace) or government purchase of physical capital (which would then be rented to entrepreneurs for use in the production process in each future period, with the rental payments distributed to the individuals as a social dividend). Combining these two forms of expenditures with the alternatives of tax or debt

¹¹ For a detailed discussion of these concepts, see Musgrave [5].

¹² The failure to distinguish between these separate questions has been the cause of some of the confusion in the literature on the public debt. See, e.g., Mishan [3].

finance gives four possible questions of balanced-budget incidence which might be asked.

However, although answers to some of these questions will arise, analysis will be restricted to the differential incidence question of substituting debt for tax finance for a given government expenditure.

This substitution could be employed while financing the purchase of physical capital. The long-run incidence question would then be resolved by comparing the long-run equilibrium arising when there is government capital and government debt with the long-run equilibrium occurring when there is only government capital. Since the simultaneous issuance of debt and purchase of capital would merely make the government a middleman between entrepreneurs and savers, this action would have no effect on the economy in either the short or long run. Thus the initial equilibrium would be compared to that arising when there is government-owned capital, but no debt outstanding.

Alternatively, the government could finance some windfall payment (such as veterans' bonuses) either from concurrent taxes or debt issuance. While tax-financed transfer payments would have an effect in the short run (depending on the relation between the recipients and the taxpayers), since it shifts neither ϕ nor ψ it would have no effect on the long-run equilibrium. Thus the original long-run equilibrium could be compared to the one arising when debt exists (which shifts ψ) but the expenditures had no permanent effect. Either of these differential incidence frameworks would lead to the same qualitative solutions, and the second one will be adopted.

12. National Debt

To avoid the problem of expected capital gains, it will be assumed that all government debt has a one-period maturity. It will also be assumed that the debt, which is refloated each period simultaneously with the achievement of equilibrium in the capital market, pays the current interest rate. For internally held debt this assumption is necessary, given the assumption of perfect certainty, for wealth owners to be willing to hold both debt and physical capital in their portfolios. The assumption is also made for externally held debt for the sake of symmetry in the comparison of the two types of debt. The, perhaps, more natural assumption of a supply curve of external capital is discussed in Appendix B. With the assumption of a horizontal supply curve at an interest rate equaling the equilibrium domestic rate before the issuance of further debt, the qualitative results of this case are identical to those of the case considered in the text.

Since the long-run effects of the debt depend on a permanent shift in ϕ or ψ , a fixed absolute amount of debt, in a growing economy, would

asymptotically have no effect. Therefore it will be assumed that the debt-labor ratio is held constant (that the quantity of debt grows at n per cent) by financing part of the interest cost by additional debt, while financing the remainder by taxes. (It should be noted that in the case of an inefficient competitive solution, where the rate of growth exceeds the rate of interest, this implies negative taxes.) The measure of the quantity of debt outstanding in any period will be the quantity outstanding at the start of the period (or equivalently, at the time of the production process), which is therefore the quantity issued in the previous period. Thus the denominator of the debt-labor ratio refers to the number of individuals in the tax base for financing the debt, rather than the number of savers entering the capital market to purchase the debt.

The taxes employed to finance interest costs (which are paid concurrently with the receipt of factor payments) will be assumed to be lump-sum taxes on the younger generation.¹³

13. External Debt

The effects of the existence of externally held debt on the domestic economy arise solely from the taxes needed to finance that part of the interest cost not covered by increased debt. Thus we would expect the utility of an individual living at the time of long-run equilibrium to decrease because of increased taxes (in the efficient case where the interest rate exceeds the growth rate) and to change because of the change in the equilibrium wage-interest rate pair caused by the impact of these taxes on the supply side of the capital market. Denoting the external debt-labor ratio by g_1 , the taxes per worker in period t are $(r_t - n)g_1$. Therefore, the equation for the ψ function must be changed to relate savings to the wage net of taxes w_t , which equals $w_t - (r_t - n)g_1$. Rewriting the condition for equilibrium in the capital market, equation (10), we have:

$$(13) \quad r_{t+1} = f' \left(\frac{s(w_t - (r_t - n)g_1, r_{t+1})}{1 + n} \right).$$

The new form of this equation implies a new stability condition which, together with the assumption on relative slopes in the capital market, is expressed in (14). (It is assumed that there is a single stable equilibrium both with and without the debt.)

¹³ The case with lump-sum taxes on the older generation is equivalent to that with lump-sum taxes on the younger generation plus intergeneration transfers in each period. This transfer scheme (from old to young) increases saving in anticipation of taxes and counteracts the decreases described in the text, although not fully.

$$(14) \quad 0 < \frac{dr_{t+1}}{dr_t} = \frac{-f''(k + g_1) \frac{\partial s}{\partial w}}{1 + n - f'' \frac{\partial s}{\partial r}} \leq 1.$$

To examine the shift in the ψ curve, we can implicitly differentiate equation (13) again, this time taking the partial derivative of r_{t+1} with respect to g_1 .

$$(15) \quad \frac{\partial r_{t+1}}{\partial g_1} = \frac{f''(n - r_t) \frac{\partial s}{\partial w}}{1 + n - f'' \frac{\partial s}{\partial r}}.$$

From equation (14) we know that the sign of this expression is the same as that of $(r - n)$. Geometrically we have ψ shifting to the new curve ψ' as shown in Diagram 4. By combining the new ψ curve with

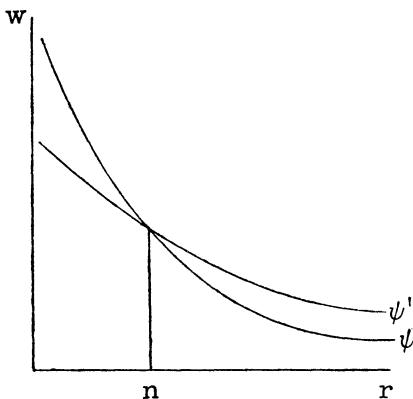


Diagram 4

the factor price frontier, which is unchanged, we can examine the change in the long-run equilibrium values of r and w . In Diagram 5 we see that, if the equilibrium interest rate was unequal to the growth rate, the existence of external debt increases the difference between the two.¹⁴

To examine the effects of external debt on utility levels in long-run equilibria, it is simplest to assume a given level of external debt and examine the changes arising from a derivative change in this quantity.

¹⁴ Since, if $r = n$, additional debt issuance exactly covers interest payments, if this were the original equilibrium, the debt has no effect.

Using equation (13) and the constancy of the interest rate in long-run equilibrium, one can express the equilibrium interest rate as an implicit function of the quantity of debt outstanding:

$$(16) \quad r = f' \left(\frac{s(\phi(r) - (r - n)g_1, r)}{1 + n} \right).$$

From this relationship one can derive the change in the equilibrium interest rate arising from a change in the debt-labor ratio:

$$(17) \quad \frac{dr}{dg_1} = \frac{-f''(r - n) \frac{\partial s}{\partial w}}{1 + n - f'' \frac{\partial s}{\partial r} + f''(k + g_1) \frac{\partial s}{\partial w}}.$$

As was described above, external debt moves the interest rate away from the Golden Rule solution. In terms of the capital market, we have that positive taxes, by decreasing the supply of capital, given any level of the wage, increase the equilibrium interest rate.

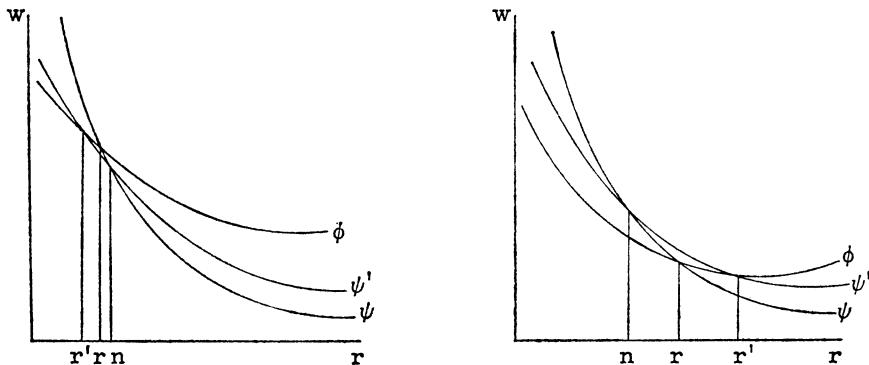


Diagram 5

The change in the utility level can be calculated by employing the expressions for the effects of changes in factor payments on the utility level, equation (7).

$$(18) \quad \frac{dU}{dg_1} = \frac{d\hat{w}}{dg_1} \frac{\partial U}{\partial w} + \frac{dr}{dg_1} \frac{\partial U}{\partial r} = \frac{\partial U}{\partial e^1} \left[\frac{d\hat{w}}{dg_1} + \frac{s}{(1+r)} \frac{dr}{dg_1} \right].$$

From the expression for the net wage, $\hat{w} = w - (r - n)g_1$, one can calculate the change in the net wage in terms of the change in the interest rate:

$$(19) \quad \frac{d\hat{w}}{dg_1} = -(k + g_1) \frac{dr}{dg_1} - (r - n).$$

Substituting this equation in the previous equation we have:

$$(20) \quad \frac{dU}{dg_1} = -\frac{\partial U}{\partial e^1} \left[(r - n) + g_1 \frac{dr}{dg_1} + \left(k - \frac{s}{1+r} \right) \frac{dr}{dg_1} \right].$$

The first term of this expression is the change in utility arising from the taxes needed to finance the addition to the outstanding debt, and is positive or negative as these taxes are positive or negative. The second term describes the change in the tax burden of existing debt occurring because of the change in the interest rate. Thus, both of these utility changes are positive if r exceeds n and negative if n exceeds r .

The third term can be explained by means of Diagram 6 containing derived indifference curves between w and r , denoted by II , and the factor-price frontier.¹⁵

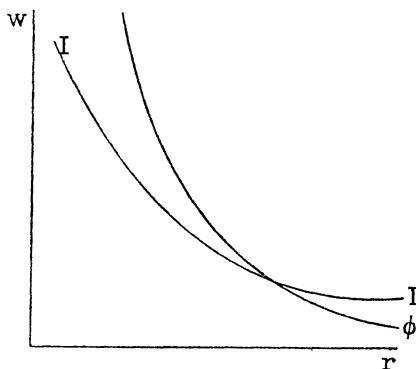


Diagram 6

The change in the interest rate is a movement along the factor-price frontier. The change in utility thus depends on the relative slopes of the factor-price frontier, the slope of which is $-k$, and of the indifference curve, the slope of which is $-s/(1+r)$. From the equilibrium condition for the capital market, $k=s/(1+n)$, this term can be rewritten as

$$\frac{dr}{dg_1} \left(\frac{k}{1+r} \right) (r - n).$$

Since dr/dg_1 has the same sign as $(r-n)$, the movement of the interest rate away from the Golden Rule level causes the utility from factor payments to fall.

Combining the three effects, we can conclude that in the "normal" case where the competitive solution is efficient, external debt causes a

¹⁵ A rigorous treatment of this approach would require changes in the indifference levels because of taxes.

fall in the utility level of an individual living in long-run equilibrium. If the competitive solution is inefficient the effects of the debt work in opposite directions and so yield no a priori conclusion.

14. Internal Debt

With internal debt, the supply side of the capital market is altered in precisely the same fashion as with external debt, since an individual taxpayer is in the same position as a taxpayer whether his tax payments flow abroad or remain in the country. Denoting the internal debt-labor ratio by g_2 , the savings function must be altered as before to read

$$s(w_t - (r_t - n)g_2, r_{t+1}).$$

It is also necessary to alter the equilibrium condition in the capital market to take account of the fact that the government enters on the demand side of this market. Denoting the quantity of internal debt to be floated in period t (and repaid in $t+1$) by G_{t+1} , the equilibrium condition becomes:

$$(21) \quad S_t = K_{t+1} + G_{t+1}.$$

Dividing this by L_{t+1} , we have the equilibrium condition expressed in terms of the ratios needed to describe the equilibrium:

$$(22) \quad \frac{s_t}{1+n} = k_{t+1} + g_2.$$

Comparing internal with external debt, we see that they both require taxes to be paid by each worker, while internal debt has a further effect in that it substitutes pieces of paper for physical capital in the portfolios of wealth owners, thus reducing output.

Recalling that the demand for capital by entrepreneurs is determined by the marginal productivity schedule of capital, we can combine this equation with equation (22) to obtain the new condition for equilibrium in the capital market:

$$(23) \quad r_{t+1} = f' \left(\frac{s(w_t - (r_t - n)g_2, r_{t+1})}{1+n} - g_2 \right).$$

As before, by implicit differentiation of this equation we can express the conditions for stability and the assumed slopes in the capital market:

$$(24) \quad 0 < \frac{dr_{t+1}}{dr_t} = \frac{-f''(k + g_2) \frac{\partial s}{\partial w}}{1 + n - f'' \frac{\partial s}{\partial r}} \leq 1.$$

To find the shift in ψ , we take the partial derivative of r with respect to g_2 :

$$(25) \quad \frac{\partial r_{t+1}}{\partial g_2} = \frac{-f''\left(\frac{\partial s}{\partial w}(r - n) + (1 + n)\right)}{1 + n - f''\frac{\partial s}{\partial r}}.$$

From (24) and the normality of present consumption, $\partial s/\partial w < 1$, we know that this expression is positive and thus that ψ shifts upward for all values of r .

Following the same analysis as with external debt, we can calculate the change in utility arising from the change in the level of internal debt. We write first the locus of equilibria for different quantities of debt:

$$(26) \quad r = f'\left(\frac{s(\phi(r) - (r - n)g_2, r)}{1 + n} - g_2\right).$$

We can then differentiate this expression with respect to g_2 to obtain the change in the equilibrium interest rate arising from the change in the quantity of debt:

$$(27) \quad \frac{dr}{dg_2} = \frac{-f''\left(1 + n + (r - n)\frac{\partial s}{\partial w}\right)}{1 + n - f''\frac{\partial s}{\partial r} + f''(k + g_2)\frac{\partial s}{\partial w}}.$$

Again, from equation (24), we know that the change in the equilibrium interest rate is positive. Employing equations (18) and (19) relating changes in utility and the net wage to changes in debt (which hold for either internal or external debt) and the equilibrium condition for the capital market, equation (22), we can express the changes in utility in two ways:

$$(28) \quad \frac{dU}{dg_2} = -\frac{\partial U}{\partial e^1}\left[(r - n) + g_2\frac{dr}{dg_2} + \left(k - \frac{s}{1 + r}\right)\frac{dr}{dg_2}\right]$$

$$(29) \quad \frac{dU}{dg_2} = -\frac{\partial U}{\partial e^1}(r - n)\left[1 + \frac{k + g_2}{1 + r}\frac{dr}{dg_2}\right].$$

Equation (28) expresses the change in utility in terms of the taxes needed to finance the increase in debt, the taxes needed to finance the increased interest payments on existing debt, and the changed value of factor payments. As before, the sign of the first two terms depends

solely on $(r-n)$. However, since $s = (1+n)(k+g_2)$, the third term, while decreasing utility when r is smaller than n , may increase or decrease utility when the competitive solution is efficient.

Equation (29), which combines the separate terms, shows that utility is decreased in the efficient case and increased in the inefficient case. Like the third term in the expression giving the change in utility from external debt, the sign of this expression is the opposite of the sign of $(r-n)dr/dg$. Separating the effects of debt issuance into those which alter the social consumption possibilities (the flow of interest payments abroad) and those that reflect a change in the allocation of consumption within the society, the total impact of all effects falling into the second group will increase or decrease utility as the interest rate is moved toward or away from the rate of growth. This is demonstrated geometrically in the next section.

15. Diagrammatic Discussion

The assumption that both present and future consumption are normal goods implies that, as one moves upward and to the left along a budget line, the indifference curves become steeper. Geometrically, this implies that the slope at A is algebraically greater than at B . See Diagram 7.

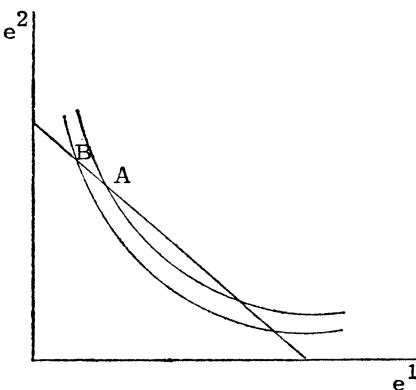


Diagram 7

Recalling equation (5), the consumption possibilities for a society with a given capital-labor ratio lie along the line $e^1 + e^2/(1+n) = y - nk$. Since the interest rate is the marginal product of capital in a competitive society, from the interest rate we know the height of the consumption constraint line. Furthermore, since consumption is allocated over time in accordance with the market interest rate, we know that the

competitive equilibrium occurs where the slope of an indifference curve equals $-(1+r)$. These two facts permit us to locate a competitive equilibrium in Diagram 7, knowing just the equilibrium interest rate (and, of course, the production function). Since internal debt does not alter the consumption possibilities available to an economy, the utility associated with the equilibrium arising from varying quantities of internal debt can be located in this diagram. (This, of course, is not true for external debt.)

Combining these two facets of a change in the interest rate we can conclude that any movement of the interest rate away from the growth rate decreases utility first by diminishing the height of the consumption constraint line and second by moving along the lowered line in the direction of decreased utility. Assuming $r' > r > n$, this is shown in Diagram 8 where A is the equilibrium point associated with r while C is the one associated with r' .

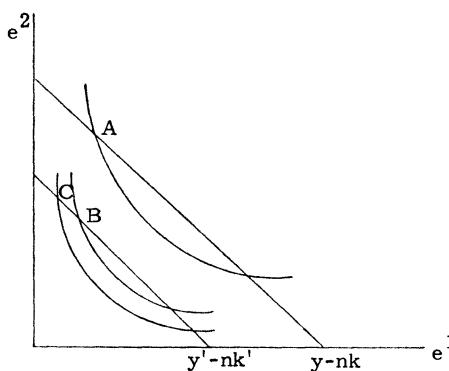


Diagram 8

Utility at B (where the slope of the indifference curve is the same as at A) is less than utility at A since B is on a lower constraint line. The slope at C , which equals $-(1+r')$, is less than at B , $-(1+r)$, implying a lower utility level at C than at B .

Thus internal debt raises or lowers the utility level as it moves the equilibrium interest rate towards or away from the growth rate. External debt has two effects, an alteration of consumption possibilities due to the flow abroad of interest payments and an alteration of utility arising from changes in the interest rate, given the level of interest payments, which is positive or negative as the interest rate is moved toward or away from the growth rate.

Thus the third term in dU/dg_1 , which is the "purely domestic" effect of issuing external debt corresponds to the entire effects of internal debt, with the sign of the expression equalling that of $-(r-n)dr/dg$.

16. Internal and External Debt

Having described the way each of them affects the equilibrium of the economy, it is now possible to turn to the complete model, in which there is both external and internal debt, and so make a direct comparison of their effects.¹⁶

Without stopping to repeat the analysis step by step, we can write down the relevant equations from the equations derived in the last three sections:

The condition for equilibrium in the capital market:

$$(30) \quad r_{t+1} = f' \left(\frac{s(w_t - (r_t - n)(g_1 + g_2), r_{t+1})}{1 + n} - g_2 \right).$$

The locus of long-run competitive equilibria with different quantities of debt outstanding:

$$(31) \quad r = f' \left(\frac{s(\phi(r) - (r - n)(g_1 + g_2), r)}{1 + n} - g_2 \right).$$

The necessary condition for stability and the assumption on the demand and supply curves for capital:

$$(32) \quad 0 < \frac{dr_{t+1}}{dr_t} = \frac{-f''(k + g_1 + g_2) \frac{\partial s}{\partial w}}{1 + n - f'' \frac{\partial s}{\partial r}} \leq 1.$$

The change in the equilibrium interest rate arising from changes in debt:

$$(33) \quad \begin{aligned} \frac{dr}{dg_1} &= \frac{-f''(r - n) \frac{\partial s}{\partial w}}{1 + n - f'' \frac{\partial s}{\partial r} + f''(k + g_1 + g_2) \frac{\partial s}{\partial w}} \\ \frac{dr}{dg_2} &= \frac{-f'' \left((r - n) \frac{\partial s}{\partial w} + (1 + n) \right)}{1 + n - f'' \frac{\partial s}{\partial r} + f''(k + g_1 + g_2) \frac{\partial s}{\partial w}}. \end{aligned}$$

¹⁶ In Section 14, the change from external to internal debt was described as adding the effect arising from the substitution of paper for physical capital in portfolios. Reversing this comparison, external debt is internal debt plus annual foreign borrowing, with foreign capital receiving its marginal product. This does not directly alter net output, but it does alter relative factor prices which directly affects utility and which alters savings.

The changes in utility arising from changes in the quantity of debt:

$$(34) \quad \begin{aligned} \frac{dU}{dg_1} &= -\frac{\partial U}{\partial e^1} \left((r - n) + \frac{dr}{dg_1} (g_1 + g_2) + \frac{dr}{dg_1} \left(k - \frac{s}{1+r} \right) \right) \\ \frac{dU}{dg_2} &= -\frac{\partial U}{\partial e^1} \left((r - n) + \frac{dr}{dg_2} (g_1 + g_2) + \frac{dr}{dg_2} \left(k - \frac{s}{1+r} \right) \right). \end{aligned}$$

With these relations before us, it is possible to examine the differential incidence question arising from the effects of issuing internal debt to retire external debt, and to examine the relationships between some of the articles in the literature on the burden of the debt.

From equation (33) we can calculate the change in the equilibrium interest rate arising from this debt swap:

$$(35) \quad \frac{dr}{dg_2} - \frac{dr}{dg_1} = \frac{-f''(1+n)}{1+n-f''\frac{\partial s}{\partial r}+f''(k+g_1+g_2)\frac{\partial s}{\partial w}}.$$

From the stability condition, (32), we know that the denominator of this expression is positive and thus that the interest rate always rises. The effect of the debt swap involves no change in taxes, and so no change in the supply side of the capital market. However, the demand side is altered by the increase in government demand, causing a rise in the equilibrium interest rate and a fall in the capital-labor ratio.

The change in utility can be derived from (34) and expressed in different ways:

$$(36) \quad \frac{dU}{dg_2} - \frac{dU}{dg_1} = -\frac{\partial U}{\partial e^1} \left[\frac{dr}{dg_2} - \frac{dr}{dg_1} \right] \left[(g_1 + g_2) + \left(k - \frac{s}{1+r} \right) \right],$$

$$(37) \quad \begin{aligned} \frac{dU}{dg_2} - \frac{dU}{dg_1} &= -\frac{\partial U}{\partial e^1} \left(\left[-(r-n) + \left(\frac{dr}{dg_2} - \frac{dr}{dg_1} \right) g_1 \right] \right. \\ &\quad \left. + (r-n) \left[1 + \left(\frac{dr}{dg_2} - \frac{dr}{dg_1} \right) \frac{(k+g_2)}{(1+r)} \right] \right), \end{aligned}$$

$$(38) \quad \frac{dU}{dg_2} - \frac{dU}{dg_1} = -\frac{\partial U}{\partial e^1} \left[\frac{dr}{dg_2} - \frac{dr}{dg_1} \right] \left[\frac{(k+g_2)(r-n)}{(1+r)} + g_1 \right].$$

Equation (36) divides the utility change into the part arising from the change in taxes and the part arising from the change in the utility of factor payments. Since the interest rate rises, taxes must rise, lowering utility. However, since

$$\left(k - \frac{s}{(1+r)} \right) \text{ is equal to } \frac{(r-n)k - (1+n)g_2}{1+r},$$

as in the discussion of internal debt, the change in utility coming from the change in factor payments may be positive or negative.

Equation (37) divides the utility change into the part arising from the change in the external interest payments, which may be positive or negative, and the part arising, as in the last section, domestically, from the change in equilibrium values, given the level of external payments. This term has the sign

$$-(r - n) \left(\frac{dr}{dg_2} - \frac{dr}{dg_1} \right),$$

which is, therefore, the sign of $n - r$. Thus, as before, for this term the rise in interest rates raises utility in the inefficient case but lowers it in the efficient case.

The third form of the equation, (38), is expressed to most easily give the sign of the utility change. When the solution is efficient, we have an unambiguous fall in utility from this debt swap. In the inefficient case the sign depends on the relative sizes of

$$g_1, \text{ and } \frac{(k + g_2)(r - n)}{(1 + r)}.$$

17. Conclusion

Thus we have seen that, where both types of debt exist, internal debt, which raises the interest rate, lowers utility in the efficient case but may raise or lower it in the inefficient case (if there were no external debt, utility would be raised in the inefficient case). External debt, which moves the interest rate away from the growth rate, lowers utility in the efficient case and may raise or lower it in the inefficient case (this remains true whether or not internal debt exists). Finally, the substitution of internal for external debt, which raises the interest rate, lowers utility in the efficient case, while being capable of raising or lowering it in the inefficient case.

There are two ways of classifying the effects of external and internal debt which shed some light on some of the effects described in the literature.

First, as in equation (34), they can be divided into utility changes arising from changes in taxes paid and from a change in the relative factor payments. This division shows that the taxes needed to finance either internal or external debt have the same impact on individuals living during long-run equilibrium.

Second, the change in utility from internal debt can be separated into the effects of external debt plus the effects of a debt swap. This

would imply four effects, the effects of the two changes in taxes, and the two effects on factor payments. These latter two effects can be distinguished by the fact that external debt affects only the supply side of the capital market, while the debt swap affects only the demand side.

In their discussion of the effects of debt, Bowen, Davis, and Kopf [1] concentrated on the tax effects of internal debt, and so described the first two of these four effects.

Modigliani [4] and Vickrey [10] discussed the fall in the capital stock arising from the substitution of debt for capital in the portfolios of wealth owners. As such, they were discussing the change in the demand side of the capital market¹⁷ and the effects described are additive to those arising from taxes. It is only necessary to add the effects of taxes on the capital stock (and thus on factor payments) to complete the discussion.

APPENDIX A

Making the alternative assumption on the capital market, which, together with the stability condition, can be expressed as:

$$-1 \leq \frac{dr_{t+1}}{dr_t} < 0,$$

we can re-examine the signs of the equations in Section 16. The denominators of the expressions giving the change in the equilibrium interest rate are now negative (where they were positive in the text). Thus external debt moves the interest rate toward the growth rate, while internal debt lowers its equilibrium value. Consequently the debt swap lowers the rate of interest.

Therefore, in the efficient case, increased debt causes positive taxes for the additional debt but lowers the taxes on existing debt and so may raise or lower the utility level. The debt swap raises utility by decreasing taxes

¹⁷ Modigliani described a one-for-one replacement of capital by debt, assuming that total wealth remained constant. However, the fall in the capital stock, which causes a fall in output, would affect the equilibrium quantity of total wealth. (Modigliani acknowledges this but ignores its effects.) The change in the capital stock can be derived from equation (33):

$$\frac{dk}{dg_1} - \frac{dk}{dg_2} = \frac{1}{f''} \left(\frac{dr}{dg_1} - \frac{dr}{dg_2} \right) = \frac{-1-n}{1+n-f'' \left(\frac{\partial s}{\partial r} - (k+g_1+g_2) \frac{\partial s}{\partial w} \right)}.$$

This differs from -1 because of the term

$$\left(\frac{\partial s}{\partial r} - (k+g_1+g_2) \frac{\partial s}{\partial w} \right).$$

This latter expression represents the partial effect on desired wealth (which is equal to savings) arising from the fall in the capital stock:

$$\frac{\partial s}{\partial k} = \frac{\partial s}{\partial r} \frac{\partial r}{\partial k} + \frac{\partial s}{\partial w} \frac{\partial w}{\partial k} = f'' \frac{\partial s}{\partial r} + \frac{\partial s}{\partial w} f'' \frac{\partial w}{\partial r} = f'' \left(\frac{\partial s}{\partial r} - (k+g_1+g_2) \frac{\partial s}{\partial w} \right).$$

and increasing the utility of factor payments by moving the interest rate toward the Golden Rule level.

APPENDIX B

Assuming a supply curve of external debt which can be expressed by writing the interest rate, ρ , as a function of g_1 (this assumes that the source of foreign capital is also increasing at n per cent per year), the net wage can be expressed:

$$\hat{w} = w - (\rho - n)g_1 - (r - n)g_2.$$

This implies that the change in the net wage can be expressed as:

$$\begin{aligned}\frac{d\hat{w}}{dg_1} &= - (k + g_2) \frac{dr}{dg_1} - (\rho - n) - g_1 \frac{d\rho}{dg_1}, \\ \frac{d\hat{w}}{dg_2} &= - (k + g_2) \frac{dr}{dg_2} - (r - n).\end{aligned}$$

Thus the utility change becomes:

$$\begin{aligned}\frac{dU}{dg_1} &= - \frac{\partial U}{\partial e^1} \left(\rho - n + g_1 \frac{d\rho}{dg_1} + \frac{dr}{dg_1} \left(k + g_2 - \frac{s}{1+r} \right) \right), \\ \frac{dU}{dg_2} &= - \frac{\partial U}{\partial e^1} \left((r - n) + \frac{dr}{dg_2} \left(k + g_2 - \frac{s}{1+r} \right) \right).\end{aligned}$$

If the supply curve of capital is horizontal at the prevailing internal interest rate, these two expressions differ from equation (34) only in the disappearance of the term $g_1(dr/dg)$ from both equations (and the somewhat different interest rate derivative). Thus the effect of a debt swap becomes:

$$\frac{dU}{dg_2} - \frac{dU}{dg_1} = - \frac{\partial U}{\partial e^1} \left(\frac{dr}{dg_2} - \frac{dr}{dg_1} \right) \left(k + g_2 - \frac{s}{1+r} \right),$$

which depends in sign solely on whether the difference between the interest and growth rates is increased.

The change in the equilibrium interest rates can be derived from the locus of equilibria:

$$\begin{aligned}r &= f' \left(\frac{s(w - (\rho - n)g_1 - (r - n)g_2, r)}{1+n} - g_2 \right) \\ \frac{dr}{dg_1} &= \frac{-f'' \frac{\partial s}{\partial w} \left((\rho - n) + g_1 \frac{d\rho}{dg_1} \right)}{1+n - f'' \frac{\partial s}{\partial r} + f''(k+g_2) \frac{\partial s}{\partial w}}\end{aligned}$$

$$\frac{dr}{dg_2} = \frac{-f'' \left(1 + n + (r - n) \frac{\partial s}{\partial w} \right)}{1 + n - f'' \frac{\partial s}{\partial w} + f''(k + g_2) \frac{\partial s}{\partial w}}$$

Again assuming that $\rho=r$ and $d\rho/dg=0$, these derivatives are qualitatively the same as those described in the text.

REFERENCES

1. W. G. BOWEN, R. G. DAVIS, and D. H. KOPF, "The Public Debt: A Burden on Future Generations?" *Am. Econ. Rev.*, Sept. 1960, 50, 701-6.
2. T. C. KOOPMANS, *Three Essays on the State of the Economic Science*. New York 1957.
3. E. J. MISHAN, "How to Make a Burden of the Public Debt," *Jour. Pol. Econ.*, Dec. 1963, 71, 529-42.
4. F. MODIGLIANI, "Long-Run Implications of Alternative Fiscal Policies and the Burden of the National Debt," *Econ. Jour.*, Dec. 1961, 71, 730-55.
5. R. A. MUSGRAVE, *The Theory of Public Finance*. New York 1959.
6. E. S. PHELPS, "The Golden Rule of Accumulation: A Fable for Growthmen," *Am. Econ. Rev.*, Sept. 1961, 51, 638-43.
7. ———, "Second Essay on the Golden Rule," Cowles Foundation Discussion Paper 173.
8. P. A. SAMUELSON, "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money," *Jour. Pol. Econ.*, Dec. 1958, 66, 467-82.
9. ———, "Parable and Realism in Capital Theory: The Surrogate Production Function," *Rev. Econ. Stud.*, June 1962, 29, 193-206.
10. W. VICKREY, "The Burden of the Public Debt: Comment," *Am. Econ. Rev.*, March 1961, 51, 132-37.

The American Economic Review

Volume LVIII

MARCH 1968

Number 1

THE ROLE OF MONETARY POLICY*

*By MILTON FRIEDMAN***

There is wide agreement about the major goals of economic policy: high employment, stable prices, and rapid growth. There is less agreement that these goals are mutually compatible or, among those who regard them as incompatible, about the terms at which they can and should be substituted for one another. There is least agreement about the role that various instruments of policy can and should play in achieving the several goals.

My topic for tonight is the role of one such instrument—monetary policy. What can it contribute? And how should it be conducted to contribute the most? Opinion on these questions has fluctuated widely. In the first flush of enthusiasm about the newly created Federal Reserve System, many observers attributed the relative stability of the 1920s to the System's capacity for fine tuning—to apply an apt modern term. It came to be widely believed that a new era had arrived in which business cycles had been rendered obsolete by advances in monetary technology. This opinion was shared by economist and layman alike, though, of course, there were some dissonant voices. The Great Contraction destroyed this naive attitude. Opinion swung to the other extreme. Monetary policy was a string. You could pull on it to stop inflation but you could not push on it to halt recession. You could lead a horse to water but you could not make him drink. Such theory by aphorism was soon replaced by Keynes' rigorous and sophisticated analysis.

Keynes offered simultaneously an explanation for the presumed impotence of monetary policy to stem the depression, a nonmonetary interpretation of the depression, and an alternative to monetary policy

* Presidential address delivered at the Eightieth Annual Meeting of the American Economic Association, Washington, D.C., December 29, 1967.

** I am indebted for helpful criticisms of earlier drafts to Armen Alchian, Gary Becker, Martin Bronfenbrenner, Arthur F. Burns, Phillip Cagan, David D. Friedman, Lawrence Harris, Harry G. Johnson, Homer Jones, Jerry Jordan, David Meiselman, Allan H. Meltzer, Theodore W. Schultz, Anna J. Schwartz, Herbert Stein, George J. Stigler, and James Tobin.

for meeting the depression and his offering was avidly accepted. If liquidity preference is absolute or nearly so—as Keynes believed likely in times of heavy unemployment—interest rates cannot be lowered by monetary measures. If investment and consumption are little affected by interest rates—as Hansen and many of Keynes' other American disciples came to believe—lower interest rates, even if they could be achieved, would do little good. Monetary policy is twice damned. The contraction, set in train, on this view, by a collapse of investment or by a shortage of investment opportunities or by stubborn thriftiness, could not, it was argued, have been stopped by monetary measures. But there was available an alternative—fiscal policy. Government spending could make up for insufficient private investment. Tax reductions could undermine stubborn thriftiness.

The wide acceptance of these views in the economics profession meant that for some two decades monetary policy was believed by all but a few reactionary souls to have been rendered obsolete by new economic knowledge. Money did not matter. Its only role was the minor one of keeping interest rates low, in order to hold down interest payments in the government budget, contribute to the "euthanasia of the rentier," and maybe, stimulate investment a bit to assist government spending in maintaining a high level of aggregate demand.

These views produced a widespread adoption of cheap money policies after the war. And they received a rude shock when these policies failed in country after country, when central bank after central bank was forced to give up the pretense that it could indefinitely keep "the" rate of interest at a low level. In this country, the public denouement came with the Federal Reserve-Treasury Accord in 1951, although the policy of pegging government bond prices was not formally abandoned until 1953. Inflation, stimulated by cheap money policies, not the widely heralded postwar depression, turned out to be the order of the day. The result was the beginning of a revival of belief in the potency of monetary policy.

This revival was strongly fostered among economists by the theoretical developments initiated by Haberler but named for Pigou that pointed out a channel—namely, changes in wealth—whereby changes in the real quantity of money can affect aggregate demand even if they do not alter interest rates. These theoretical developments did not undermine Keynes' argument against the potency of orthodox monetary measures when liquidity preference is absolute since under such circumstances the usual monetary operations involve simply substituting money for other assets without changing total wealth. But they did show how changes in the quantity of money produced in other ways could affect total spending even under such circumstances. And, more

fundamentally, they did undermine Keynes' key theoretical proposition, namely, that even in a world of flexible prices, a position of equilibrium at full employment might not exist. Henceforth, unemployment had again to be explained by rigidities or imperfections, not as the natural outcome of a fully operative market process.

The revival of belief in the potency of monetary policy was fostered also by a re-evaluation of the role money played from 1929 to 1933. Keynes and most other economists of the time believed that the Great Contraction in the United States occurred despite aggressive expansionary policies by the monetary authorities—that they did their best but their best was not good enough.¹ Recent studies have demonstrated that the facts are precisely the reverse: the U.S. monetary authorities followed highly deflationary policies. The quantity of money in the United States fell by one-third in the course of the contraction. And it fell not because there were no willing borrowers—not because the horse would not drink. It fell because the Federal Reserve System forced or permitted a sharp reduction in the monetary base, because it failed to exercise the responsibilities assigned to it in the Federal Reserve Act to provide liquidity to the banking system. The Great Contraction is tragic testimony to the power of monetary policy—not, as Keynes and so many of his contemporaries believed, evidence of its impotence.

In the United States the revival of belief in the potency of monetary policy was strengthened also by increasing disillusionment with fiscal policy, not so much with its potential to affect aggregate demand as with the practical and political feasibility of so using it. Expenditures turned out to respond sluggishly and with long lags to attempts to adjust them to the course of economic activity, so emphasis shifted to taxes. But here political factors entered with a vengeance to prevent prompt adjustment to presumed need, as has been so graphically illustrated in the months since I wrote the first draft of this talk. "Fine tuning" is a marvelously evocative phrase in this electronic age, but it has little resemblance to what is possible in practice—not, I might add, an unmixed evil.

It is hard to realize how radical has been the change in professional opinion on the role of money. Hardly an economist today accepts views that were the common coin some two decades ago. Let me cite a few examples.

In a talk published in 1945, E. A. Goldenweiser, then Director of the Research Division of the Federal Reserve Board, described the primary objective of monetary policy as being to "maintain the value of Government bonds. . . . This country" he wrote, "will have to adjust to

¹ In [2], I have argued that Henry Simons shared this view with Keynes, and that it accounts for the policy changes that he recommended.

a 2½ per cent interest rate as the return on safe, long-time money, because the time has come when returns on pioneering capital can no longer be unlimited as they were in the past" [4, p. 117].

In a book on *Financing American Prosperity*, edited by Paul Homan and Fritz Machlup and published in 1945, Alvin Hansen devotes nine pages of text to the "savings-investment problem" without finding any need to use the words "interest rate" or any close facsimile thereto [5, pp. 218-27]. In his contribution to this volume, Fritz Machlup wrote, "Questions regarding the rate of interest, in particular regarding its variation or its stability, may not be among the most vital problems of the postwar economy, but they are certainly among the perplexing ones" [5, p. 466]. In his contribution, John H. Williams—not only professor at Harvard but also a long-time adviser to the New York Federal Reserve Bank—wrote, "I can see no prospect of revival of a general monetary control in the postwar period" [5, p. 383].

Another of the volumes dealing with postwar policy that appeared at this time, *Planning and Paying for Full Employment*, was edited by Abba P. Lerner and Frank D. Graham [6] and had contributors of all shades of professional opinion—from Henry Simons and Frank Graham to Abba Lerner and Hans Neisser. Yet Albert Halasi, in his excellent summary of the papers, was able to say, "Our contributors do not discuss the question of money supply. . . . The contributors make no special mention of credit policy to remedy actual depressions. . . . Inflation . . . might be fought more effectively by raising interest rates. . . . But . . . other anti-inflationary measures . . . are preferable" [6, pp. 23-24]. *A Survey of Contemporary Economics*, edited by Howard Ellis and published in 1948, was an "official" attempt to codify the state of economic thought of the time. In his contribution, Arthur Smithies wrote, "In the field of compensatory action, I believe fiscal policy must shoulder most of the load. Its chief rival, monetary policy, seems to be disqualified on institutional grounds. This country appears to be committed to something like the present low level of interest rates on a long-term basis" [1, p. 208].

These quotations suggest the flavor of professional thought some two decades ago. If you wish to go further in this humbling inquiry, I recommend that you compare the sections on money—when you can find them—in the Principles texts of the early postwar years with the lengthy sections in the current crop even, or especially, when the early and recent Principles are different editions of the same work.

The pendulum has swung far since then, if not all the way to the position of the late 1920s, at least much closer to that position than to the position of 1945. There are of course many differences between then and now, less in the potency attributed to monetary policy than in the

roles assigned to it and the criteria by which the profession believes monetary policy should be guided. Then, the chief roles assigned monetary policy were to promote price stability and to preserve the gold standard; the chief criteria of monetary policy were the state of the "money market," the extent of "speculation" and the movement of gold. Today, primacy is assigned to the promotion of full employment, with the prevention of inflation a continuing but definitely secondary objective. And there is major disagreement about criteria of policy, varying from emphasis on money market conditions, interest rates, and the quantity of money to the belief that the state of employment itself should be the proximate criterion of policy.

I stress nonetheless the similarity between the views that prevailed in the late 'twenties and those that prevail today because I fear that, now as then, the pendulum may well have swung too far, that, now as then, we are in danger of assigning to monetary policy a larger role than it can perform, in danger of asking it to accomplish tasks that it cannot achieve, and, as a result, in danger of preventing it from making the contribution that it is capable of making.

Unaccustomed as I am to denigrating the importance of money, I therefore shall, as my first task, stress what monetary policy cannot do. I shall then try to outline what it can do and how it can best make its contribution, in the present state of our knowledge—or ignorance.

I. What Monetary Policy Cannot Do

From the infinite world of negation, I have selected two limitations of monetary policy to discuss: (1) It cannot peg interest rates for more than very limited periods; (2) It cannot peg the rate of unemployment for more than very limited periods. I select these because the contrary has been or is widely believed, because they correspond to the two main unattainable tasks that are at all likely to be assigned to monetary policy, and because essentially the same theoretical analysis covers both.

Pegging of Interest Rates

History has already persuaded many of you about the first limitation. As noted earlier, the failure of cheap money policies was a major source of the reaction against simple-minded Keynesianism. In the United States, this reaction involved widespread recognition that the wartime and postwar pegging of bond prices was a mistake, that the abandonment of this policy was a desirable and inevitable step, and that it had none of the disturbing and disastrous consequences that were so freely predicted at the time.

The limitation derives from a much misunderstood feature of the relation between money and interest rates. Let the Fed set out to keep

interest rates down. How will it try to do so? By buying securities. This raises their prices and lowers their yields. In the process, it also increases the quantity of reserves available to banks, hence the amount of bank credit, and, ultimately the total quantity of money. That is why central bankers in particular, and the financial community more broadly, generally believe that an increase in the quantity of money tends to lower interest rates. Academic economists accept the same conclusion, but for different reasons. They see, in their mind's eye, a negatively sloping liquidity preference schedule. How can people be induced to hold a larger quantity of money? Only by bidding down interest rates.

Both are right, up to a point. The *initial* impact of increasing the quantity of money at a faster rate than it has been increasing is to make interest rates lower for a time than they would otherwise have been. But this is only the beginning of the process not the end. The more rapid rate of monetary growth will stimulate spending, both through the impact on investment of lower market interest rates and through the impact on other spending and thereby relative prices of higher cash balances than are desired. But one man's spending is another man's income. Rising income will raise the liquidity preference schedule and the demand for loans; it may also raise prices, which would reduce the real quantity of money. These three effects will reverse the initial downward pressure on interest rates fairly promptly, say, in something less than a year. Together they will tend, after a somewhat longer interval, say, a year or two, to return interest rates to the level they would otherwise have had. Indeed, given the tendency for the economy to overreact, they are highly likely to raise interest rates temporarily beyond that level, setting in motion a cyclical adjustment process.

A fourth effect, when and if it becomes operative, will go even farther, and definitely mean that a higher rate of monetary expansion will correspond to a higher, not lower, level of interest rates than would otherwise have prevailed. Let the higher rate of monetary growth produce rising prices, and let the public come to expect that prices will continue to rise. Borrowers will then be willing to pay and lenders will then demand higher interest rates—as Irving Fisher pointed out decades ago. This price expectation effect is slow to develop and also slow to disappear. Fisher estimated that it took several decades for a full adjustment and more recent work is consistent with his estimates.

These subsequent effects explain why every attempt to keep interest rates at a low level has forced the monetary authority to engage in successively larger and larger open market purchases. They explain why, historically, high and rising nominal interest rates have been associated

with rapid growth in the quantity of money, as in Brazil or Chile or in the United States in recent years, and why low and falling interest rates have been associated with slow growth in the quantity of money, as in Switzerland now or in the United States from 1929 to 1933. As an empirical matter, low interest rates are a sign that monetary policy *has been* tight—in the sense that the quantity of money has grown slowly; high interest rates are a sign that monetary policy *has been* easy—in the sense that the quantity of money has grown rapidly. The broadest facts of experience run in precisely the opposite direction from that which the financial community and academic economists have all generally taken for granted.

Paradoxically, the monetary authority could assure low nominal rates of interest—but to do so it would have to start out in what seems like the opposite direction, by engaging in a deflationary monetary policy. Similarly, it could assure high nominal interest rates by engaging in an inflationary policy and accepting a temporary movement in interest rates in the opposite direction.

These considerations not only explain why monetary policy cannot peg interest rates; they also explain why interest rates are such a misleading indicator of whether monetary policy is “tight” or “easy.” For that, it is far better to look at the rate of change of the quantity of money.²

Employment as a Criterion of Policy

The second limitation I wish to discuss goes more against the grain of current thinking. Monetary growth, it is widely held, will tend to stimulate employment; monetary contraction, to retard employment. Why, then, cannot the monetary authority adopt a target for employment or unemployment—say, 3 per cent unemployment; be tight when unemployment is less than the target; be easy when unemployment is higher than the target; and in this way peg unemployment at, say, 3 per cent? The reason it cannot is precisely the same as for interest rates—the difference between the immediate and the delayed consequences of such a policy.

Thanks to Wicksell, we are all acquainted with the concept of a “natural” rate of interest and the possibility of a discrepancy between the “natural” and the “market” rate. The preceding analysis of interest rates can be translated fairly directly into Wicksellian terms. The monetary authority can make the market rate less than the natural rate

² This is partly an empirical not theoretical judgment. In principle, “tightness” or “ease” depends on the rate of change of the quantity of money supplied compared to the rate of change of the quantity demanded excluding effects on demand from monetary policy itself. However, empirically demand is highly stable, if we exclude the effect of monetary policy, so it is generally sufficient to look at supply alone.

only by inflation. It can make the market rate higher than the natural rate only by deflation. We have added only one wrinkle to Wicksell—the Irving Fisher distinction between the nominal and the real rate of interest. Let the monetary authority keep the nominal market rate for a time below the natural rate by inflation. That in turn will raise the nominal natural rate itself, once anticipations of inflation become widespread, thus requiring still more rapid inflation to hold down the market rate. Similarly, because of the Fisher effect, it will require not merely deflation but more and more rapid deflation to hold the market rate above the initial “natural” rate.

This analysis has its close counterpart in the employment market. At any moment of time, there is some level of unemployment which has the property that it is consistent with equilibrium in the structure of *real wage rates*. At that level of unemployment, real wage rates are tending on the average to rise at a “normal” secular rate, i.e., at a rate that can be indefinitely maintained so long as capital formation, technological improvements, etc., remain on their long-run trends. A lower level of unemployment is an indication that there is an excess demand for labor that will produce upward pressure on real wage rates. A higher level of unemployment is an indication that there is an excess supply of labor that will produce downward pressure on real wage rates. The “natural rate of unemployment,” in other words, is the level that would be ground out by the Walrasian system of general equilibrium equations, provided there is imbedded in them the actual structural characteristics of the labor and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labor availabilities, the costs of mobility, and so on.³

You will recognize the close similarity between this statement and the celebrated Phillips Curve. The similarity is not coincidental. Phillips' analysis of the relation between unemployment and wage change is deservedly celebrated as an important and original contribution. But, unfortunately, it contains a basic defect—the failure to distinguish between *nominal wages* and *real wages*—just as Wicksell's analysis failed to distinguish between *nominal interest rates* and *real interest rates*. Implicitly, Phillips wrote his article for a world in which everyone anticipated that nominal prices would be stable and in which that anticipation remained unshaken and immutable whatever happened to actual prices and wages. Suppose, by contrast, that everyone anticipates that prices will rise at a rate of more than 75 per cent a year—as, for exam-

³ It is perhaps worth noting that this “natural” rate need not correspond to equality between the number unemployed and the number of job vacancies. For any given structure of the labor market, there will be some equilibrium relation between these two magnitudes, but there is no reason why it should be one of equality.

ple, Brazilians did a few years ago. Then wages must rise at that rate simply to keep real wages unchanged. An excess supply of labor will be reflected in a less rapid rise in nominal wages than in anticipated prices,⁴ not in an absolute decline in wages. When Brazil embarked on a policy to bring down the rate of price rise, and succeeded in bringing the price rise down to about 45 per cent a year, there was a sharp initial rise in unemployment because under the influence of earlier anticipations, wages kept rising at a pace that was higher than the new rate of price rise, though lower than earlier. This is the result experienced, and to be expected, of all attempts to reduce the rate of inflation below that widely anticipated.⁵

To avoid misunderstanding, let me emphasize that by using the term "natural" rate of unemployment, I do not mean to suggest that it is immutable and unchangeable. On the contrary, many of the market characteristics that determine its level are man-made and policy-made. In the United States, for example, legal minimum wage rates, the Walsh-Healy and Davis-Bacon Acts, and the strength of labor unions all make the natural rate of unemployment higher than it would otherwise be. Improvements in employment exchanges, in availability of information about job vacancies and labor supply, and so on, would tend to lower the natural rate of unemployment. I use the term "natural" for the same reason Wicksell did—to try to separate the real forces from monetary forces.

Let us assume that the monetary authority tries to peg the "market" rate of unemployment at a level below the "natural" rate. For definiteness, suppose that it takes 3 per cent as the target rate and that the "natural" rate is higher than 3 per cent. Suppose also that we start out at a time when prices have been stable and when unemployment is higher than 3 per cent. Accordingly, the authority increases the rate of monetary growth. This will be expansionary. By making nominal cash

* Strictly speaking, the rise in nominal wages will be less rapid than the rise in anticipated nominal wages to make allowance for any secular changes in real wages.

⁵ Stated in terms of the rate of change of nominal wages, the Phillips Curve can be expected to be reasonably stable and well defined for any period for which the *average* rate of change of prices, and hence the anticipated rate, has been relatively stable. For such periods, nominal wages and "real" wages move together. Curves computed for different periods or different countries for each of which this condition has been satisfied will differ in level, the level of the curve depending on what the average rate of price change was. The higher the average rate of price change, the higher will tend to be the level of the curve. For periods or countries for which the rate of change of prices varies considerably, the Phillips Curve will not be well defined. My impression is that these statements accord reasonably well with the experience of the economists who have explored empirical Phillips Curves.

Restate Phillips' analysis in terms of the rate of change of real wages—and even more precisely, anticipated real wages—and it all falls into place. That is why students of empirical Phillips Curves have found that it helps to include the rate of change of the price level as an independent variable.

balances higher than people desire, it will tend initially to lower interest rates and in this and other ways to stimulate spending. Income and spending will start to rise.

To begin with, much or most of the rise in income will take the form of an increase in output and employment rather than in prices. People have been expecting prices to be stable, and prices and wages have been set for some time in the future on that basis. It takes time for people to adjust to a new state of demand. Producers will tend to react to the initial expansion in aggregate demand by increasing output, employees by working longer hours, and the unemployed, by taking jobs now offered at former nominal wages. This much is pretty standard doctrine.

But it describes only the initial effects. Because selling prices of products typically respond to an unanticipated rise in nominal demand faster than prices of factors of production, real wages received have gone down—though real wages anticipated by employees went up, since employees implicitly evaluated the wages offered at the earlier price level. Indeed, the simultaneous fall *ex post* in real wages to employers and rise *ex ante* in real wages to employees is what enabled employment to increase. But the decline *ex post* in real wages will soon come to affect anticipations. Employees will start to reckon on rising prices of the things they buy and to demand higher nominal wages for the future. “Market” unemployment is below the “natural” level. There is an excess demand for labor so real wages will tend to rise toward their initial level.

Even though the higher rate of monetary growth continues, the rise in real wages will reverse the decline in unemployment, and then lead to a rise, which will tend to return unemployment to its former level. In order to keep unemployment at its target level of 3 per cent, the monetary authority would have to raise monetary growth still more. As in the interest rate case, the “market” rate can be kept below the “natural” rate only by inflation. And, as in the interest rate case, too, only by accelerating inflation. Conversely, let the monetary authority choose a target rate of unemployment that is above the natural rate, and they will be led to produce a deflation, and an accelerating deflation at that.

What if the monetary authority chose the “natural” rate—either of interest or unemployment—as its target? One problem is that it cannot know what the “natural” rate is. Unfortunately, we have as yet devised no method to estimate accurately and readily the natural rate of either interest or unemployment. And the “natural” rate will itself change from time to time. But the basic problem is that even if the monetary authority knew the “natural” rate, and attempted to peg the market rate at that level, it would not be led to a determinate policy. The “market” rate will vary from the natural rate for all sorts of reasons other than monetary policy. If the monetary authority responds to

these variations, it will set in train longer term effects that will make any monetary growth path it follows ultimately consistent with the rule of policy. The actual course of monetary growth will be analogous to a random walk, buffeted this way and that by the forces that produce temporary departures of the market rate from the natural rate.

To state this conclusion differently, there is always a temporary trade-off between inflation and unemployment; there is no permanent trade-off. The temporary trade-off comes not from inflation per se, but from unanticipated inflation, which generally means, from a rising rate of inflation. The widespread belief that there is a permanent trade-off is a sophisticated version of the confusion between "high" and "rising" that we all recognize in simpler forms. A rising rate of inflation may reduce unemployment, a high rate will not.

But how long, you will say, is "temporary"? For interest rates, we have some systematic evidence on how long each of the several effects takes to work itself out. For unemployment, we do not. I can at most venture a personal judgment, based on some examination of the historical evidence, that the initial effects of a higher and unanticipated rate of inflation last for something like two to five years; that this initial effect then begins to be reversed; and that a full adjustment to the new rate of inflation takes about as long for employment as for interest rates, say, a couple of decades. For both interest rates and employment, let me add a qualification. These estimates are for changes in the rate of inflation of the order of magnitude that has been experienced in the United States. For much more sizable changes, such as those experienced in South American countries, the whole adjustment process is greatly speeded up.

To state the general conclusion still differently, the monetary authority controls nominal quantities—directly, the quantity of its own liabilities. In principle, it can use this control to peg a nominal quantity—an exchange rate, the price level, the nominal level of national income, the quantity of money by one or another definition—or to peg the rate of change in a nominal quantity—the rate of inflation or deflation, the rate of growth or decline in nominal national income, the rate of growth of the quantity of money. It cannot use its control over nominal quantities to peg a real quantity—the real rate of interest, the rate of unemployment, the level of real national income, the real quantity of money, the rate of growth of real national income, or the rate of growth of the real quantity of money.

II. *What Monetary Policy Can Do*

Monetary policy cannot peg these real magnitudes at predetermined levels. But monetary policy can and does have important effects on these real magnitudes. The one is in no way inconsistent with the other.

My own studies of monetary history have made me extremely sympathetic to the oft-quoted, much reviled, and as widely misunderstood, comment by John Stuart Mill. "There cannot . . .," he wrote, "be intrinsically a more insignificant thing, in the economy of society, than money; except in the character of a contrivance for sparing time and labour. It is a machine for doing quickly and commodiously, what would be done, though less quickly and commodiously, without it: and like many other kinds of machinery, it only exerts a distinct and independent influence of its own when it gets out of order" [7, p. 488].

True, money is only a machine, but it is an extraordinarily efficient machine. Without it, we could not have begun to attain the astounding growth in output and level of living we have experienced in the past two centuries—any more than we could have done so without those other marvelous machines that dot our countryside and enable us, for the most part, simply to do more efficiently what could be done without them at much greater cost in labor.

But money has one feature that these other machines do not share. Because it is so pervasive, when it gets out of order, it throws a monkey wrench into the operation of all the other machines. The Great Contraction is the most dramatic example but not the only one. Every other major contraction in this country has been either produced by monetary disorder or greatly exacerbated by monetary disorder. Every major inflation has been produced by monetary expansion—mostly to meet the overriding demands of war which have forced the creation of money to supplement explicit taxation.

The first and most important lesson that history teaches about what monetary policy can do—and it is a lesson of the most profound importance—is that monetary policy can prevent money itself from being a major source of economic disturbance. This sounds like a negative proposition: avoid major mistakes. In part it is. The Great Contraction might not have occurred at all, and if it had, it would have been far less severe, if the monetary authority had avoided mistakes, or if the monetary arrangements had been those of an earlier time when there was no central authority with the power to make the kinds of mistakes that the Federal Reserve System made. The past few years, to come closer to home, would have been steadier and more productive of economic well-being if the Federal Reserve had avoided drastic and erratic changes of direction, first expanding the money supply at an unduly rapid pace, then, in early 1966, stepping on the brake too hard, then, at the end of 1966, reversing itself and resuming expansion until at least November, 1967, at a more rapid pace than can long be maintained without appreciable inflation.

Even if the proposition that monetary policy can prevent money it-

self from being a major source of economic disturbance were a wholly negative proposition, it would be none the less important for that. As it happens, however, it is not a wholly negative proposition. The monetary machine has gotten out of order even when there has been no central authority with anything like the power now possessed by the Fed. In the United States, the 1907 episode and earlier banking panics are examples of how the monetary machine can get out of order largely on its own. There is therefore a positive and important task for the monetary authority—to suggest improvements in the machine that will reduce the chances that it will get out of order, and to use its own powers so as to keep the machine in good working order.

A second thing monetary policy can do is provide a stable background for the economy—keep the machine well oiled, to continue Mill's analogy. Accomplishing the first task will contribute to this objective, but there is more to it than that. Our economic system will work best when producers and consumers, employers and employees, can proceed with full confidence that the average level of prices will behave in a known way in the future—preferably that it will be highly stable. Under any conceivable institutional arrangements, and certainly under those that now prevail in the United States, there is only a limited amount of flexibility in prices and wages. We need to conserve this flexibility to achieve changes in relative prices and wages that are required to adjust to dynamic changes in tastes and technology. We should not dissipate it simply to achieve changes in the absolute level of prices that serve no economic function.

In an earlier era, the gold standard was relied on to provide confidence in future monetary stability. In its heyday it served that function reasonably well. It clearly no longer does, since there is scarce a country in the world that is prepared to let the gold standard reign unchecked—and there are persuasive reasons why countries should not do so. The monetary authority could operate as a surrogate for the gold standard, if it pegged exchange rates and did so exclusively by altering the quantity of money in response to balance of payment flows without "sterilizing" surpluses or deficits and without resorting to open or concealed exchange control or to changes in tariffs and quotas. But again, though many central bankers talk this way, few are in fact willing to follow this course—and again there are persuasive reasons why they should not do so. Such a policy would submit each country to the vagaries not of an impersonal and automatic gold standard but of the policies—deliberate or accidental—of other monetary authorities.

In today's world, if monetary policy is to provide a stable background for the economy it must do so by deliberately employing its powers to that end. I shall come later to how it can do so.

Finally, monetary policy can contribute to offsetting major disturbances in the economic system arising from other sources. If there is an independent secular exhilaration—as the postwar expansion was described by the proponents of secular stagnation—monetary policy can in principle help to hold it in check by a slower rate of monetary growth than would otherwise be desirable. If, as now, an explosive federal budget threatens unprecedented deficits, monetary policy can hold any inflationary dangers in check by a slower rate of monetary growth than would otherwise be desirable. This will temporarily mean higher interest rates than would otherwise prevail—to enable the government to borrow the sums needed to finance the deficit—but by preventing the speeding up of inflation, it may well mean both lower prices and lower nominal interest rates for the long pull. If the end of a substantial war offers the country an opportunity to shift resources from wartime to peacetime production, monetary policy can ease the transition by a higher rate of monetary growth than would otherwise be desirable—though experience is not very encouraging that it can do so without going too far.

I have put this point last, and stated it in qualified terms—as referring to major disturbances—because I believe that the potentiality of monetary policy in offsetting other forces making for instability is far more limited than is commonly believed. We simply do not know enough to be able to recognize minor disturbances when they occur or to be able to predict either what their effects will be with any precision or what monetary policy is required to offset their effects. We do not know enough to be able to achieve stated objectives by delicate, or even fairly coarse, changes in the mix of monetary and fiscal policy. In this area particularly the best is likely to be the enemy of the good. Experience suggests that the path of wisdom is to use monetary policy explicitly to offset other disturbances only when they offer a “clear and present danger.”

III. How Should Monetary Policy Be Conducted?

How should monetary policy be conducted to make the contribution to our goals that it is capable of making? This is clearly not the occasion for presenting a detailed “Program for Monetary Stability”—to use the title of a book in which I tried to do so [3]. I shall restrict myself here to two major requirements for monetary policy that follow fairly directly from the preceding discussion.

The first requirement is that the monetary authority should guide itself by magnitudes that it can control, not by ones that it cannot control. If, as the authority has often done, it takes interest rates or the current unemployment percentage as the immediate criterion of policy,

it will be like a space vehicle that has taken a fix on the wrong star. No matter how sensitive and sophisticated its guiding apparatus, the space vehicle will go astray. And so will the monetary authority. Of the various alternative magnitudes that it can control, the most appealing guides for policy are exchange rates, the price level as defined by some index, and the quantity of a monetary total—currency plus adjusted demand deposits, or this total plus commercial bank time deposits, or a still broader total.

For the United States in particular, exchange rates are an undesirable guide. It might be worth requiring the bulk of the economy to adjust to the tiny percentage consisting of foreign trade if that would guarantee freedom from monetary irresponsibility—as it might under a real gold standard. But it is hardly worth doing so simply to adapt to the average of whatever policies monetary authorities in the rest of the world adopt. Far better to let the market, through floating exchange rates, adjust to world conditions the 5 per cent or so of our resources devoted to international trade while reserving monetary policy to promote the effective use of the 95 per cent.

Of the three guides listed, the price level is clearly the most important in its own right. Other things the same, it would be much the best of the alternatives—as so many distinguished economists have urged in the past. But other things are not the same. The link between the policy actions of the monetary authority and the price level, while unquestionably present, is more indirect than the link between the policy actions of the authority and any of the several monetary totals. Moreover, monetary action takes a longer time to affect the price level than to affect the monetary totals and both the time lag and the magnitude of effect vary with circumstances. As a result, we cannot predict at all accurately just what effect a particular monetary action will have on the price level and, equally important, just when it will have that effect. Attempting to control directly the price level is therefore likely to make monetary policy itself a source of economic disturbance because of false stops and starts. Perhaps, as our understanding of monetary phenomena advances, the situation will change. But at the present stage of our understanding, the long way around seems the surer way to our objective. Accordingly, I believe that a monetary total is the best currently available immediate guide or criterion for monetary policy—and I believe that it matters much less which particular total is chosen than that one be chosen.

A second requirement for monetary policy is that the monetary authority avoid sharp swings in policy. In the past, monetary authorities have on occasion moved in the wrong direction—as in the episode of the Great Contraction that I have stressed. More frequently, they have

moved in the right direction, albeit often too late, but have erred by moving too far. Too late and too much has been the general practice. For example, in early 1966, it was the right policy for the Federal Reserve to move in a less expansionary direction—though it should have done so at least a year earlier. But when it moved, it went too far, producing the sharpest change in the rate of monetary growth of the post-war era. Again, having gone too far, it was the right policy for the Fed to reverse course at the end of 1966. But again it went too far, not only restoring but exceeding the earlier excessive rate of monetary growth. And this episode is no exception. Time and again this has been the course followed—as in 1919 and 1920, in 1937 and 1938, in 1953 and 1954, in 1959 and 1960.

The reason for the propensity to overreact seems clear: the failure of monetary authorities to allow for the delay between their actions and the subsequent effects on the economy. They tend to determine their actions by today's conditions—but their actions will affect the economy only six or nine or twelve or fifteen months later. Hence they feel impelled to step on the brake, or the accelerator, as the case may be, too hard.

My own prescription is still that the monetary authority go all the way in avoiding such swings by adopting publicly the policy of achieving a steady rate of growth in a specified monetary total. The precise rate of growth, like the precise monetary total, is less important than the adoption of some stated and known rate. I myself have argued for a rate that would on the average achieve rough stability in the level of prices of final products, which I have estimated would call for something like a 3 to 5 per cent per year rate of growth in currency plus all commercial bank deposits or a slightly lower rate of growth in currency plus demand deposits only.⁶ But it would be better to have a fixed rate that would on the average produce moderate inflation or moderate deflation, provided it was steady, than to suffer the wide and erratic perturbations we have experienced.

Short of the adoption of such a publicly stated policy of a steady rate of monetary growth, it would constitute a major improvement if the monetary authority followed the self-denying ordinance of avoiding wide swings. It is a matter of record that periods of relative stability in the rate of monetary growth have also been periods of relative stability in economic activity, both in the United States and other countries. Periods of wide swings in the rate of monetary growth have also been periods of wide swings in economic activity.

⁶ In an as yet unpublished article on "The Optimum Quantity of Money," I conclude that a still lower rate of growth, something like 2 per cent for the broader definition, might be better yet in order to eliminate or reduce the difference between private and total costs of adding to real balances.

By setting itself a steady course and keeping to it, the monetary authority could make a major contribution to promoting economic stability. By making that course one of steady but moderate growth in the quantity of money, it would make a major contribution to avoidance of either inflation or deflation of prices. Other forces would still affect the economy, require change and adjustment, and disturb the even tenor of our ways. But steady monetary growth would provide a monetary climate favorable to the effective operation of those basic forces of enterprise, ingenuity, invention, hard work, and thrift that are the true springs of economic growth. That is the most that we can ask from monetary policy at our present stage of knowledge. But that much—and it is a great deal—is clearly within our reach.

REFERENCES

1. H. S. ELLIS, ed., *A Survey of Contemporary Economics*. Philadelphia 1948.
2. MILTON FRIEDMAN, "The Monetary Theory and Policy of Henry Simons," *Jour. Law and Econ.*, Oct. 1967, 10, 1-13.
3. ———, *A Program for Monetary Stability*. New York 1959.
4. E. A. GOLDENWEISER, "Postwar Problems and Policies," *Fed. Res. Bull.*, Feb. 1945, 31, 112-21.
5. P. T. HOMAN AND FRITZ MACHLUUP, ed., *Financing American Prosperity*. New York 1945.
6. A. P. LERNER AND F. D. GRAHAM, ed., *Planning and Paying for Full Employment*. Princeton 1946.
7. J. S. MILL, *Principles of Political Economy*, Bk. III, Ashley ed. New York 1929.

Migration, Unemployment and Development: A Two-Sector Analysis

By JOHN R. HARRIS AND MICHAEL P. TODARO*

Throughout many less developed economies of the world, especially those of tropical Africa, a curious economic phenomenon is presently taking place. Despite the existence of positive marginal products in agriculture and significant levels of urban unemployment, rural-urban labor migration not only continues to exist, but indeed, appears to be accelerating. Conventional economic models with their singular dependence on the achievement of a full employment equilibrium through appropriate wage and price adjustments are hard put to provide rational behavioral explanations for these sizable and growing levels of urban unemployment in the absence of absolute labor redundancy in the economy as a whole. Moreover, this lack of an adequate analytical model to account for the unemployment phenomenon often leads to rather amorphous explanations such as the "bright lights" of the city acting as a magnet to lure peasants into urban areas.

In this paper we shall diverge from the usual full employment, flexible wage-price models of economic analysis by formulating a two-sector model of rural-urban migration which, among other things, recognizes the existence of a politically

determined minimum urban wage at levels substantially higher than agricultural earnings.¹ We shall then consider the effect of this parametric urban wage on the rural individual's economic behavior when the assumption of no agricultural labor surplus is made, i.e., that the agricultural marginal product is always positive and inversely related to the size of the rural labor force.² The distinguishing feature of this model is that migration proceeds in response to urban-rural differences in *expected earnings* (defined below) with the urban employment rate acting as an equilibrating force on such migration.³ We shall then use the overall model for the following purposes:

- 1) to demonstrate that given this po-

¹ For some empirical evidence on the magnitude of these real earnings differentials in less developed economies, see Reynolds, Berg, Henderson, and Ghai.

² We do not make the special assumption of an agricultural labor surplus for the following reasons: Most available empirical evidence to date tends to cast doubt on the labor surplus argument in the context of those economies of Southeast Asia and Latin America where such a surplus would be most likely to exist (see Kao, Anschel, and Eicher). Moreover, few if any economists would seriously argue that general labor surplus exists in tropical Africa, the area to which this paper is most directly related.

³ For a dynamic model of labor migration in which urban unemployment rates and expected incomes play a pivotal role in the migration process, see Todaro. However, unlike the present model which attempts to view the migration process in context of aggregate and inter-sectoral welfare considerations, Todaro's model was strictly concerned with the formulation of a positive theory of urban unemployment in developing nations. As such, it did not specifically consider the welfare of the rural sector, nor was it concerned with the broader issues of economic policy considered in the present paper.

* The authors are assistant professor of economics, Massachusetts Institute of Technology and research fellow, Institute for Development Studies, University College, Nairobi, respectively. They would like to thank the Rockefeller Foundation for making possible their research on economic problems of East Africa. Peter Diamond, Richard Eckaus, Joseph Stiglitz, two anonymous referees, and the managing editor made valuable comments on a previous draft. The authors, of course, are responsible for remaining errors.

litically determined high minimum wage, the continued existence of rural-urban migration in spite of substantial overt urban unemployment represents an economically rational choice on the part of the individual migrant;

2) to show that economists' standard policy prescription of generating urban employment opportunities through the use of "shadow prices" implemented by means of wage subsidies or direct government hiring will *not* necessarily lead to a welfare improvement and may, in fact, exacerbate the problem of urban unemployment;

3) to evaluate the welfare implications of alternative policies associated with various back-to-the-land programs when it is recognized that the standard remedy suggested by economic theory—namely, full wage flexibility—is for all practical purposes politically infeasible. Special attention will be given here to the impact of migration cum unemployment on the welfare of the rural sector as a whole which gives rise to intersectoral compensation requirements; and, finally,

4) to argue that in the absence of wage flexibility, an optimal policy is, in fact, a "policy package" including *both* partial wage subsidies (or direct government employment) and measures to restrict free migration.

I. *The Basic Model*

The basic model which we shall employ can be described as a two-sector internal trade model with unemployment. The two sectors are the permanent urban and the rural. For analytical purposes we shall distinguish between sectors from the point of view of production and income. The urban sector specializes in the production of a manufactured good, part of which is exported to the rural sector in exchange for agricultural goods. The rural sector has a choice of either using all available labor to produce a single agricultural good, some

of which is exported to the urban sector, *or* using only part of its labor to produce this good while *exporting* the remaining labor to the urban sector in return for wages paid in the form of the manufactured good. We are thus assuming that the typical migrant retains his ties to the rural sector and, therefore, the income that he earns as an urban worker will be considered, from the standpoint of sectoral welfare, as accruing to the rural sector.⁴ However, this assumption is not at all necessary for our demonstration of the rationality of migration in the face of significant urban unemployment.

The crucial assumption to be made in our model is that rural-urban migration will continue so long as the *expected* urban real income at the margin exceeds real agricultural product—i.e., prospective rural migrants behave as maximizers of *expected* utility. For analytical purposes, we shall assume that the total urban labor force consists of a permanent urban proletariat without ties to the rural sector plus the available supply of rural migrants. From this combined pool of urban labor, we assume that a periodic *random job selection process* exists whenever the number of available jobs is exceeded by the number of job seekers.⁵ Consequently, the expected

⁴ In tropical Africa especially, this notion that migrants retain their ties to the rural sector is quite common and manifested by the phenomenon of the extended family system and the flow of remittances to rural relatives of large proportions of urban earnings. However, the reverse flow, i.e., rural-urban monetary transfers is also quite common in cases where the migrant is temporarily unemployed and, therefore, must be supported by rural relatives. For an excellent discussion of this phenomenon from a sociological point of view, see Gugler (pp. 475-78).

⁵ The qualitative conclusions of the model do not depend on the precise nature of the selection process. We have assumed random selection not merely for analytic convenience but also because it directly corresponds to an appropriate dynamic construct developed in Todaro's 1969 article. There it is shown that over time expected and actual earnings will converge to a positive number even though the rate of job creation is less than the rate of migration so that unemployment is increasing.

urban wage will be defined as equal to the fixed minimum wage (expressed in terms of manufactured goods) times the proportion of the urban labor force actually employed (see equation (6)). Finally, we assume perfectly competitive behavior on the part of producers in both sectors with the further simplifying assumption that the price of the agricultural good (defined in terms of manufactured goods) is determined directly by the relative quantities of the two goods produced.

Consider now the following formulation of the model.

Agricultural Production Function:

$$(1) \quad X_A = q(N_A, \bar{L}, \bar{K}_A), \quad q' > 0, \quad q'' < 0$$

where,

X_A is output of the agricultural good,
 N_A is the rural labor used to produce
 this output,

\bar{L} is the fixed availability of land,

\bar{K}_A is the fixed capital stock,

q' is the derivative of q with respect of
 N_A , its only variable factor.

Manufacturing Production Function:

$$(2) \quad X_M = f(N_M, \bar{K}_M), \quad f' > 0, \quad f'' < 0$$

where

X_M is the output of the manufactured
 good,

N_M is the total labor (urban and rural
 migrant) required to produce this
 output.

\bar{K}_M is fixed capital stock, and

f' is the derivative of f with respect to
 N_M , its only variable factor.

It is interesting to note in this context that sociologist Gugler who has spent considerable time studying labor migration in Africa has recently concluded that rural-urban migration is essentially an economic phenomenon that can be portrayed as a "game of lottery" in which rural migrants come to the city fully aware that their chances of finding a job are low. However, the great disparity between urban and rural wages makes the successful location of an urban salaried job so attractive that unskilled migrants are willing to take a chance (pp. 472-73). See also Hutton.

Price Determination:

$$(3) \quad P = \rho \left(\frac{X_M}{X_A} \right), \quad \rho' > 0$$

where

P , the price of the agricultural good in terms of the manufactured good, (i.e., the terms of trade) is a function of the relative outputs of agricultural and manufactured good when the latter serves as numeraire.⁶

Agricultural Real Wage Determination:

$$(4) \quad W_A = P \cdot q'$$

where

W_A , the agricultural real wage, is equal to the value of labor's marginal product in agriculture expressed in terms of the manufactured good.

Manufacturing Real Wage:

$$(5) \quad W_M = f' \geq \bar{W}_M.$$

The real wage in manufacturing, expressed in terms of manufactured goods, is equated with the marginal product of labor in manufacturing because of profit maximization on the part of perfectly competitive producers. However, this wage is constrained to be greater than or equal to the fixed minimum urban wage. In our analysis, we shall be dealing only with cases in which $f' = \bar{W}_M$ (i.e., there is never an excess demand for labor at the minimum wage).

Urban Expected Wage:

$$(6) \quad W_u^* = \frac{\bar{W}_M N_M}{N_u}, \quad \frac{N_M}{N_u} \leq 1,$$

⁶ A sufficient, but not necessary, condition for this assumption is that all individuals in the economy have the same homothetic preference map. Again, the assumption is made for analytical convenience. The qualitative conclusions of our analysis will remain unaffected under several plausible assumptions about distribution of income and tastes.

where the *expected* real wage in the urban sector, W_u^e , is equal to the real minimum wage \bar{W}_M adjusted for the proportion of the total urban labor force (permanent urban plus migrants, denoted as N_u) actually employed, N_M/N_u .⁷ Only in the case of full employment in the urban sector ($N_M=N_u$) is the expected wage equal to the minimum wage (i.e., $W_u^e=\bar{W}_M$).

Labor Endowment:

$$(7) \quad N_A + N_u = \bar{N}_R + \bar{N}_u = \bar{N}$$

There is a *labor constraint* which states that the sum of workers actually employed in the agricultural sector (N_A) plus the total urban labor force (N_u) must equal the sum of initial endowments of rural (\bar{N}_R) and permanent urban (\bar{N}_u) labor which in turn equals the total labor endowment (\bar{N}).

Equilibrium Condition:

$$(8) \quad W_A = W_u^e$$

Equation (8), an equilibrium condition, is derived from the hypothesis that migration to the urban area is a positive function of the urban-rural *expected* wage differential. This can be written formally as

$$(9) \quad N_u = \psi \left(\frac{\bar{W}_M N_M}{N_u} - P \cdot q' \right),$$

$$\psi' > 0, \quad \psi(0) = 0$$

where \dot{N}_u is a time derivative. Clearly then, migration will cease only when the expected income differential is zero, the con-

⁷ This assumes a very particular form of wage expectation, namely that the expected wage is equal to the average urban wage. Although this is a convenient expression to work with, we could be more general and make the expected wage some function of the average urban wage. Indeed, the only restrictions on such a function that are necessary for our results are that, *ceteris paribus*, the expected wage varies directly with the minimum wage and inversely with the unemployment rate.

dition posited in (8).⁸ It is important to note that this assumes that a migrant gives up only his marginal product.⁹

We thus have 8 equations in 8 unknowns X_A , X_M , N_A , N_M , W_A , W_u^e , N_u and P . Given the production functions and fixed minimum wage \bar{W}_M , it is possible to solve for sectoral employment, the equilibrium unemployment rate and, consequently, the equilibrium expected wage, relative output levels and terms of trade. Let us analyze how such an unemployment equilibrium can come about.

The essence of our argument is that in many developing nations the existence of an institutionally determined urban minimum wage at levels substantially higher than that which the free market would allow can, and usually does, lead to an equilibrium with considerable urban unemployment. In our model migration is a disequilibrium phenomenon. In equilibrium $\bar{W}_M N_M / N_u = P q'$ and migration ceases. (See Appendix I for proof that this equilibrium is stable.) Now we know from equation (5) that in the competitive urban manufacturing sector, $\bar{W}_M = f'$. We also know from equation (7) that $\bar{N} - N_A = N_u$ and from equation (3) that $P = \rho(X_M /$

⁸ $\psi(0)=0$ is purely arbitrary. If, instead, we assume $\psi(\alpha)=0$ where α can take on any value, migration will cease when the urban-rural expected wage differential is equal to α . None of the subsequent analysis is affected qualitatively by specifying $\alpha=0$. Equation (8) would merely be written as $W_A + \alpha = W_u^e$.

⁹ Other assumptions could be made. Much of the literature has stressed that in peasant economies producers receive their average product which is higher than their marginal product. Indeed, this is at the heart of the well-known Lewis and Fei-Ranis models. However, these models ignore the migration decision and seem to assume that migrants continue to receive their share of peasant production yet migrate only if jobs are actually available. In much of Africa it appears that migrants continue to receive income from land after migration and commonly hire labor to work on their farms in their absence. There is also a considerable group of landless individuals who work on farms for wages. Thus it would appear that our assumption is not unreasonable. The analysis could easily be modified to make earnings foregone equal to average product, however.

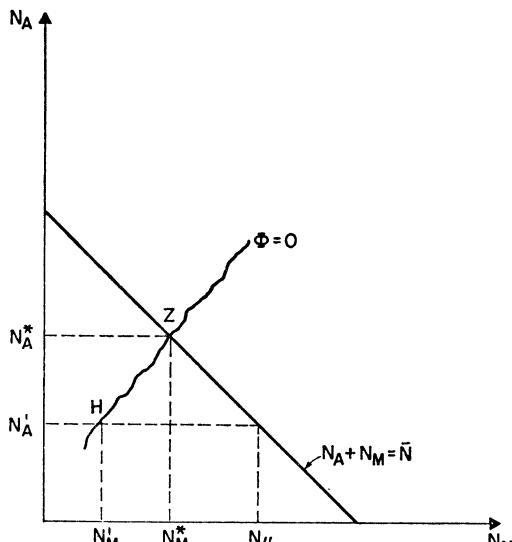


FIG 1.

X_A). Therefore, we can rewrite our equilibrium condition (8) as

$$(8') \quad \Phi = \rho(X_M/X_A)q' - \frac{f'N_M}{\bar{N} - N_A} = 0.$$

Since X_M and X_A are functions of N_M and N_A respectively, Φ is an implicit function in N_A and N_M which, for any stated minimum wage, can be solved for the equilibrium combination of agricultural and manufacturing employment. From this solution the levels of urban unemployment and commodity outputs can also be determined. There will be a unique equilibrium associated with each possible value of the minimum wage, and the locus of these equilibria is plotted in Figure 1 as the line $\Phi=0$ in N_A , N_M space.¹⁰ The line N_A

¹⁰ In Figure 1 we have assumed that

$$\frac{dN_A}{dN_M} = -[\Phi_{N_M}/\Phi_{N_A}] > 0$$

although this need not necessarily hold true. Differentiating (8') partially with respect to N_A we find that

$$\Phi_{N_A} = \frac{-\rho'fq'^2}{q^2} + \rho q'' - \frac{\rho q'}{\bar{N} - N_A}$$

$+N_M = \bar{N}$ in Figure 1 is the locus of full-employment points.

Point Z is the only equilibrium full-employment point in Figure 1 at which N_M^* workers would be employed in manufacturing and N_A^* in agriculture. Points on the locus $\Phi=0$ east of Z are infeasible and will not be considered further, while points to the west of Z are associated with min-

which is unambiguously negative since $q'' < 0$ and $\rho' > 0$. Differentiating (8') partially with respect to N_M we find that

$$\Phi_{N_M} = \frac{1}{\eta_{LW}} - \eta_P \frac{f'N_M}{X_M} + 1$$

which is less than, equal to, or greater than zero as

$$-\frac{1}{\eta_{LW}} + \eta_P \frac{f'N_M}{X_M} \geq 1,$$

where

$$\eta_{LW} = -\frac{dN_M}{dW_u} \frac{\bar{W}_u}{N_M}$$

is the wage elasticity of demand for labor and

$$\eta_P = \frac{dP}{d(\frac{X_M}{X_A})} \cdot \frac{X_M/X_A}{P}$$

is the elasticity of the terms of trade with respect to a change in relative outputs. It follows, therefore that the slope of the locus of equilibria, dN_A/dN_M depends on the respective employment and price elasticities.

A sufficient condition for Φ_{N_M} to be negative (making dN_A/dN_M positive) is for the wage elasticity of employment to be less than one, a situation which recent empirical studies suggest is likely to exist (see Erickson, Harris and Todaro (1969), and Katz). However, even if η_{LW} exceeds unity, dN_A/dN_M can still be positive providing price elasticity is sufficiently high. The logic of these conditions is clear. If η_{LW} is less than one, a decline in the minimum wage will lower the urban wage bill even though employment and output increase. This causes the expected urban wage to decline thereby reducing the expected rural-urban earnings differential which gives rise to reverse migration and increased rural employment and output. If η_{LW} exceeds unity, a fall in the minimum wage is accompanied by an increased urban wage bill and, hence, a higher expected urban wage. However, the expected rural-urban earnings differential can either increase or decrease in this case depending on the movement in terms of trade which raises the value of the marginal product in agriculture. For example, if η_{LW} were 1.5 and the wage share of manufacturing output ($f'N_M/X_M$) were .50, then an agricultural price elasticity greater than 0.67 would be sufficient to make dN_A/dN_M positive.

imum wages higher than the full-employment wage. There is a monotonic mapping such that higher minimum wages are associated with points on $\Phi=0$ lying farther to the west. Thus we can demonstrate that the setting of a minimum wage above the market-clearing level causes an economy to settle at a point such as H in Figure 1. At H , N'_A workers are employed in agriculture, N'_M in manufacturing, and $N_u - N'_M$ workers are unemployed. It is evident that the minimum wage causes a loss of employment and hence output in both sectors.¹¹

It is important to note that even though an equilibrium at point H represents a suboptimum situation for the economy as a whole, it does represent a rational, utility maximizing choice for individual rural migrants given the level of the minimum wage.

One final point might be raised at this juncture. So far we have assumed that the urban minimum wage is fixed in terms of the manufactured good. What if, instead, the minimum wage were fixed in terms of the agricultural good? We would then substitute for equation (5):

$$(5') \quad W_M = \frac{f'}{P} \geq \bar{W}_M.$$

Substituting (4), (5'), and (6) into (8) we get the equilibrium relationship

$$(11) \quad Pq' = \frac{\left(\frac{f'}{P}\right) \cdot N_M}{N_u}.$$

¹¹ If $dN_A/dN_M < 0$, which we believe to be empirically unlikely, this statement would have to be modified. In such a case, increasing the minimum wage will decrease manufacturing employment but will increase agricultural employment and output. Unemployment will result from the imposition of a minimum wage but we can no longer assert that the level of unemployment will increase concomitantly with the level of the minimum wage.

We can then imagine an economy starting initially at the point on the production possibilities frontier at which X_M is that for which equation (5') is satisfied and assume that

$$Pq' < \frac{\left(\frac{f'}{P}\right) \cdot N_M}{N_u}$$

at that point. The equilibrium point will again be reached through a simultaneous raising of Pq' and lowering of W_u^* in response to migration. As relative agricultural output falls, P will rise. This in turn will cause output of the manufactured good to fall as well, since producers will produce up to the point that $f' = \bar{W}_M P$ which rises in terms of the manufactured good. Note that f' can be raised only through output restriction (since $f'' < 0$). Therefore, in general, we would find that imposition of a minimum wage gives rise to an equilibrium characterized by unemployment and loss of potential output of both goods. A new locus $\Phi' = 0$ will be defined in Figure 1 such that the point on Φ' corresponding to any given minimum wage will be west of the corresponding point on Φ .

Although our initial assumption is a bit easier to handle, the principal conclusion remains unaffected if we make the minimum wage fixed in terms of the agricultural good. Equilibrium is only achievable with unemployment. Actual minimum wage setting is usually done with reference to some general cost of living index, and food is the largest single item in the budget of most urban workers. (See Massell and Heyer, and the Nigeria report.) Hence, the second case may be somewhat more realistic. Note that in the first case the "true" real wage was reduced somewhat by the rising agricultural price, while in the latter case it is increased by the falling relative price of the manufactured good.

III. *Implications for Development Policy*

A. *Planning in Terms of Shadow Prices*

The standard solution to the problem of an institutionally determined wage that is higher than the equilibrium level is to employ labor in the public sector according to a shadow wage and/or to grant a payroll subsidy to private employers that equates private costs with this shadow wage.¹² Two main problems arise with this prescription: first, how can one determine the appropriate shadow wage? and, secondly, what are the implications of executing such a scheme when the institutional wage will continue to be paid to the employed? Our model can shed light on both of these issues.

In a static framework the appropriate shadow wage is the opportunity cost of labor hired by the industrial sector. Hence, if labor is hired to the point that its marginal product in industry is equated with the shadow wage which in turn is equated with the marginal product in agriculture, marginal productivity of labor will be equal in both sectors, a necessary condition for an optimal allocation of resources. Naturally, this assumes a positive marginal product in agriculture and sufficient factor mobility to ensure full employment of labor. The existence of urban unemployment, however, suggests that there may be

a pool of labor that can be tapped without sacrificing output. Consequently, it might be suggested that even though agricultural labor is fully employed at peak seasons, the appropriate shadow wage for urban labor is likely to be one that is lower than the marginal product in agriculture. This would be correct if the two labor forces, urban and rural, were separate noncompeting groups. In linear programming terms, there are two labor constraints and each may well have a different associated shadow wage.

Now, the essence of our model is that the two sectors *are* intimately connected through labor migration. If one additional job is created in the industrial sector at the minimum wage, the expected wage will rise and rural-urban migration will be induced. In Appendix II it is shown that more than one agricultural worker will likely migrate in response to the creation of one additional industrial job. Hence, the opportunity cost of an industrial worker will exceed the marginal product of an agricultural worker. On the other hand, an increase in agricultural income will induce reverse migration with no diminution of industrial output. Thus, the opportunity cost of labor is lower to the agricultural than to the industrial sector!

The literature has been strangely silent for the most part about the full implications of using shadow-wage criteria. In a static context, Stolper has pointed out that financing subsidies or losses of public enterprises gives rise to fiscal problems, but unfortunately this issue has not yet been pursued in sufficient detail.¹³ If the problem is considered at all, the analyst usually assumes that a system of nondistorting lump-sum taxes is available. Little, Lefever, and

¹² Hagen (p. 498) states, "a subsidy per unit of labor equal to the wage differential [between agriculture and industry] will increase real income further [than a tariff] and if combined with free trade will permit attaining an *optimum optimorum*." Bardhan (p. 379) similarly adds. "The best remedy for the misallocation caused by a wage differential is . . . an appropriate subsidy to the use of labor in the manufacturing industry." It is important to recall that this argument is dependent on variable proportions production functions. If production coefficients are fixed, a wage subsidy will have no effect in the short run. The classic statement of this case is by Eckaus. Bardhan explores its implications for subsidy in a dynamic context. Both of these papers, however, posit surplus labor in agriculture, an assumption we do not wish to make in an African context.

¹³ Lefever assumes that a wage subsidy can be financed by a profits tax, while other writers, e.g. Hagen, Bardhan, and Chakravarty never even consider the problem. Even Little and Mirrlees who present an excellent discussion of how to calculate a shadow wage never mention the fiscal problems of implementation.

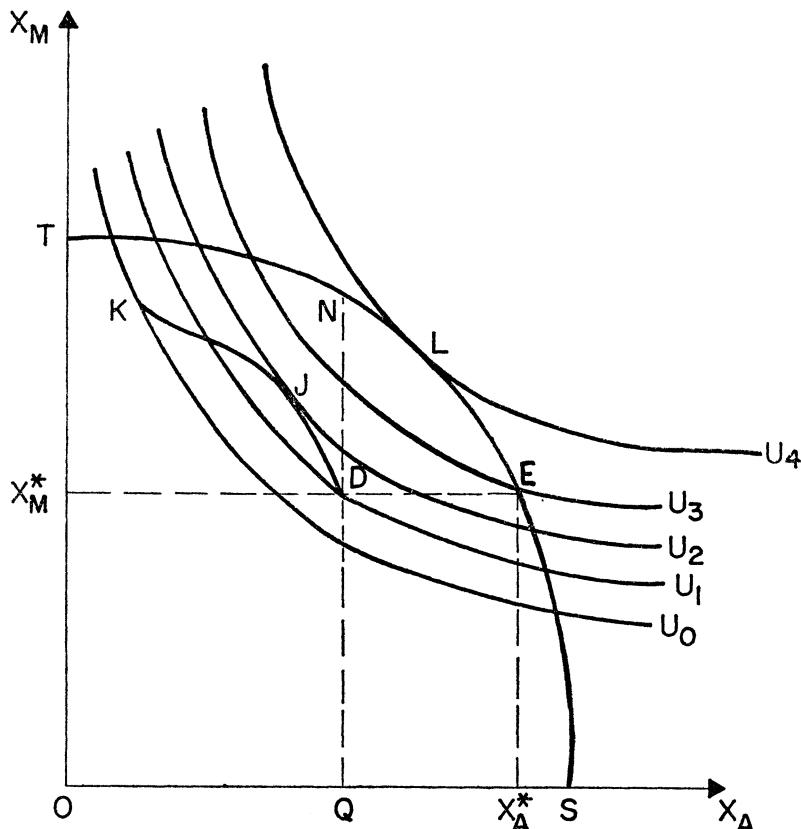


FIGURE 2

Little and Mirrlees have pointed out that in a dynamic setting, the extra consumption arising from payment of the institutional wage diverts resources from investment to consumption; thus some of the foregone future consumption should be considered in calculating the shadow wage. In our model, payment of the minimum wage to additional industrial workers will induce more rural-urban migration. Therefore, implementation of a shadow-wage employment criterion will have important effects on the level of agricultural output and on urban unemployment. The argument can be clarified with reference to Figure 2.

The initial equilibrium, given the minimum wage, is at point *D* with output of

the manufactured good restricted to OX_M^* . If individuals did not migrate in response to expected wage differentials, the economy could produce at point *E*, but migration reduces agricultural output to the level OQ . The theory of shadow pricing suggests that with an appropriate wage subsidy (or public-sector-hiring rule) the economy could move to point *L* on the production possibilities frontier which, with the posited social indifference map, is the optimum position. Welfare would be increased from a level U_1 to a higher level U_4 .

In the context of our model, such a point is unattainable. The effect of implementing a shadow wage will be to increase production of the manufactured good. But creation of an additional job at the minimum

wage will induce some additional migration (see Appendix II) from the rural sector and therefore agricultural output will fall. Hence, movement from D can only be in a northwest direction. The line DK in Figure 2 is the locus of all such attainable points and it is evident that there is only one point, K , at which there can be full employment of the economy's labor resources. At that point the expected wage will be equal to the minimum wage since there is no urban unemployment. Therefore, the marginal product in agriculture will have to be equal to the minimum wage. But, with the subsidy, the marginal product of labor in manufacturing will be lower than in agriculture, hence K lies inside the production possibilities frontier. (In the extreme case in which marginal productivity in agriculture can never be as high as the minimum wage, K will coincide with T , the point of complete specialization in manufactures.) This situation will certainly not meet the conditions for a general optimum which can be met only at L . Thus, implementing a shadow wage criterion to the point that urban unemployment is eliminated will not generally be a desirable policy.¹⁴

However, some level of wage subsidy will usually lead to an improvement. In Figure 2 it is clear that point J , with a welfare level U_2 , will be preferable to D . The criterion for welfare maximization, derived in Appendix III, is the following:

$$(12) \quad f' = Pq' \left(\frac{dN_u}{dN_M} \right).$$

Note what this means. Creating one additional job in the industrial sector increases output by f' but, since increased

¹⁴ As shown in Appendix III, DK is not uniformly convex. Therefore, K may be the best attainable point in some cases and the first-order conditions may not ensure optimality. As drawn in Figure 2, moving from D to K represents a worsening of welfare, but this clearly is not a necessary conclusion.

employment will raise the expected urban wage, migration will be induced in an amount dN_u/dN_M . The right-hand side of equation (12) states the amount of agricultural output sacrificed because of migration. Thus the shadow wage will be equal to this opportunity cost of an urban job and the amount of subsidy will be $\bar{W}_M - f'$. So long as $f' > Pq' (dN_u/dN_M)$, aggregate welfare can be increased by expanding industrial employment through subsidy or public sector hiring. Clearly the more responsive is migration to industrial employment, the higher is the social cost of industrialization and the smaller is the optimal amount of subsidy. In many African economies it is likely that dN_u/dN_M exceeds unity. If so, it will be optimal for the marginal product of labor in industry to be higher than in agriculture and urban unemployment will be a persistent phenomenon so long as minimum wages are set above a market-clearing level.

The discussion so far has ignored two other adverse effects of using a shadow wage. As mentioned earlier, several writers have noted that payment of a subsidized minimum wage to additional workers will increase total consumption, thereby reducing the level of resources available for investment. If foregone future consumption is positively valued, the opportunity cost of industrial labor will be higher than indicated in equation (12) and the shadow wage will be raised correspondingly. Furthermore, wage subsidies or public enterprise losses must be financed and if revenue cannot be raised through costless lump-sum taxes, the opportunity cost of raising taxes must be considered. Both of these effects will reduce the desirable amount of subsidized job creation in the industrial sector.

It is interesting to note that this model implies different opportunity costs of labor to the two sectors. While the creation of an additional job in the urban area reduces

agricultural output through induced migration, additional employment can be generated in the agricultural sector without reducing manufacturing output. If this phenomenon is not taken into account, standard application of investment criteria is likely to be biased in favor of urban projects.

B. Migration Restriction

An alternative approach to the problem of urban unemployment is to physically control migration from the rural areas. Such controls have recently been introduced in Tanzania and have been used for some time in South Africa.¹⁵ Other countries, such as Kenya, are giving serious consideration to instituting such a policy. Although we personally have grave reservations about the ethical issues involved in such a restriction of individual choice and the complexity and arbitrariness of administration, it seems desirable to investigate the economic implications of such a policy.

Looking at Figure 2 it is obvious that with the minimum wage such that industrial output is OX_M^* , prohibition of migration in excess of the labor required to produce that output will allow the economy to produce at point *E*. The movement from *D* to *E* arising from restriction of migration leads to an unambiguous aggregate welfare improvement providing appropriate lump-sum redistribution is effected. Since such compensation is notoriously difficult to carry out in practice, it will be useful to examine the welfare implications of such a move on each of the two sectors in the absence of compensation.

Recall that the two sectors were defined to be a permanent urban group and a rural sector that produces both agricultural goods and exports labor to the urban area

in exchange for wages in the form of manufactured goods.¹⁶ In Figure 3 the line $T'S'$ represents production possibilities for the agricultural sector when labor export is allowed. If its entire labor endowment is devoted to agricultural production, it can produce a quantity OS' . However, by exporting its labor, the agricultural sector can "produce" the manufactured good (wages are paid in the form of this good). Hence this production possibilities frontier depends on market forces (wage levels and unemployment) as well as on purely technological factors. The amount of agricultural output foregone if a unit of labor is to be "exported" is its marginal product; the amount of manufactured goods obtained by the exported labor unit depends on the wage, the amount of employment obtained by the exported unit, and its effect on employment of previously exported units.

In addition to these production possibilities, the rural sector also has the opportunity to trade some of its agricultural output with the permanent urban sector in exchange for manufactured goods. Corresponding to each point on the production possibilities frontier $T'S'$, there is a determinate price of the agricultural good. The manner in which alternative constellations of production and trade affect the

¹⁵ In considering the welfare of the rural sector as a whole we are making the tacit assumption that there is redistribution of goods between individuals in this sector. This is a very strong assumption. Yet there is considerable evidence from tropical Africa that employed urban migrants repatriate substantial portions of their earnings to their kinsmen remaining in the rural areas and conversely that income both in cash and kind is received by unemployed migrants from kinsmen remaining on the farm. To the extent that the extended family system does redistribute goods between members, this assumption may be tenable as a first approximation. As Gugler (p. 480) has pointed out, it is appropriate to view the extended family as maximizing its income by allocating its members between agriculture and urban wage employment. Although there is some evidence that growing numbers of urban workers are settling permanently and gradually eliminating rural ties, it will be many years before such ties are completely severed.

¹⁶ See Harris and Todaro (1969) for an analysis of the Tanzanian program.

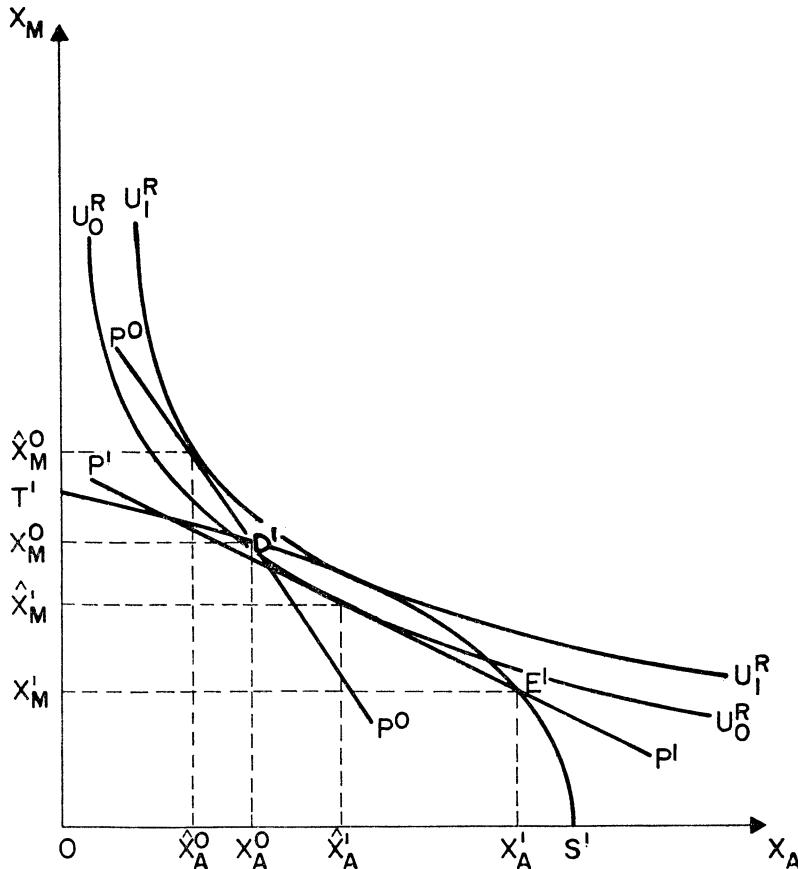


FIGURE 3

sector's welfare can be illustrated by Figure 3.

D' corresponds to the initial unemployment equilibrium D (Figure 2). At that point the rural sector as a whole "produces" X_A^0 and X_M^0 of the two goods. It also has the opportunity to trade at the price P^0 . By trading some of its agricultural output to the permanent urban sector for additional manufactured goods, it consumes \hat{X}_A^0 , \hat{X}_M^0 and achieves a welfare level of U_1^R . Restriction of migration results in the sector's producing X_A' , X_M' . If it could still trade at price P^0 , the agricultural sector would clearly be better off. But this is impossible. At E' (which corresponds to E in Figure 2), the price of

agricultural good will fall to P' and with trade the best consumption bundle attainable by the sector is \hat{X}_A , \hat{X}_M which corresponds to a lower level of welfare U_0^R . (Note that if P' did not cut $T'S'$ there could be no incentive to migrate at E' .)

It can be shown that $Pq'(1 - 1/\eta)$ (where η is the price elasticity of demand for the agricultural good) is the amount of the manufactured good sacrificed by the rural sector as a result of removing one worker from producing the agricultural good which could have been exchanged for the manufactured good at the market price $1/P$. This quantity is less than the value of labor's marginal product in agriculture (Pq') since the reduction in output has a

favorable terms-of-trade effect. If the demand for the agriculture good is inelastic ($\eta < 1$) we reach the startling conclusion that the sacrifice becomes negative! This is, of course, the familiar proposition that aggregate farm income may be increased by reducing output. The *direct* gain in manufactured goods achieved by the rural sector through exporting an additional unit of labor is $\bar{W}_M N_M / N_u$, the expected urban wage. But additional migration, by increasing unemployment, reduces the earnings of *all* migrants already in the urban labor force by a factor $(1 - R)$, where R is the fraction of the total urban labor force supplied by the rural sector.¹⁷

As long as $Pq' (1 - \eta) < \bar{W}_M N_M / N_u (1 - R)$ the welfare of the rural sector will be increased by allowing migration even though unemployment ensues and the economy as a whole sacrifices output. Since Pq' and $\bar{W}_M N_M / N_u$ are always positive and $R \leq 1$, additional migration will always benefit the rural sector when $\eta < 1$. In general, the lower is Pq' , η , or R and the higher is $\bar{W}_M N_M / N_u$, the more will the rural sector benefit from the opportunity to migrate.

From the foregoing, one can conclude that although migration restriction will improve aggregate welfare of the economy, given plausible values of η and R , substantial compensation to the rural sector will be required if it is not to be made worse off by removing the opportunity for free migration. The permanent urban labor force clearly will be made better off by becoming fully employed at the high

minimum wage while also being able to buy food at a lower price. Each unit of labor exported by the rural sector will similarly earn more but this gain will be offset by reduced total labor exports and lower agricultural prices. Whether or not this will be true depends, of course, on the values of the specific parameters of the economy. If η is sufficiently high, the rural sector could be made better off by restricting migration in the absence of compensation, but this seems very unlikely.

C. A Combination of Policies

It has been shown that either a limited wage-subsidy or a migration-restriction policy will lead to a welfare improvement. Which of the two policies will lead to the better position cannot be determined without knowing all the relevant parameters for a particular economy. It is clear, however, that neither policy alone is capable of moving the economy to the optimum that could be achieved with competitive wage determination (point *L* in Figure 2).

At first sight it may seem strange that with a single market failure, the wage level, a single policy instrument is unable to fully correct the situation.¹⁸ The reason is that the wage performs two functions in this model. It determines *both* the level of employment in the industrial sector *and* the allocation of labor between rural and urban areas. While a subsidy changes the effective wage for determination of industrial employment, so long as the wage actually received by workers exceeds agricultural earnings there will be migration and urban unemployment. Restriction of migration prevents the minimum wage having its effect on unemployment but does nothing to increase the level of industrial employment. Therefore, if the optimum position is to be achieved, a combination of both instruments will have to

¹⁷ If the urban unemployment were experienced only by migrants, this term would equal zero since the total amount of earnings through labor export would be constant. It can be positive only because the permanent urban labor force shares in unemployment, thereby reducing its share of the constant wage bill in the manufactured good industry. An interesting extension of the model would be to incorporate different employment probabilities for the permanent urban and migrant rural labor forces and then to check the sensitivity of results with our more simplified assumption of equal probabilities.

¹⁸ We wish to thank a referee of this *Review* for drawing this to our attention.

be used. In order to reach point L a wage subsidy must be instituted such that industrial employment will increase to the extent that with full employment the marginal product of labor will be equal in manufacturing and agriculture. The subsidy will be positive and equal to the difference between the minimum wage and marginal productivity. At that point $W_u^* = \bar{W}_M$ and $\bar{W}_M > Pq'$. Therefore, individuals would still find it in their interest to migrate and the point will not be attainable unless migration is restricted.

The agricultural sector has to be better off at L than at E since each additional unit of labor exported earns the full minimum wage, marginal productivity in agriculture is less than the minimum wage, and the price of the agricultural good rises. Whether the agricultural sector is better off at L than at D , however, depends again on the parametric values of the model.¹⁹ It can be stated with certainty that the amount of compensation needed to make the rural sector *no worse off* than at D will be less at L than at E , and, furthermore it should be easier to finance since total income is greater.

Even so the fiscal requirements of subsidy (or public enterprise losses) and compensation cannot be taken lightly.²⁰ A government may find it difficult to find

¹⁹ As drawn in Figure 2, L must represent a higher welfare level than D for the rural sector since P rises and the sector produces more of both goods. In fact if L lies along TS north of the ray going through D there will be an unambiguous sectoral welfare improvement. However, if L lies south of the ray on TS , the rural sector could be worse off than at D since P falls.

²⁰ This argument coincides with the statement by Stolper (p. 195), "It should be noted, however, that even at best the application of shadow prices leads to the substitution of one problem, the budget, for another one, an imperfect market."

We would not go as far as Stolper in rejecting out of hand any use of shadow pricing because of the fiscal implications. The general point is valid that one cannot disregard the consequences of implementation of shadow-price criteria if actual prices or wages continue to diverge from the shadow prices or wages.

nondistorting taxes capable of raising sufficient revenue. Perhaps a head-tax on all urban residents would be feasible although this too raises the question of how minimum wages are set (unions in tropical Africa have, in some cases, successfully fought to maintain the real after-tax wage). A tax on rural land is ruled out if there must be *net* compensation to the rural sector which, in the absence of pure profits in manufacturing, leaves an urban land tax as the remaining potential ideal tax.

All of the above suggests that altering the minimum wage may avoid the problems of taxation, administration, and interference with individual mobility attendant to the policy package just discussed. Income and wages policies designed to narrow the rural-urban wage gap have been suggested by D. P. Ghai, and Tanzania has formally adopted such a policy along with migration restriction. In the final analysis, however, the basic issue at stake is really one of political feasibility and it is not at all clear that an incomes policy is any more feasible than the alternatives.

APPENDIX I

Proof of Stability of Unemployment Equilibrium

In order to prove that our urban unemployment equilibrium is stable, we can differentiate ψ (equation (9)) with respect to N_u remembering that $dN_u = -dN_A$ according to (7). We therefore obtain

$$(1.1) \quad \frac{dN_u}{dN_u} = \psi'(\cdot) \left[-\frac{\bar{W}_M N_M}{(N_u)^2} + Pq'' \right. \\ \left. + \frac{\partial P}{\partial X_A} (q')^2 \right].$$

Stability requires $dN_u/dN_u < 0$ which is satisfied if

$$\frac{\partial P}{\partial X_A} < \frac{\frac{\bar{W}_M N_M}{(N_u)^2} - Pq''}{(q')^2}.$$

The right side of this inequality is unambiguously positive since $q'' < 0$. Hence our assumption that $\partial P/\partial X_A < 0$ will ensure stability and, indeed, is stronger than necessary. The adjustment mechanism may be made clear by the following phase diagram in which the function ψ is plotted. Its positive slope reflects the hypothesis that migration flows will increase with the magnitude of the urban-rural expected wage differential. In Figure 4, ψ is plotted under the assumption that $\psi(0) = 0$, hence the horizontal intercept is at the origin (in general the intercept would be α). Furthermore, we have arbitrarily assumed that ψ is a linear function. The arrows show the direction of adjustment in accordance with (1.1). If $\bar{W}_M N_M / N_u - P q' > 0$, then $\dot{N}_u > 0$ but we know that if $\dot{N}_u > 0$, the expected wage differential will decrease since $d\dot{N}_u/dN_u < 0$. Additional migration by increasing N_u without affecting N_M will reduce the expected urban real wage through increased unemployment. Concomitantly, the transfer of labor out of agriculture raises q' and reduced agricultural output also causes P to rise. Thus migration reduces the expected wage differential to zero and equilibrium is achieved when there is no further incentive for migration. See Todaro for a more detailed analysis of this process in a dynamic setting.

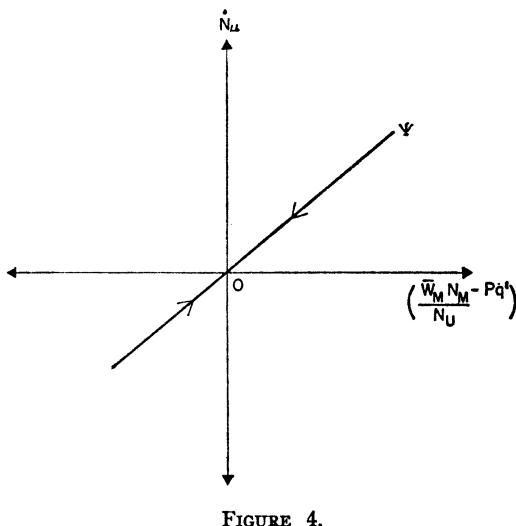


FIGURE 4.

APPENDIX II

Differentiating the equilibrium condition (8) with respect to N_M , recalling that $dN_u = -dN_A$, we obtain the expression

$$(II.1) \quad \frac{dN_u}{dN_M} = \frac{\frac{\bar{W}_M}{N_u} - q'\rho' \frac{f'}{X_A}}{\frac{\bar{W}_M N_M}{N_u^2} - \rho q'' + q'\rho' \frac{q' X_M}{X_A^2}}.$$

Defining the elasticity of demand for the agricultural good as

$$(II.2) \quad \eta_A = -\frac{\partial X_A}{\partial P} \cdot \frac{P}{X_A} = \frac{\rho X_A}{\rho' X_M},$$

(II.1) can be rewritten as

$$(II.3) \quad \frac{dN_u}{dN_M} = \frac{\frac{\bar{W}_M}{N_u} - \frac{\rho q' f'}{\eta_A X_M}}{\frac{\bar{W}_M N_M}{N_u^2} - \rho q'' + \frac{\rho (q')^2}{\eta_A X_A}}.$$

Differentiating the expression partially with respect to its various arguments it can be shown that dN_u/dN_M will vary directly with \bar{W}_M , N_M , η_A and inversely with ρ , q' , f' , N_u , and q'' . In general, the greater is the urban-rural wage differential, and the less sensitive are prices and marginal products in agriculture, the greater will be the migration induced by creation of an additional job. If the minimum wage exceeds agricultural earnings, (II.3) will generally be positive and, with parameter values relevant for many African economies, will exceed unity.

When $dN_u/dN_M > 1$, creation of an additional job at the minimum wage will increase the absolute level of unemployment although the *rate* of urban unemployment will have to fall. This can be seen by converting (II.3) to an elasticity measure.

$$(II.4) \quad \frac{dN_u}{dN_M} \cdot \frac{N_M}{N_u} =$$

$$\frac{\frac{\bar{W}_M N_M}{N_u^2} - \frac{N_M \rho q' f'}{N_u \eta_A X_M}}{\frac{\bar{W}_M N_M}{N_u^2} - \rho q'' + \frac{\rho (q')^2}{\eta_A X_A}} < 1$$

since $q'' < 0$.²¹ To give an example of what this means, suppose that an economy initially has an urban unemployment rate of 25 percent. If in response to the creation of 100 additional industrial jobs, 125 additional individuals migrate to the urban area, the absolute number unemployed increases by 25 although the unemployment rate will drop, since the marginal unemployment rate is only 20 percent.

APPENDIX III

If minimum wages are maintained and migration takes place in accordance with equation (8), aggregate welfare will be maximized if the following Lagrangean expression is maximized:

$$(III.1) \quad \begin{aligned} \Omega = & U(X_A, X_M) \\ & + \lambda_1 [q(\bar{N} - N_u) - X_A] \\ & + \lambda_2 [f(N_M) - X_M] \\ & + \lambda_3 \left\{ \rho \left(\frac{f(N_M)}{q(\bar{N} - N_u)} \right) \right. \\ & \left. - q'(N - N_u) - \frac{\bar{W}_M N_M}{N_u} \right\} \end{aligned}$$

where U is the social welfare function and the succeeding terms are the constraints imposed by equations (1), (2), and (8) (recall that $N_A = \bar{N} - N_u$ from equation (7)).

Maximizing (III.1) we get the following first-order conditions:

$$(III.2) \quad \frac{\partial \Omega}{\partial X_A} = \frac{\partial U}{\partial X_A} - \lambda_1 = 0$$

$$(III.3) \quad \frac{\partial \Omega}{\partial X_M} = \frac{\partial U}{\partial X_M} - \lambda_2 = 0$$

²¹ We are grateful to Peter Diamond for deriving this expression.

$$(III.4) \quad \frac{\partial \Omega}{\partial N_u} = -\lambda_1 q' + \lambda_3 \left[\rho' \frac{fq'}{q^2} - \rho q'' + \frac{\bar{W}_M N_M}{N_u^2} \right] = 0$$

$$(III.5) \quad \frac{\partial \Omega}{\partial N_M} = \lambda_2 f' + \lambda_3 \left[\rho' \frac{f' q'}{q} - \frac{\bar{W}_M}{N_u} \right] = 0$$

and the $\partial \Omega / \partial \lambda_i = 0$ ($i = 1, 2, 3$) which ensures that the constraints hold.

Substituting (III.2) and (III.3) into (III.4) and (III.5) we get

$$(III.6) \quad \begin{aligned} \frac{\frac{\partial U}{\partial X_M} f'}{\frac{\partial U}{\partial X_A} q'} = & \frac{\frac{\bar{W}_M}{N_u} - q' \rho' \frac{f'}{q}}{\frac{\bar{W}_M N_M}{N_u^2} - \rho q'' + q' \rho' \frac{fq'}{q^2}} . \end{aligned}$$

We know that in equilibrium $(\partial U / \partial X_M) / (\partial U / \partial X_A) = 1/P$ and it has been shown in Appendix II that the right-hand side of (III.6) is equal to dN_u / dN_M . Therefore (III.6) can be rewritten as

$$(III.7) \quad f' = P q' \frac{dN_u}{dN_M},$$

which is the condition used in the text to determine the optimal wage subsidy.

Condition (III.7) can also be written as

$$(III.8) \quad -P = \frac{-f'}{q' \frac{dN_u}{dN_M}} = \frac{dX_M}{dX_A} .$$

We know that $-P$ is equal to the marginal rate of substitution between the two commodities and dX_M / dX_A is the marginal rate of transformation. Hence (III.8) states the familiar condition for optimality: equate marginal rates of substitution and transformation. dX_M / dX_A is the slope of the line DK in Figure 2 and it clearly will be nega-

tive. However, its derivative with respect to N_M ,

$$(III.9) \quad \frac{d\left(\frac{dX_M}{dX_A}\right)}{dN_M} = \frac{-q' \frac{dN_u}{dN_M} f'' - f'\left(\frac{dN_u}{dN_M}\right)^2 q'' + f' q' \frac{d^2 N_u}{dN_M^2}}{\left(q' \frac{dN_u}{dN_M}\right)^2}$$

is of indeterminate sign since f'' , $q'' < 0$ and $d^2 N_u/dN_M^2$ will generally be negative as well. (III.9) must be positive if the effective production possibilities frontier (DK) is to be convex, a condition that is likely to hold but the possibility of concavity as full employment is approached must be considered. The slope of DK in Figure 2 seems plausible on a priori grounds.

REFERENCES

- P. K. Bardham, "Factor Market Disequilibrium and the Theory of Protection," *Oxford Econ. Pap.* (New Series), Oct. 1964, 16, 375-88.
- E. J. Berg, "Wage Structure in Less Developed Countries," in A. D. Smith, ed., *Wage Policy Issues in Economic Development*, London 1969.
- A. Callaway, "From Traditional Crafts to Modern Industries," *ODU: University of Ife Journal of African Studies*, July 1965, 2.
- S. Chakravarty, "The Use of Shadow Prices in Programme Evaluation," in Rosenstein-Rodan, ed., *Capital Formation and Economic Development*, London 1964.
- Y. S. Cho, *Disguised Unemployment in Developing Areas, with Special Reference to South Korean Agriculture*, Berkeley 1960.
- R. S. Eckaus, "The Factor-Proportions Problem in Underdeveloped Areas," *Amer Econ. Rev.*, Sept. 1955, 45, 539-65.
- J. Erickson, "Wage Employment Relationships in Latin American Industry: A Pilot Study of Argentina, Brazil, and Mexico," International Labour Office, 1969, typescript.
- J. Fei and G. Ranis, *Development of the Labor Surplus Economy*, Illinois 1964.
- D. P. Ghai, "Incomes Policy in Kenya: Need, Criteria and Machinery," *East Afr. Econ. Rev.*, June 1968, 4, 19-35.
- J. Gugler, "The Impact of Labour Migration on Society and Economy in Sub-Saharan Africa. Empirical Findings and Theoretical Considerations," *African Social Research*, Dec. 1968, 6, 463-86.
- E. E. Hagen, "An Economic Justification of Protectionism," *Quart. J. Econ.*, Nov. 1958, 72, 496-514.
- J. R. Harris and M. P. Todaro, "Urban Unemployment in East Africa: An Economic Analysis of Policy Alternatives," *East Afr. Econ. Rev.*, Dec. 1968, 4, 17-36.
- _____ and _____, "Wages, Industrial Employment, and Labour Productivity: The Kenyan Experience," *East Afr. Econ. Rev.* (New Series), June 1969, 1, 29-46.
- J. P. Henderson, "Wage Policy in Africa," Paper prepared for delivery at the African Conference on Economics, Temple University, mimeo, April 1968.
- C. R. Hutton, "The Causes of Labour Migration," in Gugler, ed., *Urbanization in Sub-Saharan Africa*, Kampala 1969.
- C. H. C. Kao, K. R. Anschel, and C. K. Eicher, "Disguised Unemployment in Agriculture: A Survey," in C. K. Eicher and L. W. Witt, eds., *Agriculture in Economic Development*, New York 1964, 129-44.
- J. M. Katz, "Verdoorn Effects; Returns to Scale, and the Elasticity of Factor Substitution," *Oxford Econ. Pap.*, Nov. 1968, 20, 342-52.
- L. Lefebvre, "Planning in a Surplus Labor Economy," *Amer. Econ. Rev.*, June 1968, 58, 343-73.
- W. A. Lewis, "Economic Development with Unlimited Supplies of Labour," *Manchester Sch. Econ. Soc. Stud.*, May 1954, 22, 139-91.
- I. M. D. Little, "The Real Cost of Labour, and the Choice Between Consumption and

- Investment," in P. N. Rosenstein-Rodan, ed., *Pricing and Fiscal Policies: A Study in Method*, Cambridge 1964, 77-91.
- ____ and J. A. Mirrlees, *Manual of Industrial Project Analysis*, Vol. II, "Social Cost Benefit Analysis," Paris 1969.
- B. F. Massell and J. Heyer, "Household Expenditure in Nairobi: A Statistical Analysis of Consumer Behaviour," *Econ. Develop. Cult. Change*, Jan. 1969, 17, 212-34.
- L. G. Reynolds, "Wages and Employment in a Labor-Surplus Economy," *Amer. Econ. Rev.*, Mar. 1965, 55, 19-39.
- W. F. Stolper, *Planning Without Facts: Lesson's in Resource Allocation from Nigeria's Development*. Cambridge 1966.
- M. P. Todaro, "A Model of Labor Migration and Urban Unemployment in Less Developed Countries," *Amer. Econ. Rev.*, Mar. 1969, 59, 138-48.
- Nigeria, *Report of the Commission on the Review of Wages, Salary and Conditions of Service of the Junior Employees of the Governments of the Federation and in Private Establishments 1963-64*.

Optimal Taxation and Public Production

I: Production Efficiency

By PETER A. DIAMOND AND JAMES A. MIRRLEES*

Theories of optimal production in a planned economy have usually assumed that the tax system can allow the government to achieve any desired redistribution of property.¹ On the other hand, some recent discussions of public investment criteria have tended to ignore taxation as a complementary method of controlling the economy.² Although lump sum transfers of the kind required for full optimality³ are not feasible today, commodity and income taxes can certainly be used to increase welfare.⁴ We shall therefore examine the maximization of social welfare using

both taxes and public production as control variables. In doing so, we intend to bring together the theories of taxation, public investment, and welfare economics.

There are two main results of the study: the demonstration of the desirability of aggregate production efficiency in a wide variety of circumstances provided that taxes are set at the optimal level; and an examination of that optimal tax structure. It is widely known that aggregate production efficiency is desired as one part of achieving a Pareto optimum. It is also widely known that when the desired Pareto optimum cannot be achieved, aggregate production efficiency may not be desirable. Our conclusion differs from these results in that production efficiency is desirable although a full Pareto optimum is not achieved. In the optimum position, the presence of commodity taxes implies that marginal rates of substitution are not equal to marginal rates of transformation. Furthermore, the absence of lump sum taxes implies that the income distribution is not the best that can be conceived. Yet, the presence of optimal commodity taxes will be shown to imply the desirability of aggregate production efficiency.

This result is similar to that derived by Marcel Boiteux, although he considered an economy where lump sum redistributions of income were possible. Boiteux also examined the optimal tax structure that was necessary for this result. The optimal tax structure for the case of a single consumer (or equivalently with lump sum redistribution) has also been examined by Frank

* The authors are at Massachusetts Institute of Technology and Nuffield College, Oxford, respectively. During some of the work, Diamond was at Churchill College, Cambridge and Nuffield College, Oxford and Mirrlees was at M.I.T. Earlier versions of this paper were given at Econometric Society winter meetings at Washington and Blaricum, 1967, at the University Social Science Council Conference, Kampala, Uganda, December 1968, and to the Game Theory and Mathematical Economics Seminar, Hebrew University, Jerusalem. The authors wish to thank M.A.H. Dempster, D. K. Foley, P. A. Samuelson, K. Shell, and participants in these seminars for helpful discussions on this subject, and referees for valuable comments. Diamond was supported in part by the National Science Foundation under grant GS 1585. The authors bear sole responsibility for opinions and errors.

¹ For a discussion of this literature, see Abram Bergson.

² For a survey of this literature, see Alan Prest and Ralph Turvey.

³ We wish to distinguish here between lump sum taxes, which may vary from individual to individual while being unaffected by the individual's behavior, and poll taxes which are the same for all individuals, or perhaps for all individuals within several large groups, distinguished perhaps by age, sex, or region.

⁴ For another study of the general equilibrium impact of taxation, which does not explore the optimality question, see Gerard Debreu (1954).

Ramsey and Paul Samuelson.⁵ Our results move beyond theirs in considering the problem of income redistribution together with that of raising revenue. Even in the absence of government revenue requirements, if lump sum redistribution is impossible, the government will want to use its excise tax powers to improve income distribution. It will subsidize and tax different goods so as to alter individual real incomes. The optimal redistribution by this method occurs when there is a balance between the equity improvements and the efficiency losses from further taxation.

The general situation we want to discuss is an economy in which there are many consumers, public and private production, public consumption, and many different kinds of feasible tax instruments. We think that it is easier to understand the problem if we present the analysis first for a single consumer, no public consumption, and only commodity taxation, although this case has little intrinsic interest. The main point of the paper is that the analysis of this special case carries over in the main to the general case.

The first two sections are devoted to this special case. In the first, the situation is portrayed geometrically (for a two-commodity world with no private production); in the second, production efficiency and conditions for the optimal taxes are derived by application of the calculus. The use of the calculus here and elsewhere is not perfectly rigorous for the usual reasons. These issues are taken up in Section IV. In the third section, we extend the analysis of production to an economy with many consumers, elucidating precise conditions under which production efficiency is desirable (and presenting certain exceptions).

⁵ For a detailed history of analysis of this problem, see William Baumol and David Bradford. A summary and discussion of the work of Boiteux has been given by Jacques Drèze.

Section IV provides a rigorous statement of the theorems. In the fifth section, we discuss briefly certain applications and extensions of the basic efficiency result.

A following paper, referred to here as Diamond-Mirrlees II, will appear in the June 1971 *Review*. In it we will examine the optimality rules for commodity taxes, for other taxes including income taxes, and for public consumption. We will also give a rigorous statement of conditions under which the first-order conditions obtained (heuristically) below are indeed necessary conditions.

I. One-Consumer Economy— Geometric Analysis

We begin by considering an economy with a single, price-taking consumer and two commodities. We assume, for the moment, that all production possibilities are controlled by the government. While there is no scope for redistribution of income in this economy, the government might need to raise revenue to cover losses if there are increasing returns to scale or if there are fixed expenditures (such as defense) and constant returns to scale. Alternatively, the technology might exhibit decreasing returns to scale, facing the government with the problem of disposing of a surplus if all transactions are carried out at market prices. The optimal solution to either raising or disposing of revenue is well known. A poll tax or subsidy, as the case may be, will permit the hiring of the needed resources and permit the economy to achieve a Pareto optimum, which, in a one-consumer economy, is equivalent to the maximization of the consumer's utility. While this is a reasonable possibility in a one-consumer economy, lump sum taxes varying from individual to individual do not seem feasible in a much larger economy. An identical problem of distributing a surplus among many people arises if it is desired to improve income distribution.

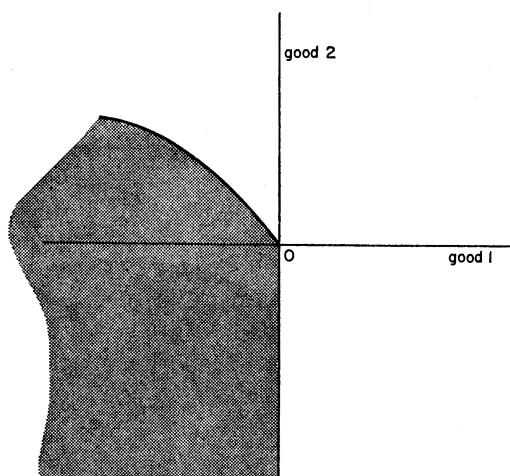


FIGURE 1

Thus we shall consider the use of commodity taxes when lump sum taxes are not permitted to the government, not for the intrinsic interest of this question in a one-consumer economy, but as an introduction to the many-consumer case. Furthermore we shall hold constant the government expenditure pattern which directly affects consumer utility. Thus we can ignore it, since the utility function already reflects its impact. The addition of choice for public consumption will be considered in Diamond-Mirrlees II.

Assuming free disposal, the technological constraint on the planner is that the government supply be on or under the production frontier. Such a constraint is shown by the shaded area in Figure 1. Let us measure on the axes the quantities supplied to the consumer. Thus, the output being produced (good 2) is measured positively, while the input (good 1) is measured negatively. The case drawn is the familiar one of decreasing returns to scale. If the government needed a fixed bundle of resources, for national defense say, then the production possibility frontier (describing the potential transactions with the consumer) would not pass through

the origin. With constant returns to scale this might appear as in Figure 2, where a units of good 1 are needed for defense. (It is perhaps convenient to think of good 1 as labor and good 2 as a consumption good.)

In a totally planned economy, where the planner selects a fixed consumption bundle (including labor to be supplied) for each consumer, the planner would have no further constraint and could choose any point that was technologically feasible. Again, this is not implausible for the planner in a one-consumer economy, but becomes so as the number of households grows. A more realistic assumption, then, is to assume that the planner can only deal with consumers through the market place, hiring labor and selling the consumer good. Assume further that the planner is constrained to charge uniform prices. The planner must now set the price of the consumer good relative to the wage (or inversely the real wage), and is constrained to transactions which the consumer is willing to undertake at some relative price. The locus of consumption bundles which the consumer is willing to achieve by trade from the origin is the offer curve or price-consumption locus. It represents the

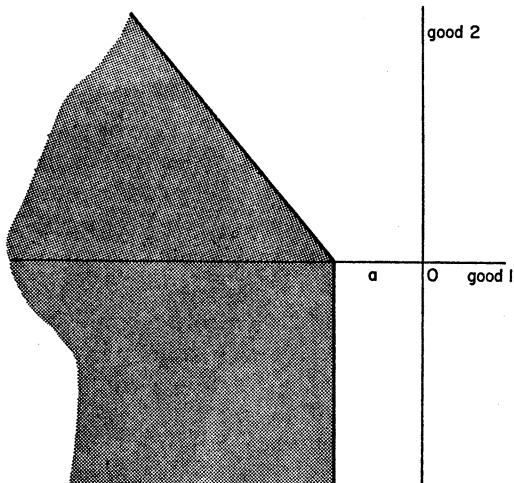


FIGURE 2

bundles of goods that the consumer would purchase at different possible price ratios. Figure 3 contains an example of an offer curve with several hypothetical budget lines and the corresponding indifference curves drawn in. The planner thus has two constraints: he must choose a point which is both technologically feasible and an equilibrium bundle from the point of view of the consumer. Combining these two constraints, the range of consumption bundles which are both feasible and potential consumer equilibria is shown as the heavy line in Figure 4.

We can state these two constraints algebraically. Let us denote by $z = (z_1, \dots, z_n)$ the vector of government supply. The production constraint is then written

$$(1) \quad G(z) \leq 0, \text{ or, equivalently,}$$

$$z_1 \leq g(z_2, z_3, \dots, z_n)$$

The constraint that the government sup-

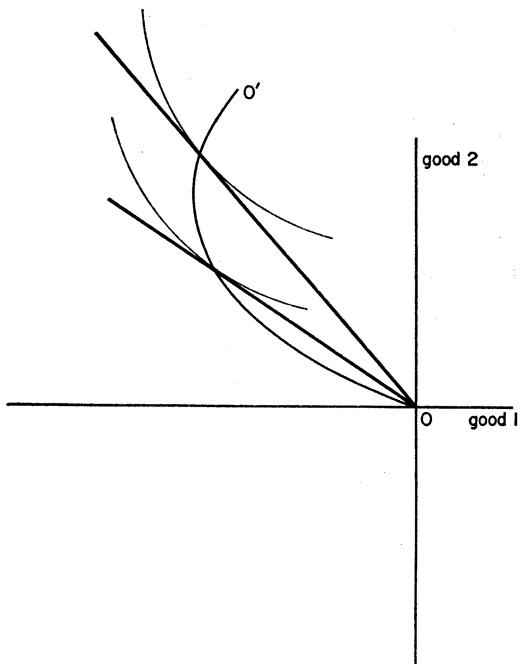


FIGURE 3

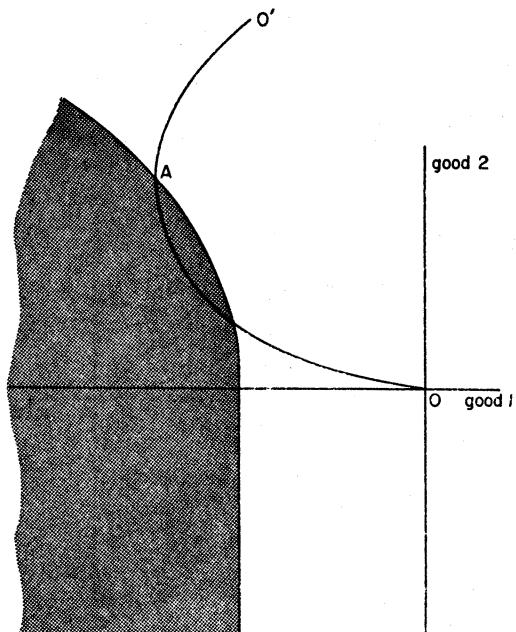


FIGURE 4

ply equal the consumer demand for some price can be written in vector notation

$$(2) \quad x(q) = z,$$

where $x = (x_1, \dots, x_n)$ is the vector of consumer demands and $q = (q_1, \dots, q_n)$ is the vector of prices faced by the consumer.

Now consider the government's objectives. Since the consumer's equilibrium position is determined by the prices he faces, we can, in the usual circumstances, describe the objective function as a function of prices, say $v(q)$. The problem is to choose q so as to

$$(3) \quad \text{Maximize } v(q)$$

$$\text{subject to } G(x(q)) \leq 0$$

This simply formulated problem is the focus of attention of the paper and can take on a variety of interpretations. The reader may note that the consideration of many consumers does not alter the form of this problem. This is a major advantage

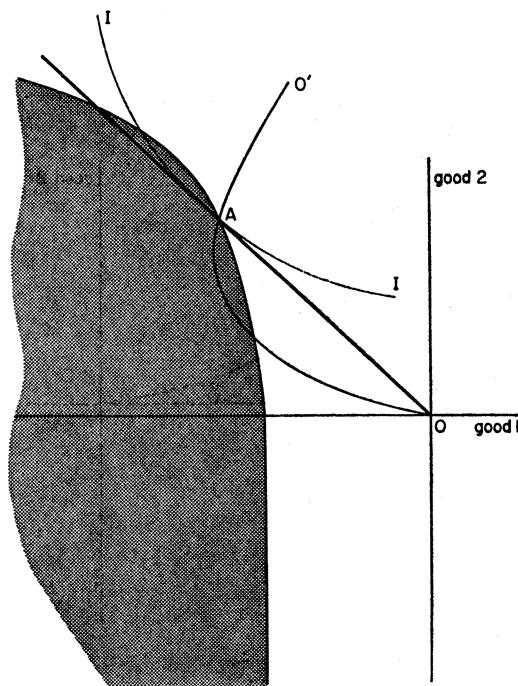


FIGURE 5

of using prices rather than quantities as the focus of the analysis.

Let us consider the case where the planner seeks to maximize the same function of consumption as the consumer's utility function. The welfare function is said to be *individualistic*, or to respect individual preferences, since welfare can be written as a function of individual utility. Returning to Figure 3 we see that the consumer moves to higher indifference curves as he proceeds along the offer curve away from the origin. Thus, in Figure 4 we wish to move as far along OO' as possible, subject to the constraint of the shaded production possibility set. The optimal point is therefore A , where the offer curve and the production frontier intersect.

The prices which will induce the consumer to purchase the optimal consumption bundle are defined by the budget line OA . In Figure 5 we show the optimal point and the implied budget line, and indiffer-

ence curve II . All the points above II and in the shaded production set are Pareto-superior to A and technologically feasible, but not attainable by market transactions without lump sum transfers. For contrast, in Figure 6, we show the Pareto optimal point, B , and the implied budget line, and indifference curve $I'I'$, which will permit decentralization. In the case drawn, the consumer's budget line does not pass through the origin; this represents his payment of a lump sum tax to cover government expenditures in excess of profits from production.

We see that the optimal point is on the production possibility frontier of the economy, not inside it. This important property of the optimum can easily be seen to carry over to the case of many commodities, but still one consumer. With many commodities, the offer curve is a union of loci, each of which is obtained by holding the prices of all but one commodity constant and varying the price of that one commodity. Doing this for each com-

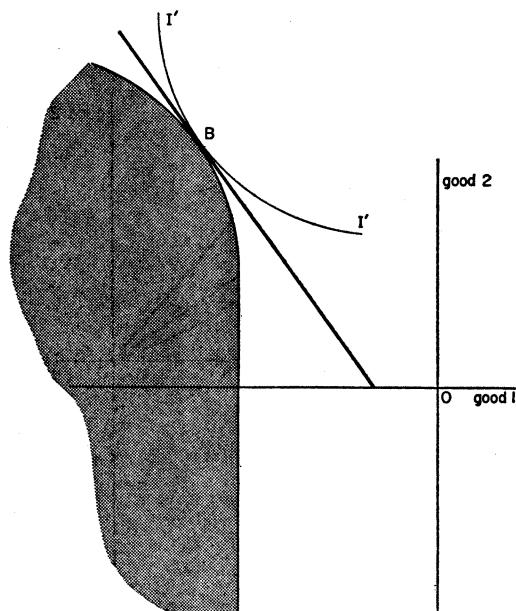


FIGURE 6

modity, and for all possible configurations of prices for the other commodities, generates all the loci. The offer curve is the union of such loci. On each locus, the point which is also on the production frontier is better than the other points on the locus. Thus, any point which is not on the production frontier is dominated by some point which is on the frontier. Therefore, the optimal point is one of the points on the frontier. The implications of this result will be seen more clearly below, when we consider both public and private production. For this result to carry over to the case of many consumers requires one further, mild assumption which will be discussed in the third section. First, we treat the one consumer economy algebraically, with both public and private production, showing by calculus the desirability of aggregate production efficiency, and obtaining the optimal relationship between consumer prices and the slope of the production possibilities. This relationship defines the optimal tax structure.

II. One-Consumer Economy—Algebraic Analysis

We assume constant returns to scale in the private production sector and the presence of competitive conditions there. In equilibrium there are, therefore, no profits. (This is a critical assumption for the efficiency analysis.) We also assume, for the present, that the only taxes used by the government are commodity taxes.⁶ Consumer prices, q , therefore determine the choices available to the consumer, and we may write the welfare function as a function of consumer prices, $v(q)$. Notice that this covers the case where the government's assessment of welfare does not coincide with the consumer's utility, al-

⁶This assumption is made solely for simplicity. In Diamond-Mirrlees II, the general principles will be seen to carry over with additional taxes, including a progressive income tax.

though depending on what he consumes. In the special case where social preferences coincide with those of the single consumer, his utility may be taken to measure welfare. Then we have

$$(4) \quad v(q) = u(x(q))$$

We shall not use this special form for $v(q)$ in the analysis below until we come to evaluate the tax structure explicitly. Until that point, the analysis applies also to welfare functions that are not individualistic. For later use let us express the derivatives of v in this special case. Writing $v_k = \partial v / \partial q_k$, $u_i = \partial u / \partial x_i$, and using (4), we have

$$(5) \quad v_k = \sum u_i \frac{\partial x_i}{\partial q_k} = -\alpha x_k,$$

where α is a positive constant (i.e., independent of k), the marginal utility of income. Equation (5) follows from the budget constraint,

$$(6) \quad \sum q_i x_i = 0,$$

which on differentiation with respect to q_k yields

$$(7) \quad x_k + \sum q_i \frac{\partial x_i}{\partial q_k} = 0$$

Since utility-maximization implies that $u_i = \alpha q_i$, (5) now follows from (7).

Production

Let us denote the vector of prices faced by private producers by $p = (p_1, \dots, p_n)$. Because of taxes, t , these may differ from the prices faced by consumers: $q_i = p_i + t_i$ ($i = 1, \dots, n$). $y = (y_1, \dots, y_n)$ is the vector of commodities privately supplied (inputs will thus appear as negative supplies), and we write the private production constraint,

$$(8) \quad y_1 = f(y_2, \dots, y_n)$$

Notice that we assume *equality* in the

production constraint, that is, that production is efficient in the private sector. This follows from profit maximization if there are no zero prices. We assume that f is a differentiable function, and that $y_i \neq 0$ ($i = 1, \dots, n$). Then, profit maximization means that

$$(9) \quad p_i = -p_1 f_i(y_2, \dots, y_n), \quad (i = 2, \dots, n)$$

where f_i denotes the derivative of f with respect to y_i . Also, by the assumption of constant returns to scale, maximized profits are zero in equilibrium:

$$(10) \quad \sum p_i y_i = 0$$

So that we may conveniently employ calculus, we shall assume that the government production constraint, (1), is satisfied with an equality rather than an inequality:

$$(11) \quad z_1 = g(z_2, \dots, z_n)$$

Thus we do not give the government the option of inefficient government production. Rather, we shift our attention to aggregate production efficiency. Efficiency will be present if marginal rates of transformation are the same in publicly and privately controlled production. It will then be seen quite easily that the assumption of efficiency in the public sector is justified.

Walras' Law

We have chosen an objective function and expressed the government's production constraint above. To complete the formulation of the maximization problem, it remains to add the requirement that the economy be in equilibrium. The conditions that all markets clear can be stated in terms of the vectors x , y , and z .

$$(12) \quad x(q) = y + z$$

The reader may be puzzled that at no place in this formulation has a budget

constraint been introduced for the government. (Other readers may be puzzled by our failure to include only $n-1$ markets in our market clearance equations. These are aspects of the same phenomenon.) Walras' Law implies that if all economic agents satisfy their budget constraints and all markets but one are in equilibrium, then the last market is also in equilibrium. It also implies that when all markets clear and all economic agents but one are on their budget constraints, then the last economic agent is on his budget constraint. In setting up our problem, we have assumed that the household and the private firms are on their budget constraints. Thus, if we assume that all markets clear, this will imply that the government is satisfying its budget constraint,⁷ which we can express as

$$(13) \quad \sum (q_i - p_i)x_i + \sum p_i z_i = 0 \\ = \sum t_i x_i + \sum p_i z_i$$

Alternatively, if we consider the government budget balance as one of the constraints, then it is only necessary to impose market clearance in $n-1$ of the markets.

In this model we can make two price normalizations, one for each price structure. Since both consumer demand and firm supply are homogeneous of degree zero in their respective prices, changing either price level without altering relative prices leaves the equilibrium unchanged. As normalizations let us assume,

$$(14) \quad p_1 = 1, \quad q_1 = 1, \quad t_1 = 0$$

It may seem surprising that it does not matter whether the government can tax good one. But the reader should remember the budget balance of the consumer. Since there are no lump sum transfers to the

⁷ In an intertemporal interpretation of this model, the government budget is in balance over the horizon of the model, not year by year.

consumer, net expenditures are zero. Thus, levying a tax at a fixed proportional rate on all consumer transactions results in no revenue. (It should be noticed that a positive tax rate applied to a good supplied by the consumer is in effect a subsidy and results in a loss of revenue to the government.)

Welfare Maximization

We can now state the maximization problem. In the statement we shall use the two sets of prices as control variables. It would be a more natural approach to use the taxes which the government actually controls as decision variables. However, once we have determined the optimal p and q vectors we have determined the optimal taxes. Using taxes as decision variables complicates the mathematical formulation and leads to a control problem since the tax vector may not uniquely determine equilibrium.

Rather than calculate the first-order conditions from the formulation spelled out above, we shall alter the problem to simplify the derivation. We have to choose

$$(15) \quad q_2, \dots, q_n, \quad p_2, \dots, p_n, \quad z_1, \dots, z_n$$

to maximize $v(q)$ subject to

$$x_i(q) - y_i - z_i = 0 \quad (i = 1, 2, \dots, n),$$

where y maximizes $\sum p_i y_i$ subject to

$$y_1 = f(y_2, \dots, y_n),$$

and

$$z_1 = g(z_2, \dots, z_n)$$

Since the choice of producer prices can be used to obtain any desired behavior on the part of private producers, we can use any vector y consistent with the production constraint (8). Producer prices are then determined by equation (9). Using the equations

$$y_2 = x_2 - z_2, \dots, y_n = x_n - z_n,$$

we reduce the constraints in (15) to the

single constraint

$$\begin{aligned} x_1(q) &= y_1 + z_1 \\ &= f(x_2 - z_2, \dots, x_n - z_n) + g(z_2, \dots, z_n) \end{aligned}$$

We have therefore simplified the problem (15) to:

$$(16) \quad \text{Choose } q_2, \dots, q_n, \quad z_2, \dots, z_n$$

to maximize $v(q)$ subject to

$$\begin{aligned} x_1(q) - f(x_2(q) - z_2, \dots, x_n(q) - z_n) \\ - g(z_2, \dots, z_n) = 0 \end{aligned}$$

Forming a Lagrangian expression from (16), with multiplier λ ,

$$\begin{aligned} L &= v(q) - \lambda[x_1(q) \\ (17) \quad &\quad - f(x_2 - z_2, \dots, x_n - z_n)] \\ &\quad - g(z_2, \dots, z_n)], \end{aligned}$$

we can differentiate with respect to q_k :

$$\begin{aligned} (18) \quad v_k - \lambda \left(\frac{\partial x_1}{\partial q_k} - \sum_{i=2}^n f_i \frac{\partial x_i}{\partial q_k} \right) &= 0 \\ k &= 2, 3, \dots, n \end{aligned}$$

Making use of the equations (9) for producer prices, this can be written

$$\begin{aligned} (19) \quad v_k - \lambda \sum_{i=1}^n p_i \frac{\partial x_i}{\partial q_k} &= 0 \\ k &= 2, 3, \dots, n \end{aligned}$$

Differentiating L with respect to z_k we have

$$(20) \quad \lambda(f_k - g_k) = 0 \quad k = 2, 3, \dots, n$$

Provided that λ is unequal to zero (i.e., that there is a social cost to a marginal need for additional resources), equation (20) implies equal marginal rates of transformation in public and private production and thus aggregate production efficiency as was argued above. The assumption that $\lambda \neq 0$ needs justification. This is provided by the rigorous arguments of Sections III and IV.

If we had introduced *several* public

production sectors, each described by a constraint like (11), we should have obtained an equation of the form (20) for each sector. Thus marginal rates of transformation in all public sectors should be equal, since they are all to be equal to the private marginal rates of transformation. This argument—which we only sketch here, since the conclusion will be proved more directly in the next section—justifies our assumption that there should be production efficiency in the public sector.

Optimal Tax Structure

The relations (19) determine the optimal tax structure, since they show how producer and consumer prices should be related. These equations show that consumer prices should be at a level such that further increases in any price result in an increase in social welfare, v_k , which is the same ratio, λ , to the cost of satisfying the change in demand arising from the price increase. Reintroducing taxes explicitly into the problem we can obtain an alternative interpretation for the first-order conditions.

Since x_i is a function of $p+t$,

$$\frac{\partial x_i}{\partial q_k} = \frac{\partial x_i}{\partial t_k}$$

(p is held constant in this latter derivative.) Consequently, the optimal tax structure, (19), can be rewritten:

$$(21) \quad v_k = \lambda \sum p_i \frac{\partial x_i}{\partial t_k} = \lambda \frac{\partial}{\partial t_k} \sum p_i x_i$$

Since $\sum p_i x_i = \sum q_i x_i - \sum t_i x_i = - \sum t_i x_i$ (by the consumer's budget constraint (6)), we have

$$(22) \quad v_k = - \lambda \frac{\partial}{\partial t_k} (\sum t_i x_i)$$

This last set of equations asserts the

proportionality of the marginal utility of a change in the price of a commodity to the change in tax revenue resulting from a change in the corresponding tax rate, calculated at constant producer prices. Like the first-order conditions for the optimum in standard welfare economics, our first-order conditions are expressions in constant prices. The tax administrator, like the production planner, need not be concerned with the response of prices to government action when looking at the first-order conditions.

If we now make the further assumption that the welfare function is individualistic, we can use equation (5) to replace v_k . The first-order conditions then become

$$(23) \quad x_k = \frac{\lambda}{\alpha} \frac{\partial (\sum t_i x_i)}{\partial t_k}$$

Thus for all commodities the ratio of marginal tax revenue from an increase in the tax on that commodity to the quantity of the commodity is a constant. This form of the first-order conditions has the advantage of showing the information needed to test whether a tax structure is optimal. The amount of information does not seem excessive relative to the data and knowledge which a planner in an advanced country should have.

The statements of the first-order conditions thus far do not directly indicate the size of the tax rates required, nor the impact upon demand that the optimal tax rates would have. In his pioneering study of optimal tax structure, Frank Ramsey manipulated the first-order conditions so as to shed light on the latter question. He employed the concept of demand curves calculated at a constant marginal utility of income. Paul Samuelson reformulated this using the more familiar demand curves calculated at a constant level of utility. We shall return to this question in Diamond-Mirrlees II.

III. Production Efficiency in the Many-Consumer Economy

We have remarked already that many of the results carry over directly to an economy of many consumers, even when lump sum taxation is excluded. We notice at once that the device of expressing welfare as a function of the prices, q , faced by consumers can be used perfectly well. Explicitly, we assume that there are H households, with utility and demand functions u^h and x^h ($h = 1, 2, \dots, H$). If, as we may generally suppose, in the absence of externalities from producers to consumers, social welfare can be expressed as a function of the consumption of the various consumers in the economy, $U(x^1, x^2, \dots, x^H)$, it may also be written

$$(24) \quad V(q) = U(x^1(q), x^2(q), \dots, x^H(q)),$$

where we assume that there are no lump sum incomes or transfers that would be influenced by producer prices or government policy. In the case where social welfare depends only on individual utility and there are no externalities, we can write

$$(25) \quad V(q) = W[u^1(x^1(q)), u^2(x^2(q)), \dots, u^H(x^H(q))],$$

where W is presumed to be strictly increasing in each of its arguments.

Using this indirect welfare function, we can carry out the analysis already presented for the one-consumer economy, and conclude in the same way that aggregate production efficiency is desirable. For that argument to be correct, we must confirm that the Lagrange multiplier λ is not zero. Rather than attempt to do this directly, we shall present a different argument for the desirability of production efficiency. A further condition will be required to secure our conclusion. In considering this problem, we shall concentrate on the case where all production is under government con-

trol. The desirability of production efficiency in this case will be seen to imply the same conclusion when there is also a private sector (provided that private producers are price takers, and profits, if any, are transferred to the government). Assume then (as we did in Section I) that all production takes place in the public sector: our problem is to find q that will

$$(26) \quad \text{Maximize } V(q),$$

$$\text{subject to } G(X(q)) \leq 0,$$

where we define $X(q) = \sum_h x^h(q)$ as aggregate demand at prices q . We shall also express the production constraint a little more generally by saying that $X(q)$ is to belong to the production set G , the set of technologically feasible production plans. (Thus the letter G denotes both the production set, and also the function that can be used to describe it; but we shall hardly ever use the *function* G explicitly).

Suppose we establish that, at the optimum for problem (26), production is efficient. Consider an economy with the same technological possibilities, partly under the control of private, competitive producers. The government can induce private firms to produce any efficient net output bundle by suitable choice of producer prices p . In particular, it can obtain the production plan that would be optimal if the government controlled all production. The choice of p does not affect consumer demands or welfare, since pure profit arising from decreasing returns to scale go to the government, and since, any commodity taxes being possible, q can be chosen independently of p . Thus, if the solution to (26) is efficient, the same equilibrium can be achieved when some production is under private control, and is optimal in that case too. Proof that production efficiency is desirable in the "special" case (26) therefore implies that pro-

duction efficiency is desirable in the more general case.

Examples of Inefficiency

Before considering the argument for efficiency, it is useful to consider some limitations on that argument as demonstrated by the following examples of desired inefficiency. It will be recollected that a production plan is efficient if any other feasible production plan provides a smaller net supply of at least one commodity. We shall use a different concept: we say that a production plan is *weakly efficient* if it is on the production frontier. It is possible for a production plan to be weakly efficient without being efficient if the production frontier has vertical or horizontal portions. For matters of economic importance, such as the existence of shadow prices, weak efficiency is all that is required. It is easy to see that if all the prices corresponding to a weakly efficient production plan are positive, the plan is in fact efficient in the usual sense.

Even with this slightly weakened concept of efficiency, it is not necessarily true that, when an optimum exists, optimal production has to be weakly efficient. We present two examples.

Example a is portrayed in Figure 7. It is a one-consumer economy where social preferences, as depicted in the social indifference curve II' , do not coincide with individual preferences. It is evident that, in the case shown, the optimal production plan is actually in the interior of the production set.

In the second example, social preferences do respect household preferences, but again optimal production lies in the interior of the production set, and is therefore not weakly efficient: suitable producer prices cannot be found, and the social optimum cannot be obtained when there is private control of production.

Example b. There are two commodities and two households. One has utility function x^2y , the other has utility function xy^2 ; each has the nonnegative quadrant $\{(x, y) | x \geq 0, y \geq 0\}$ as consumption set. The first consumer has three units of the first commodity initially; the second, one unit of the second commodity. The welfare function is

$$-\frac{1}{x_1 y_1} - \frac{1}{x_2 y_2}$$

The second commodity can be transformed into the first according to the production relation $x + 10y \leq 0$, ($x \geq 0$). Let the prices of the commodities be q_1, q_2 . Then the first household's net demands are

- 1 of the first commodity,
- q_1/q_2 of the second commodity.

The second household has net demands

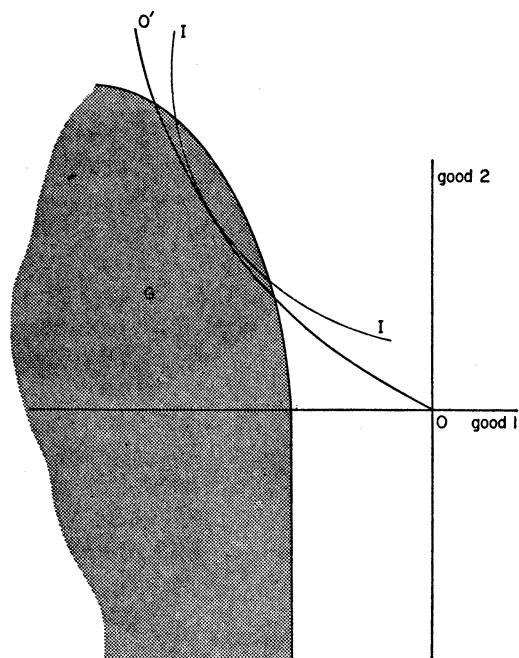


FIGURE 7

$$\frac{1}{3}(q_2/q_1) \text{ and } -\frac{1}{3}$$

Thus, the net market demand for the commodities is

$$x = \frac{1}{3}(q_2/q_1) - 1 \text{ and } y = (q_1/q_2) - \frac{1}{3}$$

These must satisfy

$$x + 10y \leq 0, \quad x \geq 0$$

Welfare is $-q_2/4q_1 - 27q_1/4q_2$ which is maximized when $q_2/q_1 = 3\sqrt{3}$: the corresponding production vector $\sqrt{3}-1, \frac{1}{3}(\sqrt{\frac{1}{3}}-1)$ is actually interior to the production set, not on the frontier. This example has the unimportant peculiarity that initial endowments of the consumers are on the frontiers of their consumption sets. More complicated examples avoiding the peculiarity have been constructed.

The Efficiency Argument

Despite these examples, the following argument shows that optimal production will generally be on the production frontier. Suppose that the aggregate demand functions, $X(q)$, are continuous. Then any small change in the prices, q , will not change aggregate production requirements by much. Therefore, if optimal production were in the interior of the production set, small changes in consumer prices would still result in technologically feasible aggregate demands. Thus, if we are at the optimum, small changes in consumer prices cannot increase welfare. If we can argue that, at the optimum, there exists a small price change which would increase $V(q)$, we can conclude that production for the optimum must occur on the production frontier. For any unsatisfied single consumer, utility can be increased either by lowering the price of a supplied good or raising the price of a demanded good (as we can see, algebraically, in equation (5)). With a single consumer, we need not argue further, provided the equilibrium involves some trade. When there are many con-

sumers, we can be certain of increasing welfare if we raise some consumer's utility without lowering that of anyone else. If there is a commodity that no consumer purchases, but some consumer supplies (such as certain labour skills); or a good (with positive price) which no consumer supplies, but some consumer purchases (such as electricity), we could alter the price of that commodity in such a way as to bring about an unambiguous increase in welfare. In that case, we conclude that efficient production is required for the maximization of individualistic social welfare. In example *b*, it will be seen that neither of the commodities is supplied, or demanded, by both consumers. The very simplicity of the case appears to be misleading.

A formal presentation of this argument is given in the next section: these technical details can be omitted without loss of continuity. We conclude this section by introducing further taxes into the discussion.

First, consider the case of a poll tax (or subsidy)—that is, a tax is paid by a household on the basis of some unalterable property, such as its sex or age distribution. Such a tax is, of course, a lump sum tax, although its availability is not, in general, sufficient to enable the full optimum to be achieved. To fix ideas, suppose there is a single transfer, τ , to be made to all households. Then welfare can be written $V(q, \tau)$, and we are to

$$(27) \text{ Maximize } V(q, \tau)$$

subject to $X(q, \tau)$ being in G

The standard efficiency argument can be used. Let (q^*, τ^*) be the optimum: if any small change in q or τ would increase V , optimal production, $X(q^*, \tau^*)$ must be on the production frontier (assuming that X is a continuous function). Now a poll subsidy must make everyone better off,

unless some are already satiated, and so must a small increase in subsidy. Thus so long as a poll subsidy is possible (and it surely is) and not every household is satiated, optimal production must be on the frontier.

Adding further tax instruments to the government's armoury in no way weakens the efficiency conclusion. We simply note that if there are other tax variables which are independent of producer prices and quantities, denoted collectively by ζ , we can hold them constant at their optimum values ζ^* , and then apply the efficiency argument to the problem (27) or (26), where V and X are evaluated for $\zeta = \zeta^*$.

Our final conclusion is that whatever the class of possible tax systems, if all possible commodity taxes are available to the government, then in general, and certainly if a poll subsidy is possible, optimal production is weakly efficient. We would not expect this conclusion to be valid if there were constraints on the possibilities of commodity taxation, or more generally, on the possible relationship between producer prices and consumer demand. The presence of pure profits is one example of such a relationship. To show what goes wrong, suppose, by way of another example, that *no* commodity taxes are possible, but a poll tax is possible, and that part of production is privately controlled, in such a way that it is uniquely determined by producer prices. Then we have to choose a public production vector z and a poll tax τ to

$$(28) \text{ Maximize } V(p, \tau)$$

subject to $X(p, \tau) - y(p) = z$ being in G , where $y(p)$ is the private production vector when prices are p . Following the argument used above, we consider τ smaller than τ^* , the optimum level, and note that $V(p^*, \tau) > V(p^*, \tau^*)$. This implies that $X(p^*, \tau) - y(p^*)$ is not in G , and therefore z^* , the optimal z , is efficient in G . But the

argument does not imply that the aggregate optimal production plan, $y(p^*) + z^*$ is efficient. Of course, in an economy where all production is under public control, these problems do not arise. Even when some of the q_k are fixed, the efficiency argument holds, for there can be no necessary relation between q and p .

IV. Theorems on Optimal Production

In this section, we explore the existence of the optimum, and the efficiency of optimal production, rigorously. We rely on Debreu (1959) for the results of general equilibrium theory that are required.

Assumptions

There are H households in the economy, each household choosing a preferred net consumption vector x from his consumption set C subject to the budget constraint $q \cdot x \leq 0$ where q is the vector of prices charged to consumers. (Consumption is measured net of initial endowment for convenience, since the latter is unaltered in the analysis.) As usual the net demand vector x has, in general, both positive and negative components corresponding to purchases and sales by the household.

The assumptions used below will be selected from the following list (the superscript h refers to the index of households; all assumptions, when made, hold for all h):

- (a.1) C^h is closed, convex, bounded below by a vector a^h , and contains a vector with every component negative.
- (a.2) The preference ordering is continuous.
- (a.3) The preference ordering is strongly convex. Formally, if x^2 is preferred or indifferent to x^1 and $0 < t < 1$, then $tx^2 + (1-t)x^1$ is strictly preferred to x^1 .
- (a.4) There is no satiation consumption in C^h .

Assumptions (a.1) and (a.2) guarantee the existence of continuous utility functions, which we shall write u^h (see Debreu Section 4.6). Furthermore, under (a.1)–(a.3), when the demand vector $x^h(q)$ is defined, it is uniquely defined. When C^h is bounded, assumptions (a.1)–(a.3) imply that $x^h(q)$ is defined and continuous at all non-zero nonnegative q . (See Debreu, Section 4.10.)

Let us denote aggregate demand by $X(q) = \sum_h x^h(q)$.

It is assumed that all production is controlled by the government. The assumptions on the production possibility set, G , will be taken from the following set:

- (b.1) Every production plan in which nothing is produced in a positive quantity is possible: i.e., if $z \leq 0$, z is in G .
- (b.2) Complete inactivity is possible: i.e., 0 is in G .
- (b.3) G is closed.
- (b.4) There exists a vector \bar{a} such that $z \leq \bar{a}$ for all nonnegative z in the convex closure of G . (i.e., the closure of the convex hull of G).⁸
- (b.5) G is convex.

The welfare function will be denoted by $U(x^1, \dots, x^H)$. When demands are functions of prices only we can define the indirect welfare function as

$$V(q) = U(x^1(q), \dots, x^H(q))$$

Similarly we can define an individual's indirect utility function by

$$v^h(q) = u^h(x^h(q))$$

We shall say that the welfare function *respects household preferences* when U can be written

$$U(x^1, \dots, x^H) = W(u^1(x^1), \dots, u^H(x^H))$$

⁸ When G is convex, this assumption is similar to the assumption that inputs are required to obtain outputs, but permits the government to own a vector of inputs.

with W increasing in each argument. We shall assume

- (c.1) U is a continuous function of (x^1, \dots, x^H)

We can now state our problem as trying to find q^* to maximize $V(q)$ subject to $X(q)$ being in G . A commodity vector will be called *attainable* if it is feasible and if there exists prices such that aggregate demand equals the vector. The set of all such vectors, the *attainable set*, is the intersection of G with the set of vectors $X(q)$ for all nonnegative q .

Existence of an Optimum

If we assume that the attainable set is nonempty and bounded, we obtain

THEOREM 1. *If assumptions (a.1)–(a.3), (b.3), and (c.1) hold, and if the attainable set is nonempty and bounded, an optimum exists.*

PROOF:

Consider an economy in which the consumption sets are truncated by removing from them all points x with $\|x\| > M$, where all vectors in the attainable set satisfy $\|x\| < M$. For this truncated economy, the demand functions are continuous at all price vectors not equal to zero. Since the attainable set, and demands for any q corresponding to an attainable vector, are the same in the original and truncated economies, an optimum for the truncated economy is an optimum for the original economy. In other words, we may, without loss of generality, assume that demands are continuous at $q \neq 0$. Since the demand functions are homogeneous of degree zero in the prices, we can restrict our attention to q satisfying $q \geq 0$ and $\sum_i q_i = 1$.

We next demonstrate that the set $\{q | X(q) \text{ in } G\}$ is closed. Let q_n be a sequence of price vectors converging to q' , with $X(q_n)$ in G for all n . Let x' be a limit point of $\{X(q_n)\}$. Since G is closed, x' is

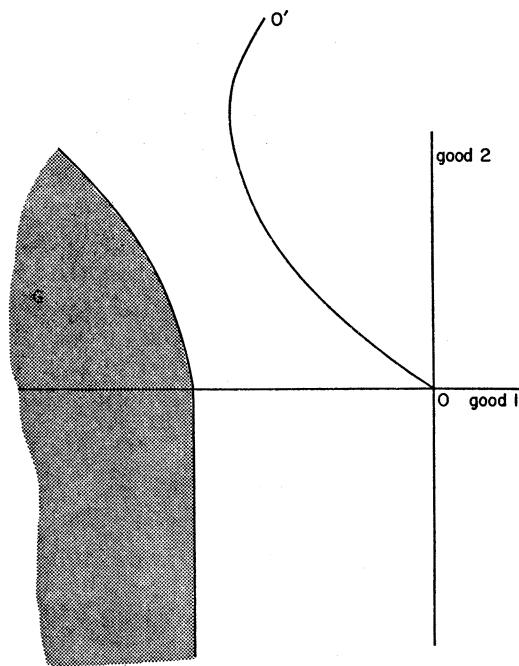


FIGURE 8

in G . At the same time, $x' = X(q')$, by the continuity of X . Thus q' is in $\{q | X(q) \text{ in } G\}$, which is therefore closed.

Since the attainable set is nonempty, and prices are in any case bounded, $\{q | X(q) \text{ in } G\}$ is closed, bounded, and nonempty. By the continuity of the demand functions, and assumption (c.1), V is a continuous function of q , which therefore attains its maximum on the set $\{q | X(q) \text{ in } G\}$.

One criterion for the attainable set to be nonempty follows immediately from the existence of competitive equilibrium in an exchange economy:

THEOREM 2. *If assumptions (a.1)–(a.4) and (b.1) hold, the attainable set is nonempty.*

PROOF:

See Debreu (Section 5.7) for a proof that there exists an equilibrium for the exchange economy with these consumers.

The equilibrium prices result in a feasible demand.

If the production set is taken to be the set of possible production vectors net of government consumption, the assumption that zero production is possible is excessively strong, especially for governments with large military establishments. But it is easy to construct examples of economies not satisfying (b.1) in which there is no attainable point. Consider the one-consumer economy depicted in example c shown in Figure 8.

The boundedness of the attainable set would be implied by the boundedness of the consumption sets, or the boundedness of production, but the following case is more appealing:

THEOREM 3. *If assumptions (a.1) and (b.2)–(b.4) hold, then the attainable set is bounded.⁹*

PROOF:

Suppose the attainable set is not bounded. Then there exists a sequence of attainable vectors x_n such that $\|x_n\|$ is an unbounded increasing sequence of real numbers. There exists an n' such that $\|x_{n'}\| > \|\bar{a}\|$, where \bar{a} is the vector employed in (b.4). Consider the sequence of vectors $(\|x_{n'}\|/\|x_n\|)x_n$ for $n \geq n'$. Each vector is in the convex hull of G (being a convex combination of the origin and x_n). Further the sequence is bounded. Thus there is a limit point, ξ , which is in the convex closure of G and satisfies $\|\xi\| > \|\bar{a}\|$. Let $b = \sum_h a_h$, where a_h are the vectors employed in (a.1). Then $x_n = \sum_h x_n^h \geq \sum_h a_h = b$. Further $(\|x_{n'}\|/\|x_n\|)x_n \geq (\|x_{n'}\|/\|x_n\|)b$. But the latter sequence of vectors converges to zero. Thus $\xi \leqq 0$. This is a contradiction.

⁹ The attainable set will also be bounded if (b.2)–(b.4) hold for the true production set, gross of government consumption, rather than the net production set, G . Thus the assumption that zero production is possible is not of great consequence.

Finally, we should remark that the strong convexity assumption, (a.3), which was made in Theorem 1 can be changed to convexity without affecting the conclusion. All that is required is to replace the continuous functions of the proof by upper semi-continuous correspondences. On the other hand, one can easily construct examples in which an optimum fails to exist because of the absence of continuity.

Efficiency

The following lemma provides two criteria for optimal production to be on the frontier of the production set. It will be used to deduce a theorem about the case where household preferences are respected.

LEMMA 1: *Assume an optimum, q^* , exists. If aggregate demand functions and the indirect welfare function are continuous in the neighborhood of the optimal prices; and if either*

- (1) *for some i , V is a strictly increasing function of q_i in the neighborhood of q^* ; or*
 - (2) *for some i with $q^* > 0$, V is a strictly decreasing function of q_i in the neighborhood of q^* ,*
- then $X(q^*)$ is on the frontier of G .*

PROOF:

Let l_i be the vector with all zero components except the i th, which is one. In case 1, for ϵ sufficiently small $V(q^* + \epsilon l_i) > V(q^*)$. Hence $X(q^* + \epsilon l_i)$ is not in G . Letting ϵ decrease to zero, the continuity of X shows that $X(q^*)$ is a limit of points not in G , and therefore belongs to the boundary of G . In case 2, a similar argument can be made using $V(q^* - \epsilon l_i)$.

These conditions are weak. They are, naturally, independent of production possibilities. It may also be noticed that, when V is a differentiable function of prices, the stated conditions are equivalent to assuming that

$$(29) \quad \text{It is not the case that } V'(q^*) \leq 0$$

Here $V'(q)$ is the vector of first derivatives of V with respect to prices. The equivalence of the conditions of the theorem and (29) is clear if we remember that

$$(30) \quad V'(q) \cdot q = \sum \frac{\partial V}{\partial q_k} q_k = 0,$$

since V is homogeneous of degree zero in q . Therefore $V' \leq 0$ if, and only if, $\partial V / \partial q_k = 0$ when $q_k > 0$ and $\partial V / \partial q_k \leq 0$ in any case.

In the following theorem, we strengthen the assumptions in a different way: they remain notably weak.

THEOREM 4. *If (a.1)–(a.4) and (c.1) hold; if social welfare respects individual preferences; and if either*

- (1) *for some i , $x_i^h \leq 0$ for all h , and $x_i^h < 0$ for some h' ; or*
- (2) *for some i , with $q_i > 0$, $x_i^h \geq 0$ for all h and $x_i^h > 0$ for some h' ;*

Then if an optimum exists, production for the optimum is on the frontier of the feasible set.

PROOF:

Individual demand functions are continuous in the neighborhood of the optimum and thus aggregate demands and the indirect welfare function are continuous. Since social welfare respects preferences, indirect social welfare can be written as an increasing function of indirect utilities. In case 1, indirect utilities are a nondecreasing function of q_i in the neighborhood of q^* for all h while the indirect utility function of h' is strictly increasing in q_i . Thus V increases with q_i . Case 2 follows similarly.

The assumption of strictly convex preferences made in Theorem 4 is required in the theorem as stated.

Example d: Consider an economy with one consumer whose indifference curves have

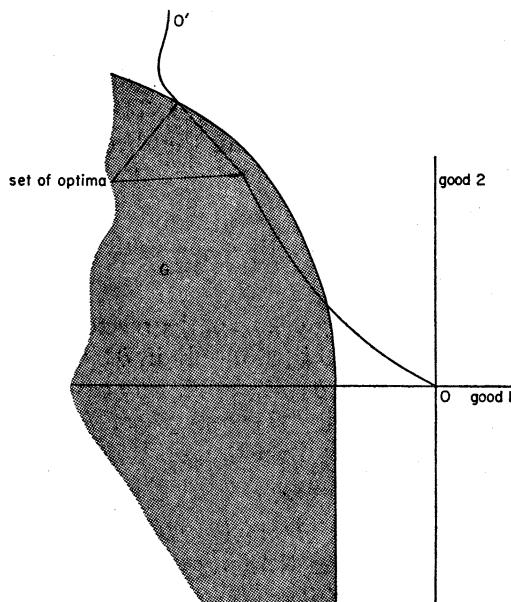


FIGURE 9

a linear section. Then the offer curve may coincide with the linear part of an indifference curve, giving a set of optima, only one of which is on the production frontier. As an illustration, see Figure 9.

The example suggests that we weaken the conclusion of Theorem 4 to say that there exists an optimum on the frontier of G : this generalization is indeed correct if we merely assume convexity of preferences. The proof follows that of Theorem 4, with upper semi-continuity of the demand correspondence replacing continuity of demand functions.

V. Extensions

We can summarize the efficiency result by considering an economy with three sectors—consumers, private producers, public producers. We assumed that only the equilibrium position of the consumer sector enters the welfare function, and that only market transactions take place between sectors, while the government has power to tax any intersector transaction

at any desired rate. One conclusion was that all sectors not containing consumers should be viewed as a single sector, and treated so that aggregate production efficiency is achieved. By regrouping the parts of the economy according to this schematic division, we can extend the efficiency result to several other problems. In each case, we indicate briefly how application of this schematic view shows the relationship of the extension to the basic model.

Intermediate Good Taxation

The model, as presented above, left no scope for intermediate good taxation. If we separate private production possibilities into two (or many) sectors, we introduce the possibility of taxing transactions between firms. In the schematic view presented above, we could consider a consumer sector and two, constant returns to scale, private production sectors. We conclude that we want efficiency for these private production possibilities taken together. Therefore the optimal tax structure includes no intermediate good taxes, since these would prevent efficiency. (Similarly we conclude that government sales to firms should be untaxed while those to consumers are taxed.)

There is a straightforward interpretation of this result, which helps to explain the desirability of production efficiency. In the absence of profits, taxation of intermediate goods must be reflected in changes in final good prices. Therefore, the revenue could have been collected by final good taxation, causing no greater change in final good prices and avoiding production inefficiency. This interpretation highlights the necessity of our assumption of constant returns to scale in privately controlled production.

However, it may well be desirable to tax transactions between consumers or to charge different taxes on producer sales to

different consumers. There are two ways in which we can consider doing this. The country might be geographically partitioned with different consumer prices in different regions. Ignoring migration, and consumers making purchases in neighboring regions, our analysis can be applied to determine taxes region by region. In general the tax structure will vary over the country.

Alternatively, we might consider taxation on all consumer-consumer transactions. Here, too, we would expect to be able to increase social welfare by having these additional tax controls. Neither addition to the available tax structure alters the desirability of production efficiency.

Untaxable Sectors

One problem that arises with a model considering taxation of all transactions is that some transactions may not be taxable, practically or legally. An example of the former might be subsistence agriculture where transactions with consumers are hard to tax while those with firms are not. If the introduction of other taxes (e.g., on land or output) is ruled out, we can accommodate this problem in the model by including subsistence agriculture in the consumer rather than producer sector (or treating it as a second consumer sector). Efficiency would then be desired for the modern and government production sectors taken together; while the tax structure rules would be stated in terms of demand derivatives of the augmented consumer sector rather than of just the true consumers.

Similarly, in an economy without taxes, a public producer subject to a budget constraint is unable to charge different prices to consumers and producers. Lumping together the entire private sector as a single consumer sector, we obtain the conditions for optimal public production of an industry regulated in this manner. This

is the problem considered by Boiteux in the context of costless income redistribution. He also analyzed such an economy with several firms, each limited by a budget constraint.

Foreigners

It is not easy to provide a satisfactory welfare economics for a world of many countries. The study of world welfare maximization is interesting, and, one may hope, "relevant." But it has the serious limitation that its results can seldom be applied to the actions of governments. However altruistic the principles on which a government seeks to act, it has to allow for the actions other governments may take, based on different principles, or for different reasons. (A somewhat analogous problem arises in intertemporal welfare economics.) In the following two subsections, we shall, in order to keep the discussion brief, refer only to the case where the reactions of all other countries are well-defined functions of the actions of the country directly considered. Thus we neglect, reluctantly, those situations that have come to be called "game-theoretic." Also, we shall not consider the problem of formulating a social welfare function in an international setting.

International Trade

So long as we are completely indifferent to the welfare of the rest of the world, and so long as the reactions of other countries are well-defined, international trade simply provides us with additional possibilities for transforming some goods and services into others. The efficiency result then implies that we would want to equate marginal rates of transformation between producing and importing. If there is a monopoly position to be exploited, it should be. If international prices are unaffected by this country's demand, intermediate goods should not be subject to a tariff, but final

good sales direct to consumers should be subject to a tariff equal to the tax on the same sale by a domestic producer.

Sometimes it is not possible to sell goods to foreigners at prices different from those at which they are sold to domestic consumers, although the theory just outlined suggests that foreigners should be treated like producers. As examples, we may cite tourists and commodities covered by special kinds of international agreement. If tourism, say, is an important trading opportunity for the country, and tourists have to be charged the same prices as domestic consumers, this will affect the optimal level of taxes on certain commodities. The general efficiency result is not upset, however. The analysis can be performed by treating tourists as consumers whose income does not affect social welfare.

The authors do not, of course, recommend indifference to the welfare of the rest of the world; although it happens to make the results somewhat neater. International trade provides the country with another set of consumers who can trade with it at prices different from its own consumers: the case (when foreign reactions are well-defined) is similar to the possibility of using different consumer prices in different regions of the same economy. In that case, there is no reason why optimal international trade prices should be the same as producer prices, p , or domestic consumer prices, q .

Migration

In all that has gone before, we have been holding constant the set of consumers in the economy. We can introduce migration in a straightforward manner. Social welfare may be a function of the consumption of every household in the world. Changes in the consumer prices charged in the home country cause migration in one direction or another, and therefore affect wel-

fare in ways we have not previously discussed (such as the effect on the inhabitants of another country of having additional taxpayers join them). But we can still define an indirect welfare function $V(q)$, so long as the reactions of the rest of the world are well-defined. Similarly we can define aggregate demand functions $X(q)$, but these are no longer continuous. For, when a man decides to emigrate, his contribution to aggregate demand changes from x^h to 0.¹⁰ But the number of migrants arising from a small price change may, quite reasonably, be assumed small relative to the population as a whole. We can therefore adequately approximate this situation by considering a continuum of consumers. In this way we can restore continuity to aggregate demand, and to the indirect welfare function. It is to be expected, then, that production efficiency is still desired. Since the derivatives of the demand functions, and possibly also the derivatives of V , will be different when the possibility of migration is allowed for, the optimal tax structure will be changed to reflect the loss of tax revenue when net taxpayers, for example, leave the country. While we do not wish to examine this problem in detail here, we believe that these ideas provide an interesting approach to the analysis.

Consumption Externalities

The schematic view of this problem given above suggests that the basic structure of the results, although not the specific optimal taxes, are unchanged by complications which occur wholly within the consumer sector. Thus, if we introduce consumption externalities that leave aggregate demand continuous we will still obtain production efficiency at the optimum, if we can argue that $V(q)$ has no unconstrained local maximum for finite q .

¹⁰ A similar discontinuity problem arises in the case of tourists' decisions not to visit the country.

The conditions used above are no longer sufficient for this argument since the direct effects of a price change might be offset by the change in the pattern of externalities induced by the price change. Although we have not examined this case in detail, there are a number of cases where arguments similar to those in the no-externality case will be valid.¹¹ Furthermore it seems quite likely to us that efficiency will be desired in realistic settings.

Capital Market Imperfections

While some capital market imperfections affecting firms are complicated to deal with, some imperfections relevant only for consumers can be described as elements solely within the consumer sector. For example, consider the constraint that consumers can lend but not borrow. We must then rewrite consumer utility maximization as subject to a set of budget constraints for the different time periods. In the case of two periods, for example, it would appear as

$$(31) \quad \begin{aligned} & \text{Maximize } u(x^1, x^2) \\ & \text{subject to } q^1x^1 + s \leq 0 \\ & \quad q^2x^2 - s \leq 0 \\ & \quad s \geq 0 \end{aligned}$$

where s represents first period savings. From this consumer problem, we still have utility and demand expressible in terms of

¹¹ We have benefited from discussions with Elisha Pazner on this subject.

prices. We expect that the efficiency result continues to hold. In calculating the optimal formula, though, it becomes necessary to distinguish the time period of the good in question for there are now two Lagrange multipliers giving the marginal utility of income in each of the two periods. For this consumer we have

$$(32) \quad \frac{\partial v}{\partial q_k^1} = -\alpha^1 x_k^1, \quad \frac{\partial v}{\partial q_k^2} = -\alpha^2 x_k^2$$

Since savings are allowed $\alpha^1 \geq \alpha^2$. If the consumer would borrow if he could, $\alpha^1 > \alpha^2$ and the optimal tax structure is altered by this market limitation.

REFERENCES

- W. Baumol and D. Bradford, "Optimal Departures From Marginal Cost Pricing," *Amer. Econ. Rev.*, June 1970, 69, 265-83.
- A. Bergson, "Market Socialism Revisited," *J. Polit. Econ.*, Oct. 1967, 75, 431-49.
- M. Boiteux, "Sur la gestion des monopoles public astreints à l'équilibre budgétaire," *Econometrica*, Jan. 1956, 24, 22-40.
- G. Debreu, "A Classical Tax-Subsidy Problem," *Econometrica*, Jan. 1954, 22, 14-22.
- , *Theory of Value*, New York 1959.
- J. Drèze, "Postwar Contributions of French Economists," *Amer. Econ. Rev. Supp.*, June 1964, 54, 1-64.
- A. Prest and R. Turvey, "Cost-Benefit Analysis: A Survey," *Econ. J.*, Dec. 1965, 75, 683-735.
- F. Ramsey, "A Contribution to the Theory of Taxation," *Econ. J.*, Mar. 1927, 37, 47-61.
- P. Samuelson, "Memorandum for U.S. Treasury, 1951," unpublished.

Optimal Taxation and Public Production II: Tax Rules

By PETER A. DIAMOND AND JAMES A. MIRRLEES*

In Part I of this paper which appeared in the March 1971 issue of this *Review*, we set out the problem of using taxation and government production to maximize a social welfare function. We derived the first-order conditions, and considered the argument for efficiency in aggregate production. Here in Part II we consider the structure of optimal taxes in more detail. Part I contained five sections, and Part II begins at Section VI. In the sixth and seventh sections we consider commodity taxation in one- and many-consumer economies. In the eighth section we consider other kinds of taxes; and in the ninth, public consumption. In the tenth section we consider a rigorous treatment of the problem, giving a sufficient condition for the validity of the first-order conditions. To begin, we shall restate the notation and basic problem.

Notation

p	producer prices
q	consumer prices
t	taxes ($t = q - p$)
$x^h(q)$	net demand by consumer h (incomes are assumed to equal zero) $h = 1, 2, \dots, H$
$u^h(x^h)$	utility function of consumer h
$v^h(q)$	indirect utility function of consumer h $v^h(q) = u^h(x^h(q))$
$X(q)$	aggregate net demand $X(q) = \sum_h x^h(q)$

* Massachusetts Institute of Technology and Nuffield College, respectively. The remainder of the matching footnote in Part I is appropriate here too.

$U(x^1, \dots, x^H)$	social welfare function
$V(q)$	indirect social welfare function $V(q) = U(x^1(q), \dots, x^H(q))$
$W(u^1, \dots, u^H)$	special case of an individualistic social welfare function, assumed for some of the analysis below.

With this notation before us again, we can restate the welfare maximization problem as that of selecting q to

$$(33) \quad \begin{aligned} &\text{Maximize } V(q) \\ &\text{subject to } G(X(q)) \leq 0 \end{aligned}$$

where G represents the aggregate production constraint. This problem gave rise to the first-order conditions ((19) and (22)) which were equivalently stated as

$$(34) \quad \begin{aligned} \frac{\partial V}{\partial q_k} &= \lambda \sum_i p_i \frac{\partial X_i}{\partial q_k} \\ &= -\lambda \frac{\partial}{\partial t_k} \left(\sum_i t_i X_i \right) \\ &\quad (k = 1, 2, \dots, n) \end{aligned}$$

Equations (34) were derived only for $k = 2, \dots, n$. But we can see that they hold also for $k = 1$; for, on multiplying by q_k and adding, we have

$$\sum_{k=1}^n \left[\frac{\partial V}{\partial q_k} - \lambda \sum_i p_i \frac{\partial X_i}{\partial q_k} \right] q_k = 0$$

by the homogeneity of degree 0 of V and the X_i . Equation (34) states that the impact of a price rise on social welfare is proportional to the cost of meeting the change

in demand induced by the price rise. Alternatively the impact of a tax increase on social welfare is proportional to the induced change in tax revenue (all calculated at fixed producer prices).

VI. Optimal Tax Structure— One-Consumer Economy

For one consumer and an individualistic welfare function (so that V coincides with v , the indirect utility function of the only consumer in the economy), we can express directly the derivative of social welfare with respect to q_k ($v_k = -\alpha x_k$ where α is the marginal utility of income—see equation (5) of Part I). For this case we can then explore the structure of taxation in more detail. The formulation of the first-order conditions using compensated demand derivatives is due to Paul Samuelson (1951). We begin by stating the familiar Slutsky equation:

$$(35) \quad \frac{\partial x_i}{\partial q_k} = s_{ik} - x_k \frac{\partial x_i}{\partial I}$$

where s_{ik} is the derivative of the compensated demand curve for i with respect to q_k , and $\partial x_i / \partial I$ is the derivative of the uncompensated demand with respect to income (evaluated at $I=0$ in our case). We shall make use of the well-known result that $s_{ik} = s_{ki}$.

Substituting into the first-order conditions (34) we have:

$$(36) \quad \begin{aligned} -\alpha x_k &= -\lambda \frac{\partial}{\partial t_k} \left(\sum_i t_i x_i \right) \\ &= -\lambda \left(x_k + \sum_i t_i \frac{\partial x_i}{\partial t_k} \right) \\ &= -\lambda x_k - \lambda \sum_i t_i s_{ik} \\ &\quad + \lambda x_k \sum_i t_i \frac{\partial x_i}{\partial I} \\ k &= 1, 2, \dots, n \end{aligned}$$

Rearranging terms, we can write this in the form:

$$(37) \quad \frac{\sum_i t_i s_{ik}}{x_k} = \frac{\alpha + \lambda - \lambda \sum_i t_i \frac{\partial x_i}{\partial I}}{\lambda}$$

The point to be noticed is that the right-hand side of this equation is independent of k . Call it $-\theta$. Finally, using the symmetry of the Slutsky matrix, we write the first-order conditions as:

$$(38) \quad \frac{\sum_i s_{ki} t_i}{x_k} = -\theta$$

Multiplying by $t_k x_k$ and summing, we obtain

$$(39) \quad \theta \sum_k t_k x_k = - \sum_{k,i} t_k s_{ki} t_i \geq 0,$$

by the negative semi-definiteness of the Slutsky matrix. Thus θ has the same sign as net government revenue.

The left-hand side of (38) is the percentage change in the demand for good k that would result from the tax change if producer prices were constant, the consumer were compensated so as to stay on the same indifference curve, and the derivatives of the compensated demand curves were constant at the same level as at the optimum point:

$$(40) \quad \begin{aligned} \Delta x_k &= \sum_i \int_0^{t_i} \frac{\partial x_k}{\partial t_i} dt_i = \sum_i \int_0^{t_i} s_{ki} dt_i \\ &= \sum_i s_{ki} \int_0^{t_i} dt_i = \sum_i s_{ki} t_i \end{aligned}$$

In fact, it is not possible for all these derivatives to be constant. But if the optimal taxes are small, it is approximately true that the optimal tax structure implies an equal percentage change in compensated demand at constant producer prices.

We can also calculate the actual changes in demand arising from the tax structure (assuming price derivatives of demand and production prices are constant) by resubstituting from the Slutsky equation (35). Then, upon substitution, we have:

$$\sum_i \frac{\partial x_k}{\partial q_i} t_i + \frac{\partial x_k}{\partial I} \sum_i t_i x_i = -\theta x_k;$$

or

$$(41) \quad \frac{\sum_i \frac{\partial x_k}{\partial q_i} t_i}{x_k} = -\theta - x_k^{-1} \frac{\partial x_k}{\partial I} \sum_i t_i x_i$$

The actual changes in demand (again assuming constant derivatives) induced by the tax structure differ from proportionality with a larger than average percentage fall in demand for goods with a large income derivative.

Three-Good Economy

In the case of a three-good economy, we can obtain an expression for the relative ad valorem tax rates of the two taxed goods. This argument is similar to that of W. J. Corlett and D. C. Hague, who discussed the direction of movement away from proportional taxation that would increase utility. In the three-good case, with good one untaxed, the first-order conditions (38) become

$$(42) \quad \begin{aligned} s_{22}t_2 + s_{23}t_3 &= -\theta x_2 \\ s_{32}t_2 + s_{33}t_3 &= -\theta x_3 \end{aligned}$$

Solving these equations we have

$$(43) \quad t_2 = \theta \frac{s_{23}x_3 - s_{33}x_2}{s_{22}s_{33} - s_{23}^2}, \quad t_3 = \theta \frac{s_{32}x_2 - s_{22}x_3}{s_{22}s_{33} - s_{23}^2}$$

Notice that the denominator here is positive, by the properties of the Slutsky matrix. We convert these into elasticity expressions, defining the elasticity of compensated demand by

$$(44) \quad \sigma_{ij} = \frac{q_j s_{ij}}{x_i}$$

Equation (43) can then be written

$$(45) \quad \frac{t_2}{q_2} = \theta'(\sigma_{23} - \sigma_{33}), \quad \frac{t_3}{q_3} = \theta'(\sigma_{32} - \sigma_{22}),$$

where

$$\theta' = \frac{\theta x_2 x_3}{q_2 q_3 (s_{22}s_{33} - s_{23}^2)}$$

We now substitute for σ_{23} and σ_{33} , using the adding-up properties of compensated elasticities,

$$(46) \quad \begin{aligned} \sigma_{23} &= -\sigma_{22} - \sigma_{21}, \\ \sigma_{32} &= -\sigma_{33} - \sigma_{31} \end{aligned}$$

This gives us

$$(47) \quad \begin{aligned} \frac{t_2}{q_2} &= \theta'(\sigma_{21} + \sigma_{22} + \sigma_{33}), \\ \frac{t_3}{q_3} &= \theta'(\sigma_{31} + \sigma_{32} + \sigma_{22}) \end{aligned}$$

The interesting case to consider is where labor ($x_1 < 0$) is the untaxed good, while goods 2 and 3 are consumer goods ($x_2 > 0$, $x_3 > 0$). Then θ' has the same sign as net government revenue. For definiteness, suppose that government revenue is positive so that $\theta' > 0$. Equation (47) shows that

$$(48) \quad \frac{t_2}{q_2} \stackrel{>}{<} \frac{t_3}{q_3} \text{ according as } \sigma_{21} \stackrel{<}{>} \sigma_{31}$$

The tax rate is proportionally greater for the good with the smaller cross-elasticity of compensated demand with the price of labor. (It is possible that one commodity is subsidized, but it has to be the one with the greater cross-elasticity.)

Examples

The implications of the above model are very diverse, depending upon the nature of the demand functions. A simple example will show how the theory can be used. If we define ordinary demand elasticities by the usual formula

$$(49) \quad \epsilon_{ik} = q_k x_i^{-1} \frac{\partial x_i}{\partial q_k},$$

we can rewrite the optimal taxation formula in the form

$$(50) \quad v_k = q_k^{-1} \lambda \sum p_i x_i \epsilon_{ik}$$

When the welfare function is individualistic, equation (5) applies, so that equation (50) may be written as

$$(51) \quad -\alpha q_k x_k = \lambda \sum_{i \neq k} p_i x_i \epsilon_{ik}$$

or

$$q_k p_k^{-1} = -\frac{\lambda}{\alpha} \sum_i \frac{p_i x_i}{p_k x_k} \epsilon_{ik}$$

If we have a good whose price does not affect other demands (implying a unitary own price elasticity), equation (51) simplifies to yield the optimal tax of that good:

$$(52) \quad \text{If } \epsilon_{ik} = 0 \ (i \neq k) \quad \text{and} \quad \epsilon_{kk} = -1, \\ \text{then } q_k p_k^{-1} = \lambda \alpha^{-1}$$

where $q_k p_k^{-1}$ equals one plus the percentage tax rate. Recalling that α is the marginal utility of income while λ reflects the change in welfare from allowing a government deficit financed from some outside source, their ratio gives a marginal cost (in terms of the numeraire good) of raising revenue. Thus the optimal tax rate on such a good gives the cost to society of raising the marginal dollar of tax.

An example of a utility function exhibiting such demand curves is the Cobb-Douglas, where only labor is supplied. As an example consider:

$$(53) \quad u(x) = b_1 \log(x_1 + \omega_1) + \sum_{i=2}^n b_i \log x_i$$

If we choose labor as the untaxed numeraire, all other goods satisfy (52) and we see that the optimal tax structure is a proportional tax structure.

It is easy to exhibit examples where the optimal tax structure is not proportional. Consider the example:

$$(54) \quad u(x) = \sum b_i \log(x_i + \omega_i), \\ \sum b_i = 1, \omega_i \neq 0$$

The demands arising from these preferences are:

$$(55) \quad x_i = q_i^{-1} b_i \sum_j q_j \omega_j - \omega_i$$

Therefore the demand elasticities are:

$$(56) \quad \epsilon_{ik} = b_i \omega_k x_i^{-1} \frac{q_k}{q_i} \quad (k \neq i)$$

$$\epsilon_{kk} = -b_k x_k^{-1} \sum_{j \neq k} \omega_j \frac{q_j}{q_k}$$

Substituting in the formula for the optimal taxes,

$$(57) \quad -\alpha q_k x_k = \\ \lambda \left[\sum_{j \neq k} b_j \frac{p_j}{q_j} \omega_k q_k - b_k \frac{p_k}{q_k} \sum_{j \neq k} \omega_j q_j \right] \\ = \lambda \sum_j \left[b_j \omega_k \frac{p_j q_k}{q_j} - b_k \omega_j \frac{p_k q_j}{q_k} \right]$$

Since the assumption $\sum b_j = 1$ allows us to write the demand functions (55) in the form:

$$(58) \quad q_k x_k = \sum_j [b_k \omega_j q_j - b_j \omega_k q_k],$$

we can deduce from (57) and (58) that

$$(59) \quad \sum_j \left[b_j \omega_k q_k \left(\frac{p_j}{q_j} - \frac{\alpha}{\lambda} \right) - b_k \omega_j q_j \left(\frac{p_k}{q_k} - \frac{\alpha}{\lambda} \right) \right] = 0$$

These equations allow us to calculate p for any given q , and in that way give the optimal taxation rules. In general, taxes will not be proportional. As one example of this, consider the following three-good case.

Sample Calculation

Let us combine the above two examples by considering a three-good economy (one-consumer good and two types of labor) with preferences as in (54). This example will be used to show that limited tax possibilities (represented by the same proportional tax on goods 2 and 3) intro-

duces the desirability of aggregate production inefficiency.

Example e. Assume that preferences satisfy

$$(60a) \quad u =$$

$$\log x_1 + \log (x_2 + 1) + \log (x_3 + 2)$$

$$x_1 > 0, \quad x_2 > 1, \quad x_3 > -2;$$

while private production possibilities are

$$(60b) \quad y_1 + y_2 + y_3 \leq 0,$$

$$y_1 \geq 0, \quad y_2 \leq 0, \quad y_3 \leq 0;$$

and the government constraint is

$$(60c) \quad 1.02z_1 + z_2 \leq 0$$

$$z_1 \geq 0, \quad z_2 \leq 0, \quad z_3 \leq -0.1$$

Thus the government needs good 3 for public use and can produce good 1 from good 2, but only less efficiently than the private sector can.

Since we know that production efficiency is desired, we have

$$q_1 = p_1 = p_2 = 1, \quad z_1 = z_2 = 0$$

From the first-order conditions (59) and market clearance given the demands (58), we obtain two equations to determine q_2 and q_3 :

$$q_2(q_3^{-1} - 1) = 2q_3(q_2^{-1} - 1)$$

$$(q_2 + 2q_3)(q_2^{-1} + q_3^{-1} + 1) = 8.7$$

These have a unique positive solution

$$q_2 = 0.94494, \quad q_3 = 0.90008$$

which give

$$x_1 = 0.9150, \quad x_2 = -0.0316, \quad x_3 = -0.9834$$

$$u = -0.1045$$

If we now require the same tax rate on goods 2 and 3 and at the same time impose production efficiency, then $q_2 = q_3 = q$, and the tax rate is determined by the market clearance equation. We obtain

$$3q + 6 = 8.7; \quad \text{i.e., } q = 0.9$$

Then demands are

$$x_1 = 0.9, \quad x_2 = 0, \quad x_3 = -1$$

and

$$u = -0.1054$$

Notice that the economy is still on the production frontier even though both input prices are lower in this case. If we introduce inefficiency with $p_2 > 1$, so that $y_2 = 0$ and $x_2 = z_2$, we can increase utility. Market clearance now requires

$$(q_2 + 2q_3)((1.02)^{-1}q_2^{-1} + q_3^{-1} + 1) = 8.7$$

At prices $q_2 = .92$, $q_3 = .90008$ for example, we have, $x_1 = 0.9067$, $x_2 = -0.0144$, $x_3 = -0.9926$, and $u = -0.1051$.

VII. Optimal Tax Structure—Many-Consumer Economy

As we noted in Section III of Part I, the equations for optimal taxation with a single consumer which do not reflect the particular form of V are also valid for many consumers. To pursue the analysis further, we must find an expression for V_k , the derivative of social welfare with respect to the k th consumer price.

With an individualistic welfare function, we have

$$(61) \quad V(q) = W(v^1(q), v^2(q), \dots, v^H(q))$$

Differentiating with respect to q_k , we obtain

$$(62) \quad V_k = \sum_h \frac{\partial W}{\partial u^h} v_k^h = - \sum_h \frac{\partial W}{\partial u^h} \alpha^h x_k^h$$

The term α^h is the marginal utility of income of consumer h . Therefore

$$(63) \quad \beta^h = \frac{\partial W}{\partial u^h} \alpha^h$$

is the increase in social welfare from a unit increase in the income of consumer h . We have

$$(64) \quad -V_k = \sum_h \beta^h x_k^h,$$

or the derivative of welfare with respect to a price equals the “welfare-weighted” net consumer demand for commodity k . The necessary condition for optimal taxation makes V_k proportional to the marginal contribution to tax revenue from raising the tax on good k .

$$(65) \quad \sum_h \beta^h x_k^h = \lambda \frac{\partial T}{\partial t_k},$$

where $T = \sum_i t_i X_i$ is total tax revenue, and the derivative is evaluated at constant producer prices (i.e., on the basis of consumer excess demand functions alone). We also have the alternative formula

$$(66) \quad \sum_h \beta^h x_k^h = -\lambda \sum_i p_i \frac{\partial X_i}{\partial q_k}$$

Example f. Before turning to interpretations of the optimal tax formulae like those above, let us consider an example.

We will assume that each consumer has a Cobb-Douglas utility function,

$$(67) \quad u^h = b_1^h \log(x_1^h + \omega^h) + \sum_2^n b_i^h \log x_i^h, \quad \sum_1^n b_i^h = 1$$

Choosing good 1 as numeraire, we saw in Section VI that with a one-consumer economy, taxation would be proportional. This will not, in general, be true in a many-consumer economy where each consumer has this utility function. The individual demand curves arising from this utility function are:

$$(68) \quad \begin{aligned} x_i^h &= q_i^{-1} b_i^h q_1 \omega^h, & i &= 2, 3, \dots, n \\ x_1^h &= -(1 - b_1^h) \omega^h \end{aligned}$$

Notice that $\partial x_i^h / \partial q_k = 0$ ($k \neq i \neq 1$) and $\partial x_i^h / \partial q_1 = -x_i^h / q_i$ ($i \neq 1$).

Assuming an individualistic welfare function, the first-order conditions (66) are in this case

$$(69) \quad \sum_h \beta^h x_k^h = \lambda p_k q_k^{-1} \sum_h x_k^h \quad (k = 2, \dots, n)$$

This implies the following formula:

$$(70) \quad \frac{q_k}{p_k} = \lambda \frac{\sum_h x_k^h}{\sum_h \beta^h x_k^h} = \lambda \frac{\sum_h b_k^h \omega^h}{\sum_h \beta^h b_k^h \omega^h} \quad (k = 2, \dots, n)$$

To complete the determination of the optimal taxes, we must find the relationship between λ , p_1 , and q_1 . This is obtained from the Walras identity. The value of net consumer demand in producer prices is equal to minus the profit in production. (Alternatively, we could determine λ so that the government budget is balanced.) That is

$$(71) \quad -p_1 \sum_h (1 - b_1^h) \omega^h + \sum_{i=2}^n \sum_h p_i q_i^{-1} b_i^h q_1 \omega^h = \gamma,$$

where γ is the maximized profit of production net of government needs ($= \sum_{i=1}^n p_i z_i$). Substituting from (70) and rearranging, we obtain

$$(72) \quad \begin{aligned} \frac{q_1}{p_1} &= \lambda \frac{\sum_h (1 - b_1^h) \omega^h + \gamma p_1^{-1}}{\sum_{i=2}^n \sum_h \beta^h b_i^h \omega^h} \\ &= \lambda \frac{\sum_h (1 - b_1^h) \omega^h + \gamma p_1^{-1}}{\sum_h \beta^h (1 - b_1^h) \omega^h} \end{aligned}$$

The number γp_1^{-1} is determined by the technology and the government expenditure decision, and therefore depends on p (unless $\gamma = 0$).

Equations (70) and (72) determine the optimal tax rates. If the social marginal utilities, β^h , are independent of taxation, the optimal tax rates can be read off at

once. This is true if W has the special form $\sum_h v^h$; for in that case $\beta^h = 1/\omega^h$. It should be noticed that, although each household's social marginal utility of income is unaffected by taxation, it is desirable to have taxation in general. If households with relatively low social marginal utility of income predominate among the purchasers of a commodity, that commodity should be relatively highly taxed. Although such taxation does nothing to bring social marginal utilities of income closer together, it does increase total welfare.

In general, taxation does affect social marginal utilities of income. The β^h depend on the tax rates, and equations (70) do not, therefore, give explicit formulae for the optimum taxes. In the case $W = -\mu^{-1} \sum_h e^{-\mu v^h}$, $\mu > 0$, so that there is a stronger bias toward equality than in the additive case, it can be verified quite easily that the optimum taxes have to satisfy

$$(73) \quad \frac{q_k}{p_k} \sum_h b_k^h (\omega^h)^{-\mu} \prod_{i=2}^n (b_i^h)^{-\mu b_i^h} q_i^h \\ = \lambda \sum_h b_k^h \omega^h \quad (k = 2, 3, \dots, n)$$

In this case, marginal utilities of income are brought closer together.¹ It is not immediately obvious from the equations (10) that the q are determined given the p . However, it can be shown that, in the present example, the first-order conditions must have a unique solution.² In fact, the

¹ If $\mu < 0$, utilities and marginal utilities are moved further apart.

² It is easily verified that $v^h = \delta_h + \sum_i b_i \log(q_i/q_i)$, where the δ_h are constants. Consequently

$$V(q) = -\mu^{-1} \sum_h e^{-\mu \delta_h} \prod_i (q_i/q_i)^{-\mu b_i^h}$$

which is a concave function of $(q_1/q_2, q_1/q_2, \dots, q_1/q_n)$. Also, aggregate demand is

$$X_i(q) = \sum_h b_i^h \omega^h \cdot (q_1/q_i), \quad X_1(q) = - \sum_h (1 - \alpha_i^h) \omega^h$$

If the production set is convex, the set of $(q_1/q_2, \dots, q_1/q_n)$ for which (X_1, X_2, \dots, X_n) is feasible is also convex. Thus the optimum q is obtained by maximizing a

relations (70) (along with (72)) would, if followed by government, certainly lead to maximum welfare if production were perfectly competitive, since any state of the economy satisfying these conditions maximizes welfare, and the maximum is unique for the welfare function considered. Unfortunately this convenient property is not general.

From equation (70) we can identify two cases where optimal taxation is proportional. If the social marginal utility of income is the same for everyone ($\beta^h = \beta$, for all h), then equation (70) reduces to $q_k p_k^{-1} = \lambda/\beta$. In this case there is no welfare gain to be achieved by redistributing income, and so no need to tax differently (on average) the expenditures of different individuals. Thus the optimal tax formula has the same form as in the one-consumer case. When the β^h do differ, taxes are greater on commodities purchased more heavily by individuals with a low social marginal utility of income. If, for example, the welfare function treats all individuals symmetrically and if there is diminishing social marginal utility with income, then there is greater taxation on goods purchased more heavily by the rich.

The second case leading to proportional taxation occurs when demand vectors are proportional for all individuals, $x^h = \rho^h x$, and thus $b_k^h = b_k$ for all h . With all individuals demanding goods in the same proportions, it is impossible to redistribute income by commodity taxation implying that the tax structure again assumes the form it has in a one-consumer economy.

Optimal Tax Formulae

The description in Section VI of some possible interpretations of the optimal tax formula carries over to the many-consumer case. Thus, as was true there con-

concave function of $(q_1/q_2, \dots, q_1/q_n)$ over a convex set, and is therefore uniquely defined by the first-order conditions.

sumer price elasticities but not producer price elasticities enter the equations, and at the optimum the social marginal utility of a price change is proportional to the marginal change in tax revenue from raising that tax, calculated at constant producer prices. Analysis of the change in demand can also be carried out, but is naturally more complicated. Assuming an individualistic welfare function, the first-order conditions can be written³

$$(74) \quad \sum_h \beta^h x_k^h = \lambda \sum_h \sum_i t_i \frac{\partial x_i^h}{\partial q_k} + \lambda \sum_h x_k^h$$

From the Slutsky equation, we know that

$$(75) \quad \begin{aligned} \frac{\partial x_i}{\partial q_k} &= s_{ik} - x_k \frac{\partial x_i}{\partial I} = s_{ki} - x_k \frac{\partial x_i}{\partial I} \\ &= \frac{\partial x_k}{\partial q_i} - x_k \frac{\partial x_i}{\partial I} + x_i \frac{\partial x_k}{\partial I} \end{aligned}$$

Substituting from (75) in (74) we can write the optimal tax formula as equation (76). Rearranging terms we can write equation (76) as (77). With constant producer prices, equation (77) gives the change in demand as a result of taxation for a good with constant price-derivatives of the demand function (or for small taxes). Considering two such goods, we see that the percentage decrease in demand is greater for the good the demand for which is concentrated among:

³ We neglect the possibility of a free good when the first-order condition would be an inequality.

- (1) individuals with low social marginal utility of income,
- (2) individuals with small decreases in taxes paid with a decrease in income,
- (3) individuals for whom the product of the income derivative of demand for good k and taxes paid are large.

VIII. Other Taxes

Thus far we have examined the combined use of public production and commodity taxation as control variables. It is natural to reexamine the analysis when additional tax variables are included in those controlled by the government. In particular, in the next subsection we will briefly consider income taxation; but first, let us examine a general class of taxes such that the consumer budget constraint depends on consumer prices and on tax variables. We shall replace the budget constraint $\sum q_i x_i = 0$ by the more general constraint $\phi(x, q, \zeta) = 0$, where ζ represents a shift parameter to reflect the choice among different systems of additional taxation (for example, the degree of progression in the income tax). Let us note that this formulation continues to assume that all taxes are levied on consumers and that there are no profits in the economy.

The key assumption to permit an extension of the analysis above is an independence of the two constraints on the planner. We need to assume that the choice of tax variables does not affect the production

$$(76) \quad \sum_h \beta^h x_k^h = \lambda \sum_h \sum_i t_i \frac{\partial x_k^h}{\partial q_i} + \lambda \sum_h \sum_i t_i \left(x_i^h \frac{\partial x_k^h}{\partial I} - x_k^h \frac{\partial x_i^h}{\partial I} \right) + \lambda \sum_h x_k^h$$

$$(77) \quad \frac{\sum_h \sum_i t_i \frac{\partial x_i^h}{\partial q_i}}{\sum_h x_k^h} = \frac{1}{\lambda} \frac{\sum_h \beta^h x_k^h}{\sum_h x_k^h} - 1 + \frac{\sum_h \left(\sum_i t_i \frac{\partial x_i^h}{\partial I} \right) x_k^h}{\sum_h x_k^h} - \frac{\sum_h \left(\sum_i t_i x_i^h \right) \frac{\partial x_k^h}{\partial I}}{\sum_h x_k^h}$$

possibilities, and further that the choice of a production point does not affect the set of possible demand configurations. In particular, this formulation implies that producer prices do not affect consumer budget constraints. Thus the income tax, to fit this formulation, needs to be levied on the wages that consumers receive, not on the cost of wages to the firm. Similarly it is assumed that there are no sales tax deductions from the income tax base.

We know already that in such a case, optimal production is efficient. We may therefore concentrate upon the case in which all production is controlled by the government, and the production constraint is that $x_1 = g(x_2, x_3, \dots, x_n)$. We have to choose $q_2, q_3, \dots, q_n, \xi$ to

$$(78) \quad \begin{aligned} & \text{maximize } V(q, \xi) \text{ subject to } X_1(q, \xi) \\ & = g(X_2(q, \xi), \dots, X_n(q, \xi)) \end{aligned}$$

As before we introduce a Lagrange multiplier λ . Differentiation with respect to q_k yields the familiar

$$(79) \quad V_k = \lambda \sum_i p_i \frac{\partial X_i}{\partial q_k},$$

where the producer price p_i is $\partial g / \partial x_i$ ($i = 2, 3, \dots, n$), and $p_1 = 1$. Differentiation with respect to the new tax variable provides the similar equation

$$(80) \quad \frac{\partial V}{\partial \xi} = \lambda \sum_i p_i \frac{\partial X_i}{\partial \xi}$$

We have an alternative form for (79), namely,

$$(81) \quad V_k = -\lambda \frac{\partial T}{\partial t_k}$$

In exactly the same way, we obtain from (80) a formula involving the effect of the new tax on total tax revenue,

$$(82) \quad V_\xi = -\lambda \frac{\partial T}{\partial \xi}$$

Income Taxation

Nothing that we have said suggests that commodity taxation is superior to income taxation. The analysis has only considered the best use of commodity taxation. It is natural to go on to ask how one employs both commodity taxation and income taxation. The formulation of income taxation raises a problem. If the planners are free to select any income tax structure and if there are a finite number of tax payers, the tax structure can be selected so that the marginal tax rate is zero for each taxpayer at his equilibrium income (although this does not necessarily bring the economy to the full welfare maximum). This eliminates much of our problem, but like lump sum taxation, seems to be beyond the policy tools available in a large economy. The natural formulation of this problem is for a continuum of tax payers, since then no man can have a tax schedule tailor-made for him. (This approach is taken by Mirrlees.) However, we shall here take the alternative route by assuming a limited set of alternatives for the income tax structure.

If only commodity taxation is possible, the tax paid by a household that purchases a vector x^h is

$$(83) \quad T^h = \sum_i t_i x_i^h$$

To add income taxation to the tax structure, we can select a subset of commodities, L , e.g., labor services, and tax the value of transactions on this subset, so that

$$I^h = \sum_{i \in L} q_i x_i^h$$

where I is "taxable income." Then

$$(84) \quad T^h = \sum_i t_i x_i^h + \tau(I^h, \xi),$$

where τ is a fixed continuously differentiable function depending on a parameter ξ , and is the same for all consumers. With a

tax on services (x_i negative) we would expect τ to be decreasing in its tax base, with a derivative between zero and minus one. In terms of the notation employed above, we can define the budget constraint $\phi(x^h, q, \xi)$ by

$$(85) \quad \begin{aligned} \phi(x^h, q, \xi) &= \sum p_i x_i^h + T^h \\ &= \sum q_i x_i^h + \tau \left(\sum_{i \text{ in } L} q_i x_i^h, \xi \right) \end{aligned}$$

Here we can regard q and ξ as the policy variables. Thus the consumer's budget constraint can be expressed in a form depending on consumer prices and independent of producer prices.

The first-order conditions for optimal income taxation are just the conditions (79) and (80), interpreted for this special case. The social marginal utility of a tax variable change is proportional to the marginal change in tax revenue calculated at constant producer prices. In the case of an individualistic welfare function, we can give more explicit formulae for the welfare derivatives, V_k and V_ξ :

$$(86) \quad V_k = \sum_h \beta^h x_k^h \left(1 + \delta_k \frac{\partial \tau^h}{\partial I} \right)$$

$$(87) \quad V_\xi = \sum_h \beta^h \frac{\partial \tau^h}{\partial \xi},$$

where $\delta_k = 1$ if k is in L , 0 if k is not in L ; and $\tau^h = \tau(I^h, \xi)$.

These equations are derived from the first-order conditions for maximizing u^h subject to $\phi=0$, noticing that, for example, the budget constraint implies that

$$\sum_k \frac{\partial \phi}{\partial x_k} \frac{\partial x_k}{\partial \xi} + \frac{\partial \phi}{\partial \xi} = 0$$

Combining (82) and (87), we obtain

$$(88) \quad \sum \beta^h \frac{\partial \tau^h}{\partial \xi} = \lambda \frac{\partial T}{\partial \xi}$$

Thus, at the optimum, for any two different kinds of change in the income tax structure, the social-marginal-utility weighted changes in taxation (consumer behavior held constant) are proportional to the changes in total tax revenue (both income and commodity tax revenue, calculated at fixed producer prices, with consumer behavior responding to the price change).

IX. Public Consumption

From the start, we have considered the government production decision as constrained by $G(z) \leq 0$. The presence of a fixed bundle of public consumption was therefore included in the model (and would show itself by $G(0)$ being positive). This is unsatisfactory and was assumed to keep as uncluttered as possible a naturally complicated problem. We can now consider a choice among vectors of public consumption which affect social welfare directly. (We shall assume that the government controls all production, thus ignoring public expenditures which affect private production rather than consumer utility.) Let us denote by e the vector of public consumption expenditures. (Items of public consumption which are difficult to measure can be described by the inputs into their production.) The presence of public consumption alters our problem in three ways. First, public consumption represents public production (or purchases) which are not supplied to the market. Thus market clearance becomes $X = z - e$.

Second, the presence of public consumption affects private net demand, which must now be written $X(q, e)$. Third, the level of public consumption directly affects the social welfare function (by affecting individual utility in the case of an individualistic welfare function).

We can restate the basic maximization problem as

(89) Maximize $V(q, e)$

q, e

subject to $G(X(q, e) + e) \leq 0$

The presence of e in the problem will not affect the equations obtained by differentiating a Lagrangian expression with respect to q . Thus the presence of alternative bundles of public consumption does not alter the rules for the optimal tax structure. Nor would we expect it to affect the conditions which imply production efficiency at the optimum. We can therefore replace the inequality in (89) with an equality and differentiate the Lagrangian expression with respect to e_k :

$$(90) \quad \frac{\partial V}{\partial e_k} - \lambda \left[\sum G_i \frac{\partial X_i}{\partial e_k} + G_k \right] = 0$$

Since

$$\begin{aligned} (91) \quad & \sum G_i \frac{\partial X_i}{\partial e_k} \\ &= \sum p_i \frac{\partial X_i}{\partial e_k} = \sum (q_i - t_i) \frac{\partial X_i}{\partial e_k} \\ &= \frac{\partial}{\partial e_k} (\sum q_i X_i - \sum t_i X_i) \\ &= -\frac{\partial}{\partial e_k} (\sum t_i X_i), \end{aligned}$$

we can write (90) as

$$(92) \quad \frac{\partial V}{\partial e_k} = -\lambda \frac{\partial}{\partial e_k} (\sum t_i X_i) + \lambda G_k$$

Equations (92) show how the optimal level of public consumption depends on:

- (i) the direct contribution of public consumption to welfare (measured by $\partial V/\partial e_k$);
- (ii) the effect of public consumption on tax revenue (measured by $\partial \sum t_i X_i / \partial e_k$); and
- (iii) the direct cost of public consumption (G_k).

There are three differences between this

theory and that of public goods in the presence of lump sum taxation (as developed, for example, by Samuelson (1954)). Because social marginal utilities of income are not equated, the expression $\partial V/\partial e_k$ cannot be reduced to a sum of marginal rates of substitution, but depends on the weights given to the different beneficiaries of public consumption:

$$(93) \quad \frac{\partial V}{\partial e_k} = \sum_h \frac{\partial W}{\partial u^h} \frac{\partial u^h}{\partial e_k}$$

Second, the cost associated with the raising of government revenue implies that the impact of public consumption on revenue is a relevant part of the first-order conditions. Third, for the same reason, the cost of public consumption is measured in terms of the cost to the government of raising revenue to finance the expenditures (in terms of the one-consumer equation, λ may not be equal to α , the marginal utility of income).

The first-order conditions for the provision of public goods can be expressed in another way, showing the relationships between the marginal cost and "willingness to pay." Write r_k^h for the marginal rate of substitution between public good k and income for the h th household. Then $\partial u^h / \partial e_k = \alpha^h r_k^h$, where α^h is the h th household's marginal utility of income. The social marginal utility of the h th household's income, β^h , is $(\partial W / \partial u^h) \alpha^h$. Consequently, from (93)

$$(94) \quad \frac{\partial V}{\partial e_k} = \sum_h \beta^h r_k^h$$

Then, from (92)

$$(95) \quad G_k = \sum_h \left[\frac{\beta^h}{\lambda} r_k^h + \frac{\partial}{\partial e_k} \sum_i t_i x_i^h \right]$$

Thus the marginal cost of producing the public good should be equated to a sum, over all households, of the price which the household is just willing to pay for a

marginal increment in the level of provision, weighted by the marginal "social worth" of the household's income, and adjusted for the effect of the level of provision on net tax payments by the household.⁴

In the discussion of public consumption thus far it has been assumed that there were no possible fees associated with the provision of public goods. This would be appropriate for national defense or preventive medicine, but not for goods where licenses can be required from users. The optimal level of license fees will not, in general, be zero. Indeed we may be able to associate with any good more complicated pricing mechanisms than the single fixed price considered above. In particular, there are the familiar examples of two-part tariffs (a license fee for use of a facility plus a per unit charge on the amount of use), and prices depending on quantity of sales. Formally these can be treated in a fashion similar to the income taxes considered above; the set of goods over which the tax is defined is now a consumption good rather than labor. With a two-part tariff, this would imply a tax function which was not continuous at the origin.

Presumably the introduction of more general pricing and taxing schemes gives an opportunity for increasing social welfare, just as the progressive income tax gives such an opportunity. In practice, the ignored costs of tax administration may severely limit the number of complicated pricing schemes which can increase welfare. We would expect the analysis done above to be basically unchanged by the addition of these possibilities, although a

⁴ Another case can be treated in a similar manner: that of limited government production of a good, which is also being produced privately, when government production is given away rather than being sold. Since the government production rule given above does not reduce to the first-order condition in producer prices, we would not find aggregate production efficiency for the sum of these two sources of production.

two-part tariff will cause aggregate demand to have discontinuities. In practice we would expect these discontinuities to be small relative to aggregate demand, and formally, they could be eliminated by the device of a continuum of consumers.

X. The Optimal Taxation Theorem

In the earlier discussion, we employed calculus techniques to obtain the first-order conditions for the optimal tax structure. However, the valid use of Lagrange multipliers is subject to certain restrictions, which in the present case have no very obvious economic significance. This section provides a rigorous analysis of conditions under which the tax formulae (34) are indeed necessary conditions for optimality, and in particular provides economically meaningful assumptions that ensure their validity. The reader should be warned that the discussion is highly technical.

One might hope to provide a rigorous analysis by using the well-known Kuhn-Tucker theorem for differentiable (not necessarily concave) functions. This theorem requires a certain "constraint qualification" to be satisfied. Let us apply it and see how far we get. We wish to

$$\text{Maximize } V(q)$$

$$\text{subject to } g(X(q)) \leq 0 \quad \text{and} \quad q \geq 0,$$

where g is a (vector) production constraint such that $g(X) \leq 0$ if, and only if, X is in G . Given that V , X , and g are differentiable, and that the Kuhn-Tucker constraint qualification is satisfied, we have the first-order conditions

$$(96) \quad V'(q^*) = \frac{\partial V}{\partial q} \leq p \cdot \frac{\partial X}{\partial q} = p \cdot X'(q^*),$$

where $p = \lambda \cdot g'(X(q^*))$ for a vector of Lagrange multipliers λ , and is therefore a support or tangent hyperplane to G at $X(q^*)$. Since V and X are homogeneous

of degree zero, $[V'(q^*) - p \cdot X'(q^*)] \cdot q^* = 0$: consequently $\partial V/\partial q_i = p \cdot (\partial X/\partial q_i)$ for i such that $q_i^* > 0$.

To express the first-order conditions in this form, we naturally expect to assume that V and X are continuously differentiable: to that extent, the differentiability assumptions are innocuous. The assumption that the production set can be described by a finite number of continuously differentiable inequality constraints that satisfy the constraint qualification is less satisfactory. The constraint qualification is an assumption about the functions g : one can violate it by changing the functions g without changing the actual constraint set, G . Some such assumption is required to avoid not unreasonable counter-examples, as we shall see below. But it is not at all obvious how one would check whether a particular example that failed to satisfy the constraint qualification could be put right by describing G by a better behaved set of inequalities. We should like to use a constraint qualification that depends on the properties of the set G (and X) rather than the particular functions g ; and we should like the assumption to be more amenable to economic interpretation. The theorem we prove below contains such an assumption, for the case where G is convex and has an interior.

Before stating the theorem let us consider an example in which the first-order conditions are not satisfied at the optimum.

Example g. Consider the one-consumer economy. In the case shown in Figure 10, the offer curve is tangent to the production frontier at the optimum production point. As q varies, the vector $X(q)$ traces out the offer curve. Thus, holding q_2 constant, the vector $\partial X(q)/\partial q_1$ is tangent to the offer curve at $X(q^*)$. Therefore if p is the vector of producer prices, which is tangent to the

production frontier at $X(q^*)$, $p \cdot \partial X(q^*)/\partial q_1 = 0$. The same is true for the derivatives with respect to q_2 . But there is no reason why $V'(q^*)$ should be zero: therefore the above first-order conditions may not be satisfied at the optimum.

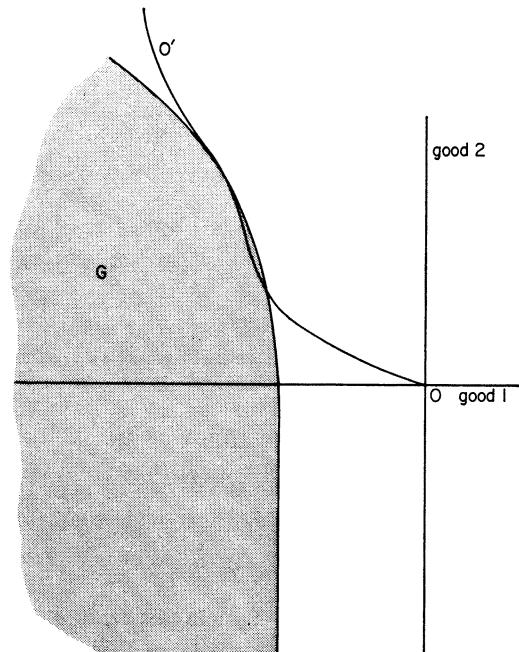


FIGURE 10

We shall make an assumption ruling out tangency between the frontier of the production set and the offer curve:

For any p, q ($q \geq 0, p \neq 0$) such that $X(q)$ is in G and $p \cdot X(q) \geq p \cdot x$ for all x in G , $p \cdot X'(q) \geq 0$.

The qualification takes this particular form because we also have the constraint $q \geq 0$. Let us note that for $q > 0$ the condition $p \cdot X(q) \geq 0$ is equivalent to $p \cdot X'(q) \neq 0$, because X is homogeneous of degree zero. The qualification asserts that *for any possible competitive equilibrium (under commodity taxation) there is a consumer price change which will decrease the value of equilibrium demand, measured in producer*

prices. By the aggregate consumer budget constraint, $q \cdot X = (p+t) \cdot X = 0$. Therefore the assumption says that at any possible equilibrium point on the production frontier, it is possible to increase tax revenue. Thus the first-order conditions may not be applicable if the optimal point represents a local tax revenue maximum. Returning to example g , we see that $p \cdot X' = 0$ at the optimum, or equivalently $\partial(t \cdot X)/\partial t = 0$, although the derivatives of V are not necessarily zero there.

We now state and prove the theorem.⁵

THEOREM 5: *Assume an optimum, (X^*, q^*) exists; that $V(q)$ and $X(q)$ are continuously differentiable; and that G is convex and has a nonempty interior. Assume furthermore that there is no pair of price vectors (p, q) for which*

$X(q)$ maximizes $p \cdot x$ for x in G ,

$$(97) \quad \begin{aligned} p &\neq 0, \text{ and} \\ p \cdot X'(q) &\geq 0 \end{aligned}$$

Then there exists p^ such that*

X^* maximizes $p^* \cdot x$ for x in G , and

$$V'(q^*) \leq p^* \cdot X'(q^*)$$

PROOF:

Let $P = \{p \mid p \cdot X^* \geq p \cdot x, \text{ all } x \text{ in } G\}$. P is the cone of normals to G at X^* , including the zero vector. It is a nonempty, closed, convex cone.

⁵ It should be noticed that when the constrained optimum is (locally) an unconstrained maximum, the producer prices satisfying the theorem are zero. This happens if optimal production is in the interior of the production set and may happen if it is on the frontier. The theorem can be weakened in a complicated manner by replacing the nontangency qualification by two conditions. One is an analog of the Kuhn-Tucker Constraint Qualification providing for the existence of an arc in the attainable set. The other use of nontangency occurs when V' is in \bar{B} but not in B . If it is assumed that when there is tangency, the cone of normals is polyhedral, B will be closed. The Kuhn-Tucker theorem is then a special case of the weakened version of theorem 5 when G is the nonnegative orthant. The Kuhn-Tucker theorem is very much easier to prove, however.

We write V' for $V'(q^*)$ and X' for $X'(q^*)$. Consider the set

$$B = \{v \mid v \leq p \cdot X', \text{ some } p \text{ in } P\}$$

We have to show that V' is in B . We do this by showing first, that if V' is in \bar{B} , the closure of B , in fact V' is in B ; and then that V' must be in \bar{B} .

If V' is in \bar{B} , there exist sequences $\{v_n\}$ and $\{p_n\}$, p_n in P , such that

$$(98) \quad \begin{aligned} v_n &\leq p_n \cdot X', \\ v_n &\rightarrow V' \quad (n \rightarrow \infty) \end{aligned}$$

Either $\{p_n\}$ is bounded or it is not. If not, we can find a subsequence on which

$$\|p_n\| \rightarrow \infty, \quad \frac{p_n}{\|p_n\|} \rightarrow \bar{p} \neq 0$$

Then, dividing (98) by $\|p_n\|$ and letting $n \rightarrow \infty$ on the subsequence, we obtain $\bar{p} \cdot X' \geq 0$ while $\bar{p} \neq 0$, is in P . This possibility is excluded by assumption (97). Therefore $\{p_n\}$ is bounded, and has a limit point p , in P . Equation (98) implies that $V' \leq p \cdot X'$. The conclusion of the theorem is thus established on the assumption that V' is in \bar{B} .

Suppose, on the contrary, that V' is not in \bar{B} . We shall derive a contradiction by a sequence of lemmas.

LEMMA 5.1:

\bar{B} is pointed. That is, v and $-v$ both belong to \bar{B} only if $v=0$.

PROOF:

If $v, -v$ is in \bar{B} , we have sequences such that

$$(99) \quad v_n^1 \leq p_n^1 \cdot X', \quad v_n^2 \leq p_n^2 \cdot X',$$

$$(100) \quad v_n^1 \rightarrow v, \quad v_n^2 \rightarrow -v$$

If $v \neq 0$, it cannot be the case that p_n^1 and p_n^2 both tend to zero. Suppose, for example, p_n^1 does not, and take a subsequence on which

$$\begin{aligned}\|p_n^1\| &\rightarrow \pi_1 \leq \infty, \\ p_n^1/\|p_n^1\| &\rightarrow p^1, \neq 0\end{aligned}$$

If $p_n^1 + p_n^2 \rightarrow 0$, $p_n^2/\|p_n^1\| \rightarrow -p^1$, and therefore $-p^1$ is in P . This is impossible, since, G having a nonempty interior, P is pointed. (If $p, -p$ are in P , $p \cdot x$ is constant for x in G , but a hyperplane has no interior.) We can therefore take a subsequence on which

$$\begin{aligned}\|p_n^1 + p_n^2\| &\rightarrow \pi, \quad 0 < \pi \leq \infty, \\ \frac{p_n^1 + p_n^2}{\|p_n^1 + p_n^2\|} &\rightarrow p, \neq 0, \in P\end{aligned}$$

From (99) (adding and dividing by $\|p_n^1 + p_n^2\|$) and (100), we now have

$$(101) \quad p \cdot X' \geq \lim \frac{v_n^1 + v_n^2}{\|p_n^1 + p_n^2\|} = 0$$

This contradicts (97), since p is in P and $p \neq 0$, and thereby establishes the lemma.

LEMMA 5.2: *If C is a pointed, closed, convex cone, there exists a vector p such that for all non-zero z in C , $p \cdot z < 0$.*

PROOF:

By the duality theorem for convex cones $C^{++}=C$, where C^+ is the dual cone, $\{p | p \cdot z \leq 0, z \text{ is in } C\}$. Clearly, if C^+ is pointed, C has a nonempty interior: for if interior C is empty, $p \cdot z=0$ for some non-zero p and all z in C , and then p and $-p$ both belong to C^+ . Under the assumptions of the theorem, C is closed and pointed. Therefore C^{++} is pointed, and C^+ has an interior point p .

$$p \cdot z < 0 \quad (\text{all nonzero } z \text{ in } C)$$

Otherwise, if $p \cdot z=0$, we can easily find a sequence $\{p_n\}$ on which $p_n \rightarrow p$ and $p_n \cdot z > 0$, so that p_n is not in C^+ .

LEMMA 5.3: *If V' is not in \bar{B} , there exists*

r such that

$$(102) \quad V' \cdot r > 0$$

$$(103) \quad v \cdot r < 0 \quad (v \in B)$$

PROOF:

The closed convex cone $\bar{B} + \{\lambda V' | \lambda \leq 0\}$ is pointed. Thus there exists an r such that

$$v \cdot r + \lambda V' \cdot r < 0$$

$$(v \in \bar{B}, \lambda \leq 0, v, \lambda \text{ not both zero})$$

Putting $v=0$ and $\lambda=-1$ we obtain (102); putting $\lambda=0$ we obtain (103).

LEMMA 5.4: *Let r be a vector satisfying (102) and (103). For some $\delta > 0$,*

$$(104) \quad X(q^* + \theta r) \in G \quad (0 \leq \theta \leq \delta)$$

PROOF:

Assume not. Then for some sequence $\{\theta_n\}$, $\theta_n > 0$, $\theta_n \rightarrow 0$,

$$X(q^* + \theta_n r) \notin G$$

Since G is convex, this implies that

$$X(q^*) + \frac{\lambda}{\theta_n} [X(q^* + \theta_n r) - X(q^*)] \notin G$$

for $\lambda \geq \theta_n$. Letting $n \rightarrow \infty$, we deduce, for any $\lambda > 0$, that

$$\begin{aligned}X(q^*) + \lambda X' \cdot r \\ = \lim_{n \rightarrow \infty} \left[X(q^*) + \lambda \frac{X(q^* + \theta_n r) - X(q^*)}{\theta_n} \right]\end{aligned}$$

is not in the interior of G . It follows that the half-line $\{X(q^*) + \lambda X' \cdot r | \lambda > 0\}$ can be separated from the interior of G by a hyperplane with normal $p \neq 0$:

$$\begin{aligned}p \cdot X(q^*) + \lambda p \cdot X' \cdot r &\geq p \cdot x \\ (\lambda > 0, x \in \text{Int } G)\end{aligned}$$

Letting $\lambda \rightarrow 0$ we have $p \in P$. Letting $x \rightarrow X^*$ we have

$$p \cdot X' \cdot r \geq 0,$$

which contradicts (103) since $p \cdot X'$ is in B . The lemma is proved.

Since q^* is optimal, (104) implies that

$$V(q^* + \theta r) \leq V(q^*) \quad (0 \leq \theta \leq \delta)$$

Therefore,

$$\begin{aligned} V' \cdot r &= \lim_{\theta \rightarrow 0} \frac{1}{\theta} [V(q^* + \theta r) - V(q^*)] \\ &\leq 0 \end{aligned}$$

This, however, contradicts (102). The hypothesis of Lemma 5.3, that $V' \in \overline{B}$, is therefore false. The proof of the theorem is thus complete.

In reaching our results that the first-order conditions for optimum taxes (96) hold in general, we have assumed that the production set, G , is convex. But one common argument for government control of production is nonconvexity of the production set. This is not a question we are primarily concerned with in this paper. However, some extensions of the theorem do hold. As an example, assume the frontier of G is differentiable at X^* , so that ρ can be uniquely defined as the normal at X^* and that G is not thin in the neighborhood of X^* —i.e., there exists a ball with center on the normal through X^* , contained in G and containing X^* . Applying the theorem to this ball we get the validity of the first-order conditions (96) using the producer prices defined by the normal.

As in general welfare economics, two uniqueness problems may arise when considering the application of the first-order conditions to achieve an optimum. In the first place, there may be more than one pair of price vectors, (ρ, q) , that satisfy the first-order conditions and allow markets to be cleared. This is similar to the problem that arises when we attempt to define optimum production and distribution by first-order conditions in the presence of a non-convex production set. It is noteworthy that, if lump sum transfers are excluded as a feasible policy, this

problem may arise even when the production set is convex. There is no reason why the demand functions should have any of the nice convexity properties which ensure that first-order conditions imply global maximization. Only in particular cases, such as that discussed in footnote 2 above (where rigorous argument is possible without appeal to theorem 5), will the first-order conditions lead to a unique solution.

The second problem is that the tax policies one might like to employ may not uniquely determine the behavior of the system. The lump sum redistribution of wealth required in standard welfare economics does not carry with it any guarantee that the desired competitive equilibrium is the unique one consistent with the optimal wealth distribution (although if the wrong equilibrium is achieved, this should be easily noticed). Similarly, in the present case, if we employ taxes rather than consumer prices as the government control variables, the equilibrium of the economy may not be unique.⁶ But if consumer prices are used as the control variables—and why not?—the demand functions give us a unique equilibrium position, so long as preferences are strictly convex.

XI. Concluding Remarks

Welfare economics has usually been concerned with characterizing the best of attainable worlds, accepting only the basic technological constraints. As economists have been aware, the omitted constraints on communication, calculation, and administration of an economy (not to mention political constraints) limit the direct applicability of the implications of this theory to policy problems, although great insight into these problems has certainly been acquired. We have not at-

⁶ For a discussion of multiple equilibria in a related problem, see E. Foster and H. Sonnenschein.

tempted to come directly to grips with the problem of incorporating these complications into economic theory. Instead, we have explored the implications of viewing these constraints as limits on the set of policy tools that can be applied. There are many sets of policy tools which might be examined in this way. Specifically, we have assumed that the policy tools available to the government include commodity taxation (and subsidization) to any extent. For these tools we have derived the rules for optimal tax policy and have shown the desirability of aggregate production efficiency, in the presence of optimal taxation. We have also considered expansion of the set of policy tools in such a way that we continue to have the condition that production decisions do not change the class of possible budget constraints. For example, this condition is still preserved when one includes poll taxes, progressive income taxation, regional differences in taxation, taxation on transactions between consumers, and most kinds of rationing. This type of expansion of the set of policy tools does not alter the desirability of production efficiency, nor does it alter the conditions for the optimal commodity tax structure, although in general the tax rates themselves will change. We have, unfortunately, ignored the cost of administering taxes. Presumably optimization by means of sets of policy tools that do not, because the cost of administration, include the full scope of commodity taxation, will not lead to the same conclusions.

Let us briefly consider the type of policy implications that are raised by our analysis. In the context of a planned economy, our analysis implies the desirability of using a single price vector in all production decisions, although these prices will, in general, differ from the prices at which commodities are sold to consumers.

As an application of this analysis to a mixed economy, let us briefly examine the

discussion of a proper criterion for public investment decisions. As has been widely noted, there are considerable differences in western economies between the intertemporal marginal rates of transformation and substitution. This has been the basis of analyses leading to investment criteria which would imply aggregate production inefficiency because they employ an interest rate for determining the margins of public production which differs from the private marginal rate of transformation. One argument used against these criteria is that the government, recognizing the divergence between rates of transformation and substitution, should use its power to achieve the full Pareto optimum, bringing these rates into equality. When this is done, the single interest rate then existing will be the appropriate rate to use in public investment decisions. We begin by presuming that the government does not have the power to achieve any Pareto optimum that it chooses. Then from the maximization of a social welfare function, we argued that the government will, in general, prefer one of the non-Pareto optima to the Pareto optima, if any, that can be achieved. At the constrained optimum, which is the social welfare function maximizing position of the economy for the available policy tools, we saw that the economy will still be characterized by a divergence between marginal rates of substitution and transformation, not just intertemporally, but also elsewhere, e.g., in the choice between leisure and goods. However, we concluded that in this situation we desired aggregate production efficiency. This implies the use of interest rates for public investment decisions which equate public and private marginal rates of transformation.

We have obtained the first-order conditions for public production, but we have not considered the correct method of evaluating indivisible investments. This

is one problem that deserves examination. In examining the optimal tax structure, we have briefly considered the tax rates implied by particular utility functions. This analysis should be extended to more general and more interesting sets of consumers. Further, we have not examined in any detail the uniqueness and stability of equilibrium, that is, the question whether there are means of achieving in practice an equilibrium which is close to the optimum.

Finally, we would like to emphasize the assumptions which seem to us most seriously to limit the applications of this theory.⁷ We have assumed no costs of tax administration and no tax evasion. And we have assumed constant-returns-to-scale and price-taking, profit-maximizing behavior in private production. Pure profits (or losses) associated with the violation of these assumptions imply that private production decisions directly influence social welfare by affecting household incomes. In such a case, it would presumably be desirable to add a profits tax to the set of policy instruments. Nevertheless, aggregate production efficiency would no longer be desirable in general; although it may be possible to get close to the opti-

⁷ These assumptions are viewed in the context of equilibrium theory. There is no need here to go into the limitations inherent in current equilibrium theory.

mum with efficient production if pure profits are small. We hope, nevertheless, that the methods and results of this paper have shown that economic analysis need not depend on the simplifying, but unrealistic, assumption that the perfect capital levy has taken place.⁸

REFERENCES

- W. J. Corlett and D. C. Hague, "Completeness and the Excess Burden of Taxation," *Rev. Econ. Stud.*, 1953, 21, No. 1, 21–30.
- E. Foster and H. Sonnenschein, "Price Distortion and Economic Welfare," *Econometrica*, Mar. 1970, 38, 281–97.
- H. Kuhn and A. Tucker, "Nonlinear Programming," in J. Neyman, ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley 1951.
- J. A. Mirrlees, "An Exploration in the Theory of Optimum Income Taxation," *Rev. Econ. Stud.*, Apr. 1971, 38, forthcoming.
- C. C. Morrison, "Marginal Cost Pricing and the Theory of Second Best," *Western Econ. J.*, June 1969, 7, 145–52.
- P. A. Samuelson, "Memorandum for U.S. Treasury, 1951," unpublished.
- , "The Pure Theory of Public Expenditure," *Rev. Econ. Statist.*, Nov. 1954, 36, 387–89.

⁸ A recent paper by Clarence Morrison also deals with marginal cost pricing as a special case of optimal pricing.

Production, Information Costs, and Economic Organization

By ARMEN A. ALCHIAN AND HAROLD DEMSETZ*

The mark of a capitalistic society is that resources are owned and allocated by such nongovernmental organizations as firms, households, and markets. Resource owners increase productivity through cooperative specialization and this leads to the demand for economic organizations which facilitate cooperation. When a lumber mill employs a cabinetmaker, cooperation between specialists is achieved within a firm, and when a cabinetmaker purchases wood from a lumberman, the cooperation takes place across markets (or between firms). Two important problems face a theory of economic organization—to explain the conditions that determine whether the gains from specialization and cooperative production can better be obtained within an organization like the firm, or across markets, and to explain the structure of the organization.

It is common to see the firm characterized by the power to settle issues by fiat, by authority, or by disciplinary action superior to that available in the conventional market. This is delusion. The firm does not own all its inputs. It has no power of fiat, no authority, no disciplinary action any different in the slightest degree from ordinary market contracting between any two people. I can "punish" you only by withholding future business or by seeking redress in the courts for any failure to honor our exchange agreement. That is exactly all that any employer can do. He

can fire or sue, just as I can fire my grocer by stopping purchases from him or sue him for delivering faulty products. What then is the content of the presumed power to manage and assign workers to various tasks? Exactly the same as one little consumer's power to manage and assign his grocer to various tasks. The single consumer can assign his grocer to the task of obtaining whatever the customer can induce the grocer to provide at a price acceptable to both parties. That is precisely all that an employer can do to an employee. To speak of managing, directing, or assigning workers to various tasks is a deceptive way of noting that the employer continually is involved in renegotiation of contracts on terms that must be acceptable to both parties. Telling an employee to type this letter rather than to file that document is like my telling a grocer to sell me this brand of tuna rather than that brand of bread. I have no contract to continue to purchase from the grocer and neither the employer nor the employee is bound by any contractual obligations to continue their relationship. Long-term contracts between employer and employee are not the essence of the organization we call a firm. My grocer can count on my returning day after day and purchasing his services and goods even with the prices not always marked on the goods—because I know what they are—and he adapts his activity to conform to my directions to him as to what I want each day . . . he is not my employee.

Wherein then is the relationship between a grocer and his employee different from that between a grocer and his cus-

* Professors of economics at the University of California, Los Angeles. Acknowledgment is made for financial aid from the E. Lilly Endowment, Inc. grant to UCLA for research in the behavioral effects of property rights.

tomers? It is in a *team* use of inputs and a centralized position of some party in the contractual arrangements of *all* other inputs. It is the *centralized contractual agent in a team productive process*—not some superior authoritarian directive or disciplinary power. Exactly what is a team process and why does it induce the contractual form, called the firm? These problems motivate the inquiry of this paper.

I. The Metering Problem

The economic organization through which input owners cooperate will make better use of their comparative advantages to the extent that it facilitates the payment of rewards in accord with productivity. If rewards were random, and without regard to productive effort, no incentive to productive effort would be provided by the organization; and if rewards were negatively correlated with productivity the organization would be subject to sabotage. Two key demands are placed on an economic organization—metering input productivity and metering rewards.¹

Metering problems sometimes can be resolved well through the exchange of products across competitive markets, because in many situations markets yield a high correlation between rewards and productivity. If a farmer increases his output of wheat by 10 percent at the prevailing market price, his receipts also increase by 10 percent. This method of organizing economic activity meters the *output directly*, reveals the marginal product and apportions the *rewards* to resource owners in accord with that direct measurement of their outputs. The success of this decentralized, market exchange in promoting productive specialization requires that changes in market rewards fall

on those responsible for changes in *output*.²

The classic relationship in economics that runs from marginal productivity to the distribution of income implicitly *assumes* the existence of an organization, be it the market or the firm, that allocates rewards to resources in accord with their productivity. The problem of economic organization, the economical means of metering productivity and rewards, is not confronted directly in the classical analysis of production and distribution. Instead, that analysis tends to assume sufficiently economic—or zero cost—means, as if productivity automatically created its reward. We conjecture the direction of causation is the reverse—the specific sys-

² A producer's wealth would be reduced by the present capitalized value of the future income lost by loss of reputation. Reputation, i.e., credibility, is an asset, which is another way of saying that reliable information about expected performance is both a costly and a valuable good. For acts of God that interfere with contract performance, both parties have incentives to reach a settlement akin to that which would have been reached if such events had been covered by specific contingency clauses. The reason, again, is that a reputation for "honest" dealings—i.e., for actions similar to those that would probably have been reached had the contract provided this contingency—is wealth.

Almost every contract is open-ended in that many contingencies are uncovered. For example, if a fire delays production of a promised product by *A* to *B*, and if *B* contends that *A* has not fulfilled the contract, how is the dispute settled and what recompense, if any, does *A* grant to *B*? A person uninitiated in such questions may be surprised by the extent to which contracts permit either party to escape performance or to nullify the contract. In fact, it is hard to imagine any contract, which, when taken solely in terms of its stipulations, could not be evaded by one of the parties. Yet that is the ruling, viable type of contract. Why? Undoubtedly the best discussion that we have seen on this question is by Stewart Macaulay.

There are means not only of detecting or preventing cheating, but also for deciding how to allocate the losses or gains of unpredictable events or quality of items exchanged. Sales contracts contain warranties, guarantees, collateral, return privileges and penalty clauses for specific nonperformance. These are means of assignment of *risks* of losses of cheating. A lower price without warranty—an "as is" purchase—places more of the risk on the buyer while the seller buys insurance against losses of his "cheating." On the other hand, a warranty or return privilege or service contract places more risk on the seller with insurance being bought by the buyer.

¹ Meter means to measure and also to apportion. One can meter (measure) output and one can also meter (control) the output. We use the word to denote both; the context should indicate which.

tem of rewarding which is relied upon stimulates a particular productivity response. If the economic organization meters poorly, with rewards and productivity only loosely correlated, then productivity will be smaller; but if the economic organization meters well productivity will be greater. What makes metering difficult and hence induces means of economizing on metering costs?

II. Team Production

Two men jointly lift heavy cargo into trucks. Solely by observing the total weight loaded per day, it is impossible to determine each person's marginal productivity. With team production it is difficult, solely by observing total output, to either define or determine *each* individual's contribution to this output of the cooperating inputs. The output is yielded by a team, by definition, and it is not a *sum* of separable outputs of each of its members. Team production of Z involves at least two inputs, X_i and X_j , with $\partial^2 Z / \partial X_i \partial X_j \neq 0$.³ The production function is *not* separable into two functions each involving only inputs X_i or only inputs X_j . Consequently there is no *sum* of Z of two separable functions to treat as the Z of the team production function. (An example of a *separable* case is $Z = aX_i^2 + bX_j^2$ which is separable into $Z_i = aX_i^2$ and $Z_j = bX_j^2$, and $Z = Z_i + Z_j$. This is not team production.) There exist production techniques in which the Z obtained is greater than if X_i and X_j had produced separable Z . Team production will be used if it yields an output enough larger than the sum of separable production of Z to cover the costs of organizing and disciplining team members—the topics of this paper.⁴

³ The function is separable into additive functions if the cross partial derivative is zero, i.e., if $\partial^2 Z / \partial X_i \partial X_j = 0$.

⁴ With sufficient generality of notation and conception this team production function could be formulated as a case of the generalized production function interpretation given by our colleague, E. A. Thompson.

Usual explanations of the gains from cooperative behavior rely on exchange and production in accord with the comparative advantage specialization principle with separable additive production. However, as suggested above there is a source of gain from cooperative activity involving working as a *team*, wherein individual cooperating inputs do not yield identifiable, separate products which can be *summed* to measure the total output. For this cooperative productive activity, here called "team" production, measuring *marginal* productivity and making payments in accord therewith is more expensive by an order of magnitude than for separable production functions.

Team production, to repeat, is production in which 1) several types of resources are used and 2) the product is not a sum of separable outputs of each cooperating resource. An additional factor creates a team organization problem—3) not all resources used in team production belong to one person.

We do not inquire into why all the jointly used resources are not owned by one person, but instead into the types of organization, contracts, and informational and payment procedures used among owners of teamed inputs. With respect to the one-owner case, perhaps it is sufficient merely to note that (a) slavery is prohibited, (b) one might assume risk aversion as a reason for one person's not borrowing enough to purchase all the assets or sources of services rather than renting them, and (c) the purchase-resale spread may be so large that costs of short-term ownership exceed rental costs. Our problem is viewed basically as one of organization among different people, not of the physical goods or services, however much there must be selection and choice of combination of the latter.

How can the members of a team be rewarded and induced to work efficiently?

In team production, marginal products of cooperative team members are not so directly and separably (i.e., cheaply) observable. What a team offers to the market can be taken as the marginal product of the team but not of the team members. The costs of metering or ascertaining the marginal products of the team's members is what calls forth new organizations and procedures. Clues to each input's productivity can be secured by observing *behavior* of individual inputs. When lifting cargo into the truck, how rapidly does a man move to the next piece to be loaded, how many cigarette breaks does he take, does the item being lifted tilt downward toward his side?

If detecting such behavior were costless, neither party would have an incentive to shirk, because neither could impose the cost of his shirking on the other (if their cooperation was agreed to voluntarily). But since costs must be incurred to monitor each other, each input owner will have more incentive to shirk when he works as part of a team, than if his performance could be monitored easily or if he did not work as a team. If there is a net increase in productivity available by team production, net of the metering cost associated with disciplining the team, then team production will be relied upon rather than a multitude of bilateral exchange of separable individual outputs.

Both leisure and higher income enter a person's utility function.⁵ Hence, each person should adjust his work and realized reward so as to equate the marginal rate of substitution between leisure and production of real output to his marginal rate of substitution in consumption. That is, he would adjust his rate of work to bring his demand prices of leisure and output to equality with their true costs. However,

with detection, policing, monitoring, measuring or metering costs, each person will be induced to take more leisure, because the effect of relaxing on *his realized* (reward) rate of substitution between output and leisure will be less than the effect on the *true* rate of substitution. His realized cost of leisure will fall more than the true cost of leisure, so he "buys" more leisure (i.e., more nonpecuniary reward).

If his relaxation cannot be detected perfectly at zero cost, part of its effects will be borne by others in the team, thus making *his* realized cost of relaxation less than the true total cost to the team. The difficulty of detecting such actions permits the private costs of his actions to be less than their full costs. Since each person responds to his private realizable rate of substitution (in production) rather than the true total (i.e., social) rate, and so long as there are costs for other people to detect his shift toward relaxation, it will not pay (them) to force him to readjust completely by making him realize the true cost. Only enough efforts will be made to equate the marginal gains of detection activity with the marginal costs of detection; and that implies a lower rate of productive effort and more shirking than in a costless monitoring, or measuring, world.

In a university, the faculty use office telephones, paper, and mail for personal uses beyond strict university productivity. The university administrators could stop such practices by identifying the responsible person in each case, but they can do so only at higher costs than administrators are willing to incur. The extra costs of identifying each party (rather than merely identifying the presence of such activity) would exceed the savings from diminished faculty "turpitudinal peccadilloes." So the faculty is allowed some degree of "privileges, perquisites, or fringe benefits." And the total of the pecuniary wages paid

⁵ More precisely: "if anything other than pecuniary income enters his utility function." Leisure stands for all nonpecuniary income for simplicity of exposition.

is lower because of this irreducible (at acceptable costs) degree of amenity-seizing activity. Pay is lower in pecuniary terms and higher in leisure, conveniences, and ease of work. But still every person would prefer to see detection made more effective (if it were somehow possible to monitor costlessly) so that he, as part of the now more effectively producing team, could thereby realize a higher pecuniary pay and less leisure. If everyone could, at zero cost, have his reward-realized rate brought to the true production possibility real rate, all could achieve a more preferred position. But detection of the responsible parties is costly; that cost acts like a tax on work rewards.⁶ Viable shirking is the result.

What forms of organizing team production will lower the cost of detecting "performance" (i.e., marginal productivity) and bring personally realized rates of substitution closer to true rates of substitution? Market competition, in principle, could monitor some team production. (It already *organizes* teams.) Input owners who are not team members can offer, in return for a smaller share of the team's rewards, to replace excessively (i.e., overpaid) shirking members. Market competition among potential team members would determine team membership and individual rewards. There would be no team leader, manager, organizer, owner, or employer. For such decentralized organizational control to work, outsiders, possibly after observing each team's total

output, can speculate about their capabilities as team members and, by a market competitive process, revised teams with greater productive ability will be formed and sustained. Incumbent members will be constrained by threats of replacement by outsiders offering services for lower reward shares or offering greater rewards to the other members of the team. Any team member who shirked in the expectation that the reduced output effect would not be attributed to him will be displaced if his activity is detected. Teams of productive inputs, like business units, would evolve in apparent spontaneity in the market—without any central organizing agent, team manager, or boss.

But completely effective control cannot be expected from individualized market competition for two reasons. First, for this competition to be completely effective, new challengers for team membership must know where, and to what extent, shirking is a serious problem, i.e., know they can increase net output as compared with the inputs they replace. To the extent that this is true it is probably possible for existing fellow team members to recognize the shirking. But, by definition, the detection of shirking by observing team output is costly for team production. Secondly, assume the presence of detection costs, and assume that in order to secure a place on the team a new input owner must accept a smaller share of rewards (or a promise to produce more). Then his incentive to shirk would still be at least as great as the incentives of the inputs replaced, because he still bears less than the entire reduction in team output for which he is responsible.

⁶ Do not assume that the sole result of the cost of detecting shirking is one form of payment (more leisure and less take home money). With several members of the team, each has an incentive to cheat against each other by engaging in more than the average amount of such leisure if the employer can not tell at zero cost which employee is taking more than average. As a result the total productivity of the team is lowered. Shirking detection costs thus change the form of payment and also result in lower total rewards. Because the cross partial derivatives are positive, shirking reduces other people's marginal products.

III. The Classical Firm

One method of reducing shirking is for someone to specialize as a monitor to check the input performance of team members.⁷

⁷ What is meant by performance? Input energy, initiative, work attitude, perspiration, rate of exhaustion?

(Continued)

But who will monitor the monitor? One constraint on the monitor is the aforesaid market competition offered by other monitors, but for reasons already given, that is not perfectly effective. Another constraint can be imposed on the monitor: give him title to the net earnings of the team, net of payments to other inputs. If owners of cooperating inputs agree with the monitor that he is to receive any residual product above prescribed amounts (hopefully, the marginal value products of the other inputs), the monitor will have an added incentive not to shirk as a monitor. Specialization in monitoring plus reliance on a residual claimant status will reduce shirking; but additional links are needed to forge the firm of classical economic theory. How will the residual claimant monitor the other inputs?

We use the term monitor to connote several activities in addition to its disciplinary connotation. It connotes measuring output performance, apportioning rewards, observing the input behavior of inputs as means of detecting or estimating their marginal productivity and giving assignments or instructions in what to do and how to do it. (It also includes, as we shall show later, authority to terminate or revise contracts.) Perhaps the contrast between a football coach and team captain is helpful. The coach selects strategies and tactics and sends in instructions about what plays to utilize. The captain is essentially an observer and reporter of

Or output? It is the latter that is sought—the *effect* or output. But performance is nicely ambiguous because it suggests both input and output. It is *nicely* ambiguous because as we shall see, sometimes by inspecting a team member's input activity we can better judge his output effect, perhaps not with complete accuracy but better than by watching the output of the *team*. It is not always the case that watching input activity is the only or best means of detecting, measuring or monitoring output effects of each team member, but in some cases it is a useful way. For the moment the word performance glosses over these aspects and facilitates concentration on other issues.

the performance at close hand of the members. The latter is an inspector-steward and the former a supervisor manager. For the present all these activities are included in the rubric "monitoring." All these tasks are, in principle, negotiable across markets, but we are presuming that such market measurement of marginal productivities and job reassessments are not so cheaply performed for team production. And in particular our analysis suggests that it is not so much the costs of spontaneously negotiating contracts in the markets among groups for team production as it is the detection of the performance of individual members of the team that calls for the organization noted here.

The specialist *who receives the residual rewards* will be the monitor of the members of the team (i.e., will manage the use of cooperative inputs). The monitor earns his residual through the reduction in shirking that he brings about, not only by the prices that he agrees to pay the owners of the inputs, but also by observing and directing the actions or uses of these inputs. *Managing or examining the ways to which inputs are used in team production is a method of metering the marginal productivity of individual inputs to the team's output.*

To discipline team members and reduce shirking, the residual claimant must have power to revise the contract terms and incentives of *individual* members without having to terminate or alter every other input's contract. Hence, team members who seek to increase their productivity will assign to the monitor not only the residual claimant right but also the right to alter individual membership and performance on the team. Each team member, of course, can terminate his own membership (i.e., quit the team), but only the monitor may unilaterally terminate the membership of any of the

other members without necessarily terminating the team itself or his association with the team; and he alone can expand or reduce membership, alter the mix of membership, or sell the right to be the residual claimant-monitor of the team. It is this entire bundle of rights: 1) to be a residual claimant; 2) to observe input behavior; 3) to be the central party common to all contracts with inputs; 4) to alter the membership of the team; and 5) to sell these rights, that defines the *ownership* (or the employer) of the *classical* (capitalist, free-enterprise) firm. The coalescing of these rights has arisen, our analysis asserts, because it resolves the shirking-information problem of team production better than does the noncentralized contractual arrangement.

The relationship of each team member to the *owner* of the firm (i.e., the party common to all input contracts *and* the residual claimant) is simply a "quid pro quo" contract. Each makes a purchase and sale. The employee "orders" the owner of the team to pay him money in the same sense that the employer directs the team member to perform certain acts. The employee can terminate the contract as readily as can the employer, and long-term contracts, therefore, are not an essential attribute of the firm. Nor are "authoritarian," "dictational," or "fiat" attributes relevant to the conception of the firm or its efficiency.

In summary, two necessary conditions exist for the emergence of the firm on the prior assumption that more than pecuniary wealth enter utility functions: 1) It is possible to increase productivity through team-oriented production, a production technique for which it is costly to directly measure the marginal outputs of the co-operating inputs. This makes it more difficult to restrict shirking through simple market exchange between cooperating inputs. 2) It is economical to estimate mar-

ginal productivity by observing or specifying input behavior. The simultaneous occurrence of both these preconditions leads to the contractual organization of inputs, known as the *classical capitalist firms* with (a) joint input production, (b) several input owners, (c) one party who is common to all the contracts of the joint inputs, (d) who has rights to renegotiate any input's contract independently of contracts with other input owners, (e) who holds the residual claim, and (f) who has the right to sell his central contractual residual status.⁸

Other Theories of the Firm

At this juncture, as an aside, we briefly place this theory of the firm in the contexts of those offered by Ronald Coase and Frank Knight.⁹ Our view of the firm is not necessarily inconsistent with Coase's; we attempt to go further and identify refutable implications. Coase's penetrating insight is to make more of the fact that markets do not operate costlessly, and he relies on the cost of using markets to *form* contracts as his basic explanation for the existence of firms. We do not disagree with the proposition that, *ceteris paribus*, the higher is the cost of transacting across markets the greater will be the comparative advantage of organizing resources within the firm; it is a difficult proposition to disagree with or to refute. We could with equal ease subscribe to a theory of the firm based on the cost of managing, for surely it is true that, *ceteris paribus*, the lower is the cost of managing the greater will be the comparative advantage of organizing resources within the firm. To move the theory forward, it is necessary to know what is meant by a firm and to

⁸ Removal of (b) converts a capitalist proprietary firm to a socialist firm.

⁹ Recognition must also be made to the seminal inquiries by Morris Silver and Richard Auster, and by H. B. Malmgren.

explain the circumstances under which the cost of "managing" resources is low relative to the cost of allocating resources through market transaction. The conception of and rationale for the classical firm that we propose takes a step down the path pointed out by Coase toward that goal. Consideration of team production, team organization, difficulty in metering outputs, and the problem of shirking are important to our explanation but, so far as we can ascertain, not in Coase's. Coase's analysis insofar as it had heretofore been developed would suggest open-ended contracts but does not appear to imply anything more—neither the residual claimant status nor the distinction between employee and subcontractor status (nor any of the implications indicated below). And it is not true that employees are generally employed on the basis of long-term contractual arrangements any more than on a series of short-term or indefinite length contracts.

The importance of our proposed additional elements is revealed, for example, by the explanation of why the person to whom the control monitor is responsible receives the residual, and also by our later discussion of the implications about the corporation, partnerships, and profit sharing. These alternative forms for organization of the firm are difficult to resolve on the basis of market transaction costs only. Our exposition also suggests a definition of the classical firm—something crucial that was heretofore absent.

In addition, sometimes a technological development will lower the cost of market transactions while, at the same time, it expands the role of the firm. When the "putting out" system was used for weaving, inputs were organized largely through market negotiations. With the development of efficient central sources of power, it became economical to perform weaving in proximity to the power source and to engage in team production. The bringing

in of weavers surely must have resulted in a reduction in the cost of negotiating (forming) contracts. Yet, what we observe is the beginning of the factory system in which inputs are organized within a firm. Why? The weavers did not simply move to a common source of power that they could tap like an electric line, purchasing power while they used their own equipment. Now team production in the joint use of equipment became more important. The measurement of marginal productivity, which now involved interactions between workers, especially through their joint use of machines, became more difficult though contract negotiating cost was reduced, while managing the *behavior* of inputs became easier because of the increased centralization of activity. The firm as an organization expanded even though the cost of transactions was reduced by the advent of centralized power. The same could be said for modern assembly lines. Hence the emergence of central power sources expanded the scope of productive activity in which the firm enjoyed a comparative advantage as an organizational form.

Some economists, following Knight, have identified the bearing of risks of wealth changes with the director or central employer without explaining why that is a viable arrangement. Presumably, the more risk-averse inputs become employees rather than owners of the classical firm. Risk averseness and uncertainty *with regard to the firm's fortunes* have little, if anything, to do with our explanation although it helps to explain why all resources in a team are not owned by one person. That is, the role of risk taken in the sense of absorbing the windfalls that buffet the firm because of unforeseen competition, technological change, or fluctuations in demand are not central to our theory, although it is true that imperfect knowledge and, therefore, risk, in *this* sense of risk, underlie the problem of

monitoring team behavior. We deduce the system of paying the manager with a residual claim (the equity) from the desire to have efficient means to reduce shirking so as to make team production economical and not from the smaller aversion to the risks of enterprise in a dynamic economy. We conjecture that "distribution-of-risk" is not a valid rationale for the *existence* and organization of the *classical* firm.

Although we have emphasized team production as creating a costly metering task and have treated team production as an essential (necessary?) condition for the firm, would not other obstacles to cheap metering also call forth the same kind of contractual arrangement here denoted as a firm? For example, suppose a farmer produces wheat in an easily ascertained quantity but with subtle and difficult to detect quality variations determined by how the farmer grew the wheat. A vertical integration could allow a purchaser to control the farmer's behavior in order to more economically estimate productivity. But this is not a case of joint or team production, unless "information" can be considered part of the product. (While a good case could be made for that broader conception of production, we shall ignore it here.) Instead of forming a firm, a buyer can contract to have his inspector on the site of production, just as home builders contract with architects to supervise building contracts; that arrangement is not a firm. Still, a firm might be organized in the production of many products wherein no team production or jointness of use of separately owned resources is involved.

This possibility rather clearly indicates a broader, or complementary, approach to that which we have chosen. 1) As we do in this paper, it can be argued that the firm is the particular policing device utilized when joint team production is present. If other sources of high policing costs arise, as in the wheat case just indicated, some other form of contractual ar-

rangement will be used. Thus to each source of informational cost there may be a different type of policing and contractual arrangement. 2) On the other hand, one can say that where policing is difficult across markets, various forms of contractual arrangements are devised, but there is no reason for that known as the firm to be uniquely related or even highly correlated with team production, as defined here. It might be used equally probably and viably for other sources of high policing cost. We have not intensively analyzed other sources, and we can only note that our current and readily revisable conjecture is that 1) is valid, and has motivated us in our current endeavor. In any event, the test of the theory advanced here is to see whether the conditions we have identified are necessary for firms to have long-run viability rather than merely births with high infant mortality. Conglomerate firms or collections of separate production agencies into one owning organization can be interpreted as an investment trust or investment diversification device—probably along the lines that motivated Knight's interpretation. A holding company can be called a firm, because of the common association of the word firm with any ownership unit that owns income sources. The term firm as commonly used is so turgid of meaning that we can not hope to explain every entity to which the name is attached in common or even technical literature. Instead, we seek to identify and explain a particular contractual arrangement induced by the cost of information factors analyzed in this paper.

IV. Types of Firms

A. Profit-Sharing Firms

Explicit in our explanation of the capitalist firm is the assumption that the cost of managing the team's inputs by a central monitor, who disciplines himself because he is a residual claimant, is low

relative to the cost of metering the marginal outputs of team members.

If we look within a firm to see who monitors—hires, fires, changes, promotes, and renegotiates—we should find him being a residual claimant or, at least, one whose pay or reward is more than any others correlated with fluctuations in the residual value of the firm. They more likely will have options or rights or bonuses than will inputs with other tasks.

An implicit “auxiliary” assumption of our explanation of the firm is that the cost of team production is increased if the residual claim is not held entirely by the central monitor. That is, we assume that if profit sharing had to be relied upon for all team members, losses from the resulting increase in central monitor shirking would exceed the output gains from the increased incentives of other team members not to shirk. If the optimal team size is only two owners of inputs, then an equal division of profits and losses between them will leave each with stronger incentives to reduce shirking than if the optimal team size is large, for in the latter case only a smaller percentage of the losses occasioned by the shirker will be borne by him. Incentives to shirk are positively related to the optimal size of the team under an equal profit-sharing scheme.¹⁰

The preceding does not imply that profit sharing is never viable. Profit sharing to encourage self-policing is more appropriate for small teams. And, indeed, where input owners are free to make whatever contractual arrangements suit them, as generally is true in capitalist economies, profit sharing seems largely limited to partner-

¹⁰ While the degree to which residual claims are centralized will affect the size of the team, this will be only one of many factors that determine team size, so as an approximation, we can treat team size as exogenously determined. Under certain assumptions about the shape of the “typical” utility function, the incentive to avoid shirking with unequal profit-sharing can be measured by the Herfindahl index.

ships with a relatively small number of *active*¹¹ partners. Another advantage of such arrangements for smaller teams is that it permits more effective reciprocal monitoring among inputs. Monitoring need not be entirely specialized.

Profit sharing is more viable if small team size is associated with situations where the cost of specialized management of inputs is large relative to the increased productivity potential in team effort. We conjecture that the cost of managing team inputs increases if the productivity of a team member is difficult to correlate with his behavior. In “artistic” or “professional” work, watching a man’s activities is not a good clue to what he is actually thinking or doing with his mind. While it is relatively easy to manage or direct the loading of trucks by a team of dock workers where input activity is so highly related in an obvious way to output, it is more difficult to manage and direct a lawyer in the preparation and presentation of a case. Dock workers can be directed in detail without the monitor himself loading the truck, and assembly line workers can be monitored by varying the speed of the assembly line, but detailed direction in the preparation of a law case would require in much greater degree that the monitor prepare the case himself. As a result, artistic or professional inputs, such as lawyers, advertising specialists, and doctors, will be given relatively freer reign with regard to individual behavior. If the management of inputs is relatively costly, or ineffective, as it would seem to be in these cases, but, nonetheless if team effort is more productive than separable production with exchange across markets, then there will develop a tendency to use profit-sharing schemes to provide incentives to avoid shirking.¹²

¹¹ The use of the word active will be clarified in our discussion of the corporation, which follows below.

¹² Some sharing contracts, like crop sharing, or rental

B. Socialist Firms

We have analyzed the classical proprietorship and the profit-sharing firms in the context of free association and choice of economic organization. Such organizations need not be the most viable when political constraints limit the forms of organization that can be chosen. It is one thing to have profit sharing when professional or artistic talents are used by small teams. But if political or tax or subsidy considerations induce profit-sharing techniques when these are not otherwise economically justified, then additional management techniques will be developed to help reduce the degree of shirking.

For example, most, if not all, firms in Jugoslavia are owned by the employees in the restricted sense that all share in the residual. This is true for large firms and for firms which employ nonartistic, or nonprofessional, workers as well. With a decay of political constraints, most of these firms could be expected to rely on paid wages rather than shares in the residual. This rests on our auxiliary assumption that general sharing in the residual results in losses from enhanced shirking by the monitor that exceed the gains from reduced shirking by residual-sharing employees. If this were not so, profit sharing with employees should have occurred more frequently in Western societies where such organizations are neither banned nor preferred politically. Where residual sharing by employees is politically imposed, as in Jugoslavia, we are led to expect that some management technique will arise to reduce the shirking by the central monitor, a technique that will not be found frequently in Western societies since the monitor retains all (or much) of the re-

sidual in the West and profit sharing is largely confined to small, professional-artistic team production situations. We do find in the larger scale residual-sharing firms in Jugoslavia that there are employee committees that can recommend (to the state) the termination of a manager's contract (veto his continuance) with the enterprise. We conjecture that the workers' committee is given the right to recommend the termination of the manager's contract precisely because the general sharing of the residual increases "excessively" the manager's incentive to shirk.¹³

C. The Corporation

All firms must initially acquire command over some resources. The corporation does so primarily by selling promises of future returns to those who (as creditors or owners) provide financial capital. In some situations resources can be acquired in advance from consumers by promises of future delivery (for example, advance sale of a proposed book). Or where the firm is a few artistic or professional persons, each can "chip in" with time and talent until the sale of services brings in revenues. For the most part, capital can be acquired more cheaply if many (risk-averse) investors contribute small portions to a large investment. The economies of raising large sums of equity capital in this way suggest that modifications in the relationship among corporate inputs are required to cope with the shirking problem

¹³ Incidentally, investment activity will be changed. The inability to capitalize the investment value as "take-home" proviate property *wealth* of the members of the firm means that the benefits of the investment must be taken as annual income by those who are employed at the time of the income. Investment will be confined more to those with shorter life and with higher rates or pay-offs if the alternative of investing is paying out the firm's income to its employees to take home and use as private property. For a development of this proposition, see the papers by Eirik Furoboth and Svetozar Pejovich, and by Pejovich.

payments based on gross sales in retail stores, come close to profit sharing. However, it is gross output sharing rather than profit sharing. We are unable to specify the implications of the difference. We refer the reader to S. N. Cheung.

that arises with profit sharing among large numbers of corporate stockholders. One modification is limited liability, especially for firms that are large relative to a stockholder's wealth. It serves to protect stockholders from large losses no matter how they are caused.

If every stock owner participated in each decision in a corporation, not only would large bureaucratic costs be incurred, but many would shirk the task of becoming well informed on the issue to be decided, since the losses associated with unexpectedly bad decisions will be borne in large part by the many other corporate shareholders. More effective control of corporate activity is achieved for most purposes by transferring decision authority to a smaller group, whose main function is to negotiate with and manage (renegotiate with) the other inputs of the team. The corporate stockholders retain the authority to revise the membership of the management group and over major decisions that affect the structure of the corporation or its dissolution.

As a result a new modification of partnerships is induced—the right to sale of corporate shares without approval of any other stockholders. Any shareholder can remove his wealth from control by those with whom he has differences of opinion. Rather than try to control the decisions of the management, which is harder to do with many stockholders than with only a few, unrestricted salability provides a more acceptable escape to each stockholder from continued policies with which he disagrees.

Indeed, the policing of managerial shirking relies on across-market competition from new groups of would-be managers as well as competition from members within the firm who seek to displace existing management. In addition to competition from outside and inside managers, control is facilitated by the temporary

congealing of share votes into voting blocs owned by one or a few contenders. Proxy battles or stock-purchases concentrate the votes required to displace the existing management or modify managerial policies. But it is more than a change in policy that is sought by the newly formed financial interests, whether of new stockholders or not. It is the capitalization of expected future benefits into stock prices that concentrates on the innovators the wealth gains of their actions if they own large numbers of shares. Without capitalization of future benefits, there would be less incentive to incur the costs required to exert informed decisive influence on the corporation's policies and managing personnel. Temporarily, the structure of ownership is reformed, moving away from diffused ownership into decisive power blocs, and this is a transient resurgence of the classical firm with power again concentrated in those who have title to the residual.

In assessing the significance of stockholders' power it is not the usual diffusion of voting power that is significant but instead the frequency with which voting congeals into decisive changes. Even a one-man owned company may have a long term with just one manager—continuously being approved by the owner. Similarly a dispersed voting power corporation may be also characterized by a long-lived management. The question is the probability of replacement of the management if it behaves in ways not acceptable to a majority of the stockholders. The unrestricted salability of stock and the transfer of proxies enhances the probability of decisive action in the event current stockholders or any outsider believes that management is not doing a good job with the corporation. We are not comparing the corporate responsiveness to that of a single proprietorship; instead, we are indicating features of the corporate structure that are induced by the problem of

delegated authority to manager-monitors.¹⁴

D. Mutual and Nonprofit Firms

The benefits obtained by the new management are greater if the stock can be purchased and sold, because this enables *capitalization* of anticipated future im-

¹⁴ Instead of thinking of shareholders as joint *owners*, we can think of them as investors, like bondholders, except that the stockholders are more optimistic than bondholders about the enterprise prospects. Instead of buying bonds in the corporation, thus enjoying smaller risks, shareholders prefer to invest funds with a greater realizable return if the firm prospers as expected, but with smaller (possibly negative) returns if the firm performs in a manner closer to that expected by the more pessimistic investors. The pessimistic investors, in turn, regard only the bonds as likely to pay off.

If the entrepreneur-organizer is to raise capital on the best terms to him, it is to his advantage, as well as that of prospective investors, to recognize these differences in expectations. The residual claim on earnings enjoyed by shareholders does not serve the function of enhancing their efficiency as monitors in the general situation. The stockholders are "merely" the less risk-averse or the more optimistic member of the group that finances the firm. Being more optimistic than the average and seeing a higher mean value future return, they are willing to pay more for a certificate that allows them to realize gain on their expectations. One method of doing so is to buy claims to the distribution of returns that "they see" while bondholders, who are more pessimistic, purchase a claim to the distribution that they see as more likely to emerge. Stockholders are then comparable to warrant holders. They care not about the voting rights (usually not attached to warrants); they are in the same position in so far as voting rights are concerned as are bondholders. The only difference is in the probability distribution of rewards and the terms on which they can place their bets.

If we treat bondholders, preferred and convertible preferred stockholders, and common stockholders and warrant holders as simply different classes of investors—differing not only in their risk averseness but in their beliefs about the probability distribution of the firm's future earnings, why should stockholders be regarded as "owners" in any sense distinct from the other financial investors? The entrepreneur-organizer, who let us assume is the chief operating officer and sole repository of control of the corporation, does not find his authority residing in common stockholders (except in the case of a take over). Does this type of control make any difference in the way the firm is conducted? Would it make any difference in the kinds of behavior that would be tolerated by competing managers and investors (and we here deliberately refrain from thinking of them as owner-stockholders in the traditional sense)?

provements into present *wealth* of new managers who bought stock and created a larger capital by their management changes. But in nonprofit corporations, colleges, churches, country clubs, mutual savings banks, mutual insurance companies, and "coops," the future consequences of improved management are not

Investment old timers recall a significant incidence of nonvoting common stock, now prohibited in corporations whose stock is traded on listed exchanges. (Why prohibited?) The entrepreneur in those days could hold voting shares while investors held nonvoting shares, which in every other respect were identical. Nonvoting share holders were simply investors devoid of ownership connotations. The control and behavior of inside owners in such corporations has never, so far as we have ascertained, been carefully studied. For example, at the simplest level of interest, does the evidence indicate that nonvoting shareholders fared any worse because of not having voting rights? Did owners permit the nonvoting holders the normal return available to voting shareholders? Though evidence is prohibitively expensive to obtain, it is remarkable that voting and nonvoting shares sold for essentially identical prices, even during some proxy battles. However, our casual evidence deserves no more than interest-initiating weight.

One more point. The facade is deceptive. Instead of nonvoting shares, today we have warrants, convertible preferred stocks all of which are solely or partly "equity" claims without voting rights, though they could be converted into voting shares.

In sum, is it the case that the stockholder-investor relationship is one emanating from the *division of ownership* among several people, or is it that the collection of investment funds from people of varying anticipations is the underlying factor? If the latter, why should any of them be thought of as the owners in whom voting rights, whatever they may signify or however exercisable, should reside in order to enhance efficiency? Why voting rights in any of the outside, participating investors?

Our initial perception of this possibly significant difference in interpretation was precipitated by Henry Manne. A reading of his paper makes it clear that it is hard to understand why an investor who wishes to back and "share" in the consequences of some new business should necessarily have to acquire voting power (i.e., power to change the manager-operator) in order to invest in the venture. In fact, we invest in some ventures in the hope that no other stockholders will be so "foolish" as to try to toss out the incumbent management. We want him to have the power to stay in office, and for the prospect of sharing in his fortunes we buy nonvoting common stock. Our willingness to invest is enhanced by the knowledge that we can act legally via fraud, embezzlement and other laws to help assure that we outside investors will not be "milked" beyond our initial discounted anticipations.

capitalized into present wealth of stockholders. (As if to make more difficult that competition by new would-be monitors, multiple shares of ownership in those enterprises cannot be bought by one person.) One should, therefore, find greater shirking in nonprofit, mutually owned enterprises. (This suggests that nonprofit enterprises are especially appropriate in realms of endeavor where more shirking is desired and where redirected uses of the enterprise in response to market-revealed values is less desired.)

E. Partnerships

Team production in artistic or professional intellectual skills will more likely be by partnerships than other types of team production. This amounts to market-organized team activity and to a non-employer status. Self-monitoring partnerships, therefore, will be used rather than employer-employee contracts, and these organizations will be small to prevent an excessive dilution of efforts through shirking. Also, partnerships are more likely to occur among relatives or long-standing acquaintances, not necessarily because they share a common utility function, but also because each knows better the other's work characteristics and tendencies to shirk.

F. Employee Unions

Employee unions, whatever else they do, perform as monitors for employees. Employers monitor employees and similarly employees monitor an employer's performance. Are correct wages paid on time and in good currency? Usually, this is extremely easy to check. But some forms of employer performance are less easy to meter and are more subject to employer shirking. Fringe benefits often are in non-pecuniary, contingent form; medical, hospital, and accident insurance, and retirement pensions are contingent payments

or performances partly in *kind* by employers to employees. Each employee cannot judge the character of such payments as easily as money wages. Insurance is a contingent payment—what the employee will get upon the contingent event may come as a disappointment. If he could easily determine what other employees had gotten upon such contingent events he could judge more accurately the performance by the employer. He could "trust" the employer not to shirk in such fringe contingent payments, but he would prefer an effective and economic monitor of those payments. We see a specialist monitor—the union employees' agent—hired by them and monitoring those aspects of employer payment most difficult for the employees to monitor. Employees should be willing to employ a specialist monitor to administer such hard-to-detect employer performance, even though their monitor has incentives to use pension and retirement funds not entirely for the benefit of employees.

V. Team Spirit and Loyalty

Every team member would prefer a team in which no one, not even himself, shirked. Then the true marginal costs and values could be equated to achieve more preferred positions. If one could enhance a common interest in nonshirking in the guise of a team loyalty or team spirit, the team would be more efficient. In those sports where team activity is most clearly exemplified, the sense of loyalty and team spirit is most strongly urged. Obviously the team is better, with team spirit and loyalty, because of the reduced shirking—not because of some other feature inherent in loyalty or spirit as such.¹⁵

¹⁵ *Sports Leagues:* Professional sports contests among teams is typically conducted by a *league* of teams. We assume that sports consumers are interested not only in absolute sporting skill but also in skills *relative* to other teams. Being slightly better than opposing teams enables one to claim a major portion of the receipts; the

Corporations and business firms try to instill a spirit of loyalty. This should not be viewed simply as a device to increase profits by *over-working* or misleading the employees, nor as an adolescent urge for belonging. It promotes a closer approximation to the employees' potentially available true rates of substitution between production and leisure and enables each team member to achieve a more preferred

inferior team does not release resources and reduce costs, since they were expected in the play of contest. Hence, absolute skill is developed beyond the equality of marginal investment in sporting skill with its true social marginal value product. It follows there will be a tendency to overinvest in training athletes and developing teams. "Reverse shirking" arises, as budding players are induced to overpractice hyperactively relative to the social marginal value of their enhanced skills. To prevent overinvestment, the teams seek an agreement with each other to restrict practice, size of teams, and even pay of the team members (which reduces incentives of young people to overinvest in developing skills). Ideally, if all the contestant teams were owned by one owner, overinvestment in sports would be avoided, much as ownership of common fisheries or underground oil or water reserve would prevent overinvestment. This hyperactivity (to suggest the opposite of shirking) is controlled by the league of teams, wherein the league adopts a common set of constraints on each team's behavior. In effect, the teams are no longer really owned by the team owners but are supervised by them, much as the franchisers of some product. They are not full-fledged owners of their business, including the brand name, and can not "do what they wish" as franchises. Comparable to the franchiser, is the league commissioner or conference president, who seeks to restrain hyperactivity, as individual team supervisors compete with each other and cause external diseconomies. Such restraints are usually regarded as anticompetitive, anti-social, collusive-cartel devices to restrain free open competition, and reduce players' salaries. However, the interpretation presented here is premised on an attempt to avoid hyperinvestment in team sports production. Of course, the team operators have an incentive, once the league is formed and restraints are placed on hyperinvestment activity, to go further and obtain the private benefits of monopoly restriction. To what extent overinvestment is replaced by monopoly restriction is not yet determinable; nor have we seen an empirical test of these two competing, but mutually consistent interpretations. (This interpretation of league-sports activity was proposed by Earl Thompson and formulated by Michael Canes.) Again, athletic teams clearly exemplify the specialization of monitoring with captains and coaches; a captain detects shirkers while the coach trains and selects strategies and tactics. Both functions may be centralized in one person.

situation. The difficulty, of course, is to create economically that team spirit and loyalty. It can be preached with an aura of moral code of conduct—a morality with literally the same basis as the ten commandments—to restrict our conduct toward what we would choose if we bore our full costs.

VI. Kinds of Inputs Owned by the Firm

To this point the discussion has examined why firms, as we have defined them, exist? That is, why is there an owner-employer who is the common party to contracts with other owners of inputs in team activity? The answer to that question should also indicate the kind of the jointly used resources likely to be owned by the central-owner-monitor and the kind likely to be hired from people who are not team-owners. Can we identify characteristics or features of various inputs that lead to their being hired or to their being owned by the firm?

How can residual-claimant, central-employer-owner demonstrate ability to pay the other hired inputs the promised amount in the event of a loss? He can pay in advance or he can commit wealth sufficient to cover negative residuals. The latter will take the form of machines, land, buildings, or raw materials committed to the firm. Commitments of labor-wealth (i.e., human wealth) given the property rights in people, is less feasible. These considerations suggest that residual claimants—owners of the firm—will be investors of resalable capital equipment in the firm. The goods or inputs more likely to be invested, than rented, by the owners of the enterprise, will have higher resale values relative to the initial cost and will have longer expected use in a firm relative to the economic life of the good.

But beyond these factors are those developed above to explain the existence of

the institution known as the firm—the costs of detecting output performance. When a durable resource is used it will have a marginal product and a depreciation. Its use requires payment to cover at least use-induced depreciation; unless that user cost is specifically detectable, payment for it will be demanded in accord with *expected* depreciation. And we can ascertain circumstances for each. An indestructible hammer with a readily detectable marginal product has zero user cost. But suppose the hammer were destructible and that careless (which is easier than careful) use is more abusive and causes greater depreciation of the hammer. Suppose in addition the abuse is easier to detect by observing the way it is used than by observing only the hammer after its use, or by measuring the output scored from a hammer by a laborer. If the hammer were rented and used in the absence of the owner, the depreciation would be greater than if the use were observed by the owner and the user charged in accord with the imposed depreciation. (Careless use is more likely than careful use—if one does not pay for the greater depreciation.) An absentee owner would therefore ask for a higher rental price because of the higher *expected* user cost than if the item were used by the owner. The expectation is higher because of the greater difficulty of observing specific user cost, by inspection of the hammer after use. Renting is therefore in this case more costly than owner use. This is the valid content of the misleading expressions about ownership being more economical than renting—ignoring all other factors that may work in the opposite direction, like tax provision, short-term occupancy and capital risk avoidance.

Better examples are tools of the trade. Watch repairers, engineers, and carpenters tend to own their own tools especially if

they are portable. Trucks are more likely to be employee owned rather than other equally expensive team inputs because it is relatively cheap for the driver to police the care taken in using a truck. Policing the use of trucks by a nondriver owner is more likely to occur for trucks that are not specialized to one driver, like public transit busses.

The factor with which we are concerned here is one related to the costs of monitoring not only the gross product performance of an input but also the abuse or depreciation inflicted on the input in the course of its use. If depreciation or user cost is more cheaply detected when the owner can see its use than by only seeing the input before and after, there is a force toward owner use rather than renting. Resources whose user cost is harder to detect when used by someone else, tend on this count to be owner-used. Absentee ownership, in the lay language, will be less likely. Assume momentarily that labor service cannot be performed in the absence of its owner. The labor owner can more cheaply monitor any abuse of himself than if somehow labor-services could be provided without the labor owner observing its mode of use or knowing what was happening. Also his incentive to abuse himself is increased if he does not own himself.¹⁶

¹⁶ Professional athletes in baseball, football, and basketball, where athletes having sold their source of service to the team owners upon entering into sports activity, are owned by team owners. Here the team owners must monitor the athletes' physical condition and behavior to protect the team owners' wealth. The athlete has *less* (not, *no*) incentive to protect or enhance his athletic prowess since capital value changes have less impact on his own wealth and more on the team owners. Thus, some athletes sign up for big initial bonuses (representing present capital value of future services). Future salaries are lower by the annuity value of the prepaid "bonus" and hence the athlete has *less* to lose by subsequent abuse of his athletic prowess. Any decline in his subsequent service value would in part be borne by the team owner who owns the players' future service. This does not say these losses of future salaries have no effect on preservation of athletic talent (we are not making a "sunk cost" error). Instead, we assert that the

The similarity between the preceding analysis and the question of absentee landlordism and of sharecropping arrangements is no accident. The same factors which explain the contractual arrangements known as a firm help to explain the incidence of tenancy, labor hiring or sharecropping.¹⁷

VII. Firms as a Specialized Market Institution for Collecting, Collating, and Selling Input Information

The firm serves as a highly specialized surrogate market. Any person contemplating a joint-input activity must search and detect the qualities of available joint inputs. He could contact an employment agency, but that agency in a small town would have little advantage over a large firm with many inputs. The employer, by virtue of monitoring many inputs, acquires special superior information about their productive talents. This aids his *directive* (i.e., market hiring) efficiency. He "sells" his information to employee-inputs as he aids them in ascertaining good input combinations for team activity. Those who work as employees or who rent services to him are using him to discern superior combinations of inputs. Not only

does the director-employer "decide" what each input will produce, he also estimates which heterogeneous inputs will work together jointly more efficiently, and he does this in the context of a privately owned market for forming teams. The department store is a firm and is a superior private market. People who shop and work in one town can as well shop and work in a privately owned firm.

This marketing function is obscured in the theoretical literature by the assumption of homogeneous factors. Or it is tacitly left for individuals to do themselves via personal market search, much as if a person had to search without benefit of specialist retailers. Whether or not the firm arose because of this efficient information service, it gives the director-employer more knowledge about the productive talents of the team's inputs, and a basis for superior decisions about efficient or profitable combinations of those heterogeneous resources.

In other words, opportunities for profitable team production by inputs already within the firm may be ascertained more economically and accurately than for resources outside the firm. Superior combinations of inputs can be more economically identified and formed from resources already used in the organization than by obtaining new resources (and knowledge of them) from the outside. Promotion and revision of employee assignments (contracts) will be preferred by a firm to the hiring of new inputs. To the extent that this occurs there is reason to expect the firm to be able to operate as a conglomerate rather than persist in producing a single product. Efficient production with heterogeneous resources is a result not of having *better* resources but in *knowing more accurately* the relative productive performances of those resources. Poorer resources can be paid less in accord with their inferiority; greater accuracy of

preservation is reduced, not eliminated, because the amount of loss of wealth suffered is smaller. The athlete will spend less to maintain or enhance his prowess thereafter. The effect of this revised incentive system is evidenced in comparisons of the kinds of attention and care imposed on the athletes at the "expense of the team owner" in the case where athletes' future services are owned by the team owner with that where future labor service values are owned by the athlete himself. Why athletes' future athletic services are owned by the team owners rather than being hired is a question we should be able to answer. One presumption is cartelization and monopsony gains to team owners. Another is exactly the theory being expounded in this paper—costs of monitoring production of athletes; we know not on which to rely.

¹⁷ The analysis used by Cheung in explaining the prevalence of sharecropping and land tenancy arrangements is built squarely on the same factors—the costs of detecting output performance of jointly used inputs in team production and the costs of detecting user costs imposed on the various inputs if owner used or if rented.

knowledge of the potential and actual productive actions of inputs rather than having high productivity resources makes a firm (or an assignment of inputs) profitable.¹⁸

VIII. Summary

While ordinary contracts facilitate efficient specialization according to comparative advantage, a special class of contracts among a group of joint inputs to a team production process is commonly used for team production. Instead of multilateral contracts among all the joint inputs' owners, a central common party to a set of bilateral contracts facilitates efficient organization of the joint inputs in team production. The terms of the contracts form the basis of the entity called the firm—especially appropriate for organizing team production processes.

Team productive activity is that in which a union, or joint use, of inputs yields a larger output than the sum of the products of the separately used inputs. This

team production requires—like all other production processes—an assessment of marginal productivities if efficient production is to be achieved. Nonseparability of the products of several differently owned joint inputs raises the cost of assessing the marginal productivities of those resources or services of each input owner. Monitoring or metering the productivities to match marginal productivities to costs of inputs and thereby to reduce shirking can be achieved more economically (than by across market bilateral negotiations among inputs) in a firm.

The essence of the classical firm is identified here as a contractual structure with: 1) joint input production; 2) several input owners; 3) one party who is common to all the contracts of the joint inputs; 4) who has rights to renegotiate any input's contract independently of contracts with other input owners; 5) who holds the residual claim; and 6) who has the right to sell his central contractual residual status. The central agent is called the firm's owner and the employer. No authoritarian control is involved; the arrangement is simply a contractual structure subject to continuous renegotiation with the central agent. The contractual structure arises as a means of enhancing efficient organization of team production. In particular, the ability to detect shirking among owners of jointly used inputs in team production is enhanced (detection costs are reduced) by this arrangement and the discipline (by revision of contracts) of input owners is made more economic.

Testable implications are suggested by the analysis of different types of organizations—nonprofit, proprietary for profit, unions, cooperatives, partnerships, and by the kinds of inputs that tend to be owned by the firm in contrast to those employed by the firm.

We conclude with a highly conjectural

¹⁸ According to our interpretation, the firm is a specialized surrogate for a market for team use of inputs; it provides superior (i.e., cheaper) collection and collation of knowledge about heterogeneous resources. The greater the set of inputs about which knowledge of performance is being collated within a firm the greater are the present costs of the collation activity. Then, the larger the firm (market) the greater the attenuation of monitor control. To counter this force, the firm will be divisionalized in ways that economize on those costs—just as will the market be specialized. So far as we can ascertain, other theories of the reasons for firms have no such implications.

In Japan, employees by custom work nearly their entire lives with one firm, and the firm agrees to that expectation. Firms will tend to be large and conglomerate to enable a broader scope of input revision. Each firm is, in effect, a small economy engaging in "international and international" trade. Analogously, Americans expect to spend their whole lives in the United States, and the bigger the country, in terms of variety of resources, the easier it is to adjust to changing tastes and circumstances. Japan, with its lifetime employees, should be characterized more by large, conglomerate firms. Presumably, at some size of the firm, specialized knowledge about inputs becomes as expensive to transmit across divisions of the firms as it does across markets to other firms.

but possibly significant interpretation. As a consequence of the flow of information to the central party (employer), the firm takes on the characteristic of an efficient market in that information about the productive characteristics of a large set of specific inputs is now more cheaply available. Better recombinations or new uses of resources can be more efficiently ascertained than by the conventional search through the general market. In this sense inputs compete with each other within and via a firm rather than solely across markets as conventionally conceived. Emphasis on interfirm competition obscures intrafirm competition among inputs. Conceiving competition as the *revelation and exchange* of knowledge or information about qualities, potential uses of different inputs in different potential applications indicates that the firm is a device for enhancing competition among sets of input resources as well as a device for more efficiently rewarding the inputs. In contrast to markets and cities which can be viewed as publicly or nonowned market places, the firm can be considered a privately owned market; if so, we could consider the firm and the ordinary market as competing types of markets, competition between private proprietary markets and public or communal markets. Could it be that the market suffers from the defects of com-

munal property rights in organizing and influencing uses of valuable resources?

REFERENCES

- M. Canes, "A Model of a Sports League," unpublished doctoral dissertation, UCLA 1970.
- S. N. Cheung, *The Theory of Share Tenancy*, Chicago 1969.
- R. H. Coase, "The Nature of the Firm," *Economica*, Nov. 1937, 4, 386-405; reprinted in G. J. Stigler and K. Boulding, eds., *Readings in Price Theory*, Homewood 1952, 331-51.
- E. Furobotn and S. Pejovich, "Property Rights and the Behavior of the Firm in a Socialist State," *Zeitschrift für Nationalökonomie*, 1970, 30, 431-454.
- F. H. Knight, *Risk, Uncertainty and Profit*, New York 1965.
- S. Macaulay, "Non-Contractual Relations in Business: A Preliminary Study," *Amer. Sociological Rev.*, 1968, 28, 55-69.
- H. B. Malmgren, "Information, Expectations and the Theory of the Firm," *Quart J. Econ.*, Aug. 1961, 75, 399-421.
- H. Manne, "Our Two Corporation Systems: Law and Economics," *Virginia Law Rev.*, Mar. 1967, 53, No. 2, 259-84.
- S. Pejovich, "The Firm, Monetary Policy and Property Rights in a Planned Economy," *Western Econ. J.*, Sept. 1969, 7, 193-200.
- M. Silver and R. Auster, "Entrepreneurship, Profit, and the Limits on Firm Size," *J. Bus. Univ. Chicago*, Apr. 1969, 42, 277-81.
- E. A. Thompson, "Nonpecuniary Rewards and the Aggregate Production Function," *Rev. Econ. Statist.*, Nov. 1970, 52, 395-404.

The Economic Theory of Agency: The Principal's Problem

By STEPHEN A. ROSS*

The relationship of agency is one of the oldest and commonest codified modes of social interaction. We will say that an agency relationship has arisen between two (or more) parties when one, designated as the agent, acts for, on behalf of, or as representative for the other, designated the principal, in a particular domain of decision problems. Examples of agency are universal. Essentially all contractual arrangements, as between employer and employee or the state and the governed, for example, contain important elements of agency. In addition, without explicitly studying the agency relationship, much of the economic literature on problems of moral hazard (see K. J. Arrow) is concerned with problems raised by agency. In a general equilibrium context the study of information flows (see J. Marschak and R. Radner) or of financial intermediaries in monetary models is also an example of agency theory.

The canonical agency problem can be posed as follows. Assume that both the agent and the principal possess state independent von Neumann-Morgenstern utility functions, $G(\cdot)$ and $U(\cdot)$ respectively, and that they act so as to maximize their expected utility. The problems of agency are really most interesting when seen as involving choice under uncertainty and this is the view we will adopt. The agent may choose an act, $a \in A$, a feasible action space, and the random payoff from

this act, $w(a, \theta)$, will depend on the random state of nature θ ($\epsilon \Omega$ the state space set), unknown to the agent when a is chosen. By assumption the agent and the principal have agreed upon a fee schedule f to be paid to the agent for his services. The fee, f , is generally a function of both the state of the world, θ , and the action, a , but we will assume that the action can influence the parties and, hence, the fee only through its impact on the payoff. This permits us to write,

$$(1) \quad f = f(w(a, \theta); \theta).$$

Two points deserve mention. Obviously the choice of a fee schedule is the outcome of a bargaining problem or, in large games, of a market process. Much of what we have to say is relevant for this view but we will not treat the bargaining problem explicitly. Second, while it is possible to conceive of the fee as being directly functionally dependent on the act, the theory loses much of its interest, since without further conditions, such a fee can always be chosen as a Dirac δ -function forcing a particular act (see S. Ross). In some sense, then, we are assuming that only the payoff is operational and we will take this point up below. Now, the agent will choose an act, a , so as to

$$(2) \quad \max_a E\{G[f(w(a, \theta); \theta)]\},$$

where the agent takes the expectation over his subjectively held probability distribution. The solution to the agent's problem involves the choice of an optimal act, a^o , conditional on the particular fee schedule, i.e., $a^o = a(\langle f \rangle)$, where $a(\cdot)$ is a

* Associate professor of economics, University of Pennsylvania. This work was supported by grants from the Rodney L. White Center for Financial Research at the University of Pennsylvania and from the National Science Foundation.

mapping from the space of fee schedules into A .

If the principal has complete information about the fee to act mapping, $a(\langle f \rangle)$, he will now choose a fee so as to

$$(3) \quad \max_{\langle f \rangle} \max_{\theta} E\{U[w(a(\langle f \rangle), \theta) - f(w(a(\langle f \rangle), \theta); \theta)]\},$$

where the expectation is taken over the principal's subjective probability distribution over states of nature. If the principal is not fully informed about $a(\cdot)$, then $a(\cdot)$ will be a random function from his point of view. Formally, at least, by appropriately augmenting the state space the criterion (3) could still be made to apply. In general some side constraints on $\langle f \rangle$ would also have to be imposed to insure that the problem possesses a solution (see Ross). A market-imposed minimum expected fee or expected utility of fee by the agent would be one economically sensible constraint:

$$(4) \quad E\{G[f(w(a, \theta); \theta)]\} \geq k.$$

Since utility functions are assumed to be independent of states, θ , one of the important reasons for a fee to depend directly on θ would be if individual subjective probability distributions differed. In what follows we will assume that both the agent and the principal share the same subjective beliefs about the occurrence of θ and write the fee as a function of the payoff only,

$$(5) \quad f = f(w(a, \theta)).$$

Notice that this interpretation would not in general be permissible if the principal lacked perfect knowledge of $a(\cdot)$. More importantly, though, surely aside from simple comparative advantage, for some questions the *raison d'être* for an agency relationship is that the agent (or the principal) may possess different (better or finer) information about the states of

the world than the principal (agent). If we abstract from this possibility we will have to show that we are not throwing out the baby with the bath water.

Under this assumption the problem is considerably simplified but much of interest does remain. Suppose, first, that we are simply interested in the properties of Pareto-efficient arrangements that the agent and the principal will strike. Notice that the optimal fee schedule as seen by the principal is found by solving (3) and is dependent on the desire to motivate the agent. In general, then, we would expect such an arrangement to be Pareto-in-efficient, but we will return to this point below. The family of Pareto-efficient fee schedules can be characterized by assuming that the principal and the agent co-operate to choose a schedule that maximizes a weighted sum of utilities

$$(6) \quad \max_{\langle f \rangle} E\{U[w - f] + \lambda G[f]\},$$

where λ is a relative weighting factor (and where strategies have been randomized to insure convexity). K. Borch recognized that the solution to (6) is obtained by maximizing the function internal to the expectation which requires setting

$$(\text{P.E.}) \quad U'[w - f] = \lambda G'[f]$$

when U and G are monotone and concave. (See H. Raiffa for a good exposition.) The P.E. condition defines the fee schedule, $f(\cdot)$, as a function of the payoff w (and the weight, λ). (See R. Wilson (1968) or Ross for a fuller discussion of this derivation and the functional aspect of the fee schedule.)

An alternative approach to finding optimal fee schedules was first proposed by Wilson in the theory of syndicates and studied by Wilson (1968, 1969) and Ross. This is the similarity condition that solves for the fee schedule by setting

$$(S) \quad U[w - f] = aG[f] + b$$

for constants $a > 0, b$. If $\langle f \rangle$ satisfies S then, given the fee schedule, it should be clear that the agent and the principal have identical attitudes towards risky payoffs and, consequently, the agent will always choose the act that the principal most desires. Ross was able to completely characterize the class of utility functions that satisfied both P.E. and S (for a range of λ) and show that in such situations the fee schedule is (affine) linear, L , in the payoff. (The class is simply that of pairs $\langle U, G \rangle$ with linear risk tolerance,

$$-\frac{U'}{U''} = cw + d \quad \text{and} \quad -\frac{G'}{G''} = cw + e,$$

where c, d and e are constants.) In fact, it can be shown that any two of S , P.E., or L imply the third.

A question of interest that naturally arises is that of the relation that S and P.E. bear to the exact solution to the principal's problem. (A comparable "agent's problem" can also be posed but we will not be concerned with that here. Some observations on such a problem are contained in Ross.) The solution to the principal's problem (3) subject to the constraint (4) and to the constraint imposed by the condition that the agent chooses the optimal act from his problem (2) can, under some circumstances, be posed as a classical variational problem. To do so we will assume that the payoff function is (twice) differentiable and that the agent chooses an optimal act, given a fee schedule, by the first order condition

$$(7) \quad \underset{\theta}{E}\{G'[f(w)]f'(w)w_a\} = 0,$$

where a subscript indicates partial differentiation. The principal's problem is now to

$$(8) \quad \max_{\langle f \rangle} E\{H\} \equiv \max_{\langle f \rangle} E\{U[w - f] \\ + \Psi G'f'w_a + \lambda G\}$$

where Ψ and λ are Lagrange multipliers associated with the constraints (7) and (4) respectively. Changing variables to $V(\theta) \equiv f(w(a, \theta))$ where we have suppressed the impact of a on V and assuming, without loss of generality, that θ is uniformly distributed on $[0, 1]$ permits us to solve (8) by the Euler-Lagrange equation. Thus, at an optimum

$$(9) \quad \frac{d}{d\theta} \left\{ \frac{\partial H}{\partial V'} \right\} - \frac{\partial H}{\partial V} \\ = U' + \Psi G' \frac{d}{d\theta} \left[\frac{w_a}{w_\theta} \right] - \lambda G' = 0;$$

or the marginal rate of substitution,

$$(10) \quad \frac{U'}{G'} = \lambda - \Psi \frac{d}{d\theta} \left[\frac{w_a}{w_\theta} \right].$$

This is an intuitively appealing result; the marginal rate of substitution is set equal to a constant as in the P.E. condition plus an additional term which captures the constraint (7) imposed on the principal by the need to motivate the agent. To determine the optimal act, a , we differentiate (8) with respect to a which yields

$$(11) \quad \underset{\theta}{E}\{U'[1 - f']w_a + \Psi G''(f'w_a)^2 \\ + \Psi G'f''(w_a)^2 + \Psi G'f'w_{aa}\} = 0,$$

where we have made use of (7). Substituting the boundary conditions permits us to solve for the multipliers Ψ and λ .

Like S or P.E. (10) defines the fee schedule as a function of w . (Notice that we are tacitly assuming that, at least for the optimal act, the payoff is (*a.e.* locally) state invertible. This allows the fee to take the form of (5).) It follows that (10) will coincide with P.E. if and only if Ψ is zero, or if $\Psi \neq 0$, we must have

$$(12) \quad \frac{d}{d\theta} \left[\frac{w_a}{w_\theta} \right] = b(a),$$

a function of a alone.

In particular, using these conditions we can ask what class of (pairs of) utility functions $\langle U, G \rangle$ has the property that, for any payoff structure, $w(a, \theta)$, the solution to the principal's problem is Pareto-efficient. Conversely, we can ask what class of payoff structures has the property that the principal's problem yields a Pareto-efficient solution for any pair of utility functions $\langle U, G \rangle$.

A little reflection reveals that the only pairs of $\langle U, G \rangle$ that could possibly belong to the first class must be those which satisfy S and P.E. for a range of schedules (indexed by the λ weight in P.E.). Clearly if (10) is to be equivalent to P.E. for all payoff functions, $w(a, \theta)$, then Ψ must be zero and the motivational constraint (7) must not be binding. For this to be the case, for an interval of values of k (in (4)), the satisfaction of P.E. must imply that the agent chooses the principal's most desired act by (7). For any fee schedule, $\langle f \rangle$, the principal wants the act to be chosen to maximize $E_\theta \{U[w-f]\}$ which implies that

$$(13) \quad E_\theta \{U'(1-f')w_a\} = 0.$$

If (13) is to be equivalent to the motivational constraint (7) for all possible payoff structures, then we must have

$$(14) \quad U'(1-f') = G'f'$$

which, with P.E. (or (10) with $\Psi=0$) yields a linear fee schedule in the payoff. But, as shown in Ross, linearity of the fee schedule and P.E. imply the satisfaction of S and the $\langle U, G \rangle$ pair must belong to the linear risk-tolerance class of utility functions described above.

Since the linear risk-tolerance class, while important, is very limited, we turn

now to the converse question of what payoff structures permit a Pareto-efficient solution for all $\langle U, G \rangle$ pairs. If $\Psi=0$ we must, as before, have that the motivational constraint is not binding for all $\langle U, G \rangle$ or (13) must always imply (7). The implication will always hold if there exists an a^* such that for all a there is some choice of the state domain, I , for which

$$(15) \quad w(a^*, \theta) \geq w(a, \theta), \quad \theta \in I.$$

Conversely, from P.E., we must have that for all $G(\cdot)$

$$(16) \quad E_\theta \{G'[f](1-f')w_a\} = 0$$

implies (7) where f is determined by P.E. Since $\langle U, G \rangle$ can always be chosen so as to attain any desired weightings of w_a in (7) and (16) the special case of (15) is the only one for which motivation is irrelevant. Given (15) all individuals have a uniquely optimal act irrespective of their attitudes towards risk.

If $\Psi \neq 0$, then to assure Pareto efficiency we must satisfy (12). This is a partial differential equation and its solution is given by

$$(17) \quad w(a, \theta) = H[\theta B(a) - C(a)],$$

where $H(\cdot)$, $B(\cdot)$ and $C(\cdot)$ are arbitrary functions. (The detailed computations are carried out in an appendix.) This is a rich and interesting class of payoff functions. In particular, (17) is a generalization of the class of functions of the form $l(\theta-a)$, where the object is to pick an act, a , so as to best guess the state θ . It therefore includes, for example, traditional estimation problems, problems with a quadratic payoff function, and all problems with payoff functions of the form $|\theta-a|^{\frac{1}{2}}h(a)$, and many asymmetric ones as well. It is not, however, difficult to find plausible payoff functions which do not take the form of (17). (The class of the form (15) will generate such functions.)

We may conclude, then, that the class of payoff structures that simultaneously solve the principal's problem and lead to Pareto efficiency for all $\langle U, G \rangle$ pairs is quite important and quite likely to arise in practice.

In general, though, it is clear that the solution to the principal's problem will not be Pareto-efficient. This is, however, a somewhat naive view to take. Pareto efficiency as defined above assumes that perfect information is held by the participants. In fact, the optimal solution to the principal's problem implied that the fee-to-act mapping induced by the agent was completely known to the principal. In such a case it might be thought that the principal could simply tell the agent to perform a particular act. The difficulty arises in monitoring the act that the agent chooses. Michael Spence and Richard Zeckhauser have examined this problem in detail in the case of insurance. In addition, if agents are numerous the fee may be the only communication mechanism. While it might in principle be feasible to monitor the agent's actions, it would not be economically viable to do so.

The format of this paper has been such as to allow us to only touch on what is surely the most challenging aspect of agency theory; embedding it in a general equilibrium market context. Much is to be learned from such attempts. One would naturally expect a market to arise in the services of agents. Furthermore, in some sense, such a market serves as a surrogate for a market in the information possessed by agents. To the extent to which this occurs, the study of agency in market contexts should shed some light on the economics of information. To mention one more path of interest—in a world of true uncertainty where adequate contingent markets do not exist, the manager of the firm is essentially an agent of the shareholders. It can, therefore, be expected that

an understanding of the agency relationship will aid our understanding of this difficult question.

The results obtained here provide some of the micro foundations for such studies. We have shown that, for an interesting class of utility functions and for a very broad and relevant class of payoff structures, the need to motivate agents does not conflict with the attainment of Pareto efficiency. At the least, a callous observer might view these results as providing some solace to those engaged in econometric activity.

APPENDIX

This appendix solves the partial differential equation (12) in the text. Integrating (12) over θ yields

$$\frac{\partial w}{\partial a} + [b(a)\theta + c(a)] \frac{\partial w}{\partial \theta} = 0.$$

Along a locus of constant w ,

$$\frac{d\theta}{da} = -\frac{\partial w/\partial a}{\partial w/\partial \theta} = b(a)\theta + c(a),$$

is a first order Bernoulli equation that integrates to

$$\theta = e^{\int b(a)} \left[\int e^{-\int b(a)} c(a) + k \right],$$

where k is a constant of integration. It follows that

$$w(a, \theta) = H[\theta B(a) - C(a)],$$

where

$$B(a) \equiv e^{-\int b(a)}$$

and

$$C(a) \equiv \int e^{-\int b(a)} c(a) + k,$$

and $H(\cdot)$ is an arbitrary function.

REFERENCES

- K. J. Arrow, *Essays in the Theory of Risk-Bearing*, Chicago 1970.
- K. Borch, "Equilibrium in a Reinsurance Market," *Econometrica*, July 1962, 30, 424-444.
- J. Marschak and R. Radner, *The Economic Theory of Teams*, New Haven and London 1972.
- H. Raiffa, *Decision Analysis; Introductory Lectures on Choices Under Uncertainty*, Reading, Mass. 1968.
- S. Ross, "On the Economic Theory of Agency: The Principle of Similarity," *Proceedings of the NBER-NSF Conference on Decision Making and Uncertainty*, forthcoming.
- M. Spence and R. Zeckhauser, "Insurance, Information and Individual Action," *Amer. Econ. Rev. Proc.*, May 1971, 61, 380-387.
- R. Wilson, "On the Theory of Syndicates," *Econometrica*, Jan. 1968, 36, 119-132.
- , "The Structure of Incentives for Decentralization Under Uncertainty," *La Decision*, Editions Du Centre National De La Recherche Scientifique, Paris 1969.

Some International Evidence on Output-Inflation Tradeoffs

By ROBERT E. LUCAS, JR.*

This paper reports the results of an empirical study of real output-inflation tradeoffs, based on annual time-series from eighteen countries over the years 1951-67. These data are examined from the point of view of the hypothesis that average real output levels are invariant under changes in the time pattern of the rate of inflation, or that there exists a "natural rate" of real output. That is, we are concerned with the questions (i) does the natural rate theory lead to expressions of the output-inflation relationship which perform satisfactorily in an econometric sense for all, or most, of the countries in the sample, (ii) what testable restrictions does the theory impose on this relationship, and (iii) are these restrictions consistent with recent experience?

Since the term "natural rate theory" refers to varied aggregation of models and verbal developments,¹ it may be helpful to sketch the key elements of the particular version used in this paper. The first essential presumption is that *nominal* output is determined on the aggregate demand side of the economy, with the division into real output and the price level largely dependent on the behavior of *suppliers* of labor and goods. The second is that the partial "rigidities" which dominate short-run supply behavior result from suppliers' lack of information on some of the prices relevant to their decisions. The third

presumption is that inferences on these relevant, unobserved prices are made optimally (or "rationally") in light of the stochastic character of the economy.

As I have argued elsewhere (1972), theories developed along these lines will *not* place testable restrictions on the coefficients of estimated Phillips curves or other single equation expressions of the tradeoff. They will not, for example, imply that money wage changes are linked to price level changes with a unit coefficient, or that "long-run" (in the usual distributed lag sense) Phillips curves must be vertical. They *will* (as we shall see below) link supply parameters to parameters governing the stochastic nature of demand shifts. The fact that the implications of the natural rate theory come in this form suggests an attempt to test it using a sample, such as the one employed in this study, in which a wide variety of aggregate demand behavior is exhibited.

In the following section, a simple aggregative model will be constructed using the elements sketched above. Results based on this model are reported in Section II, followed by a discussion and conclusions.

I. An Economic Model

The general structure of the model developed in this section may be described very simply. First, the aggregate price-quantity observations are viewed as intersection points of an aggregate demand and an aggregate supply schedule. The former is drawn up under the assumption of a cleared money market and represents the output-price level relationship implicit in

* Graduate School of Industrial Administration, Carnegie-Mellon University.

¹ The most useful, general statements are those of Milton Friedman (1968) and Edmund Phelps. Specific illustrative examples are provided by Donald Gordon and Allan Hynes and Lucas (April 1972).

the standard IS-LM diagram. It is viewed as being shifted by the usual set of demand-shift variables: monetary and fiscal policies and variation in export demands. The supply schedule is drawn under the assumption of a cleared labor market; its slope therefore reflects labor and product market "rigidities."

The structure of this model, which is essentially that suggested in Lucas and Leonard Rapping (1969), will be greatly simplified by an additional special assumption: that the aggregate demand curve is unit elastic.² In this case, the level of nominal output can be treated as an "exogenous" variable with respect to the goods market, and the entire burden of accounting for the breakdown of nominal income into real output and price is placed on the aggregate supply side. In the next subsection, *A*, a supply model designed to serve this purpose is developed. In subsection *B*, solutions to the full (demand and supply) model are obtained.

A. Aggregate Supply

All formulations of the natural rate theory postulate rational agents, whose decisions depend on *relative* prices only, placed in an economic setting in which they cannot distinguish relative from general price movements. Obviously, there is no limit to the number of models one can construct where agents are placed in this situation of imperfect information; the trick is to find tractable schemes with this feature. One such model is developed below.

We imagine suppliers as located in a large number of scattered, competitive markets. Demand for goods in each period

² An explicit derivation of the price-output relationship from the IS-LM framework is given by Frederic Raines. Of course, this framework does not imply an elasticity of unity, though it is consistent with it. Since the unit elasticity hypothesis is primarily a matter of convenience in the present study, I shall comment below on the probable consequences of relaxing it.

is distributed unevenly over markets, leading to relative as well as general price movements. As a consequence, the situation as perceived by individual suppliers will be quite different from the aggregate situation as seen by an outside observer. Accordingly, we shall attempt to keep these two points of view separate, turning first to the situation faced by individual suppliers.

Quantity supplied in each market will be viewed as the product of a normal (or secular) component common to all markets and a cyclical component which varies from market to market. Letting z index markets, and using y_{nt} and y_{ct} to denote the *logs* of these components, supply in market z is:

$$(1) \quad y_t(z) = y_{nt} + y_{ct}(z)$$

The secular component, reflecting capital accumulation and population change, follows the trend line:

$$(2) \quad y_{nt} = \alpha + \beta t$$

The cyclical component varies with perceived, *relative* prices and with its own lagged value:

$$(3) \quad y_{ct}(z) = \gamma [P_t(z) - E(P_t | I_t(z))] \\ + \lambda y_{ct-1}(z)$$

where $P_t(z)$ is the actual price in z at t and $E(P_t | I_t(z))$ is the mean current, general price level, conditioned on information available in z at t , $I_t(z)$.³ Since y_{ct} is a deviation from trend, $|\lambda| < 1$.

³ A supply function for labor which varies with the ratio of actual to expected prices is developed and verified empirically by Lucas and Rapping (1969). The effect of lagged on actual employment is also shown. In our 1972 paper, in response to Albert Rees's criticism, we found that this persistence in employment cannot be fully explained by price expectations behavior. Both these effects—an expectations and a persistence effect—will be transmitted by firms to the goods market. In addition, they are probably augmented by speculative behavior on the part of firms (as analyzed for example, by Paul Taubman and Maurice Wilkinson).

For a general equilibrium model in which suppliers behave essentially as given by (3), see my 1972 papers.

The information available to suppliers in z at t comes from two sources. First, traders enter period t with knowledge of the past course of demand shifts, of normal supply y_{nt} , and of past deviations $y_{c,t-1}, y_{c,t-2}, \dots$. While this information does not permit exact inference of the *log* of the current general price level, P_t , it does determine a "prior" distribution on P_t , common to traders in all markets. We assume that this distribution is known to be normal, with mean \bar{P}_t (depending in a known way on the above history) and a constant variance σ^2 .

Second, we suppose that the actual price deviates from the (geometric) economy-wide average by an amount which is distributed independently of P_t . Specifically, let the percentage deviation of the price in z from the average P_t be denoted by z (so that markets are indexed by their price deviations from average) where z is normally distributed, independent of P_t , with mean zero and variance τ^2 . Then the observed price in z , $P_t(z)$ (in *logs*) is the sum of independent, normal variates

$$(4) \quad P_t(z) = P_t + z$$

The information $I_t(z)$ relevant for estimation of the unobserved (by suppliers in z at t) P_t , consists then of the observed price $P_t(z)$ and the history summarized in \bar{P}_t .

To utilize this information, suppliers use (4) to calculate the distribution of P_t , conditional on $P_t(z)$ and \bar{P}_t . This distribution is (by straightforward calculation) normal with mean:

$$(5) \quad \begin{aligned} E(P_t | I_t(z)) &= E(P_t | P_t(z), \bar{P}_t) \\ &= (1 - \theta)P_t(z) + \theta\bar{P}_t \end{aligned}$$

where $\theta = \tau^2 / (\sigma^2 + \tau^2)$, and variance $\theta\sigma^2$. Combining (1), (3), and (5) yields the supply function for market z :

$$(6) \quad \begin{aligned} y_t(z) &= y_{nt} + \theta\gamma[P_t(z) - \bar{P}_t] \\ &\quad + \lambda y_{c,t-1}(z) \end{aligned}$$

Averaging over markets (integrating with respect to the distribution of z) gives the aggregate supply function:

$$(7) \quad \begin{aligned} y_t &= y_{nt} + \theta\gamma(P_t - \bar{P}_t) \\ &\quad + \lambda[y_{t-1} - y_{n,t-1}] \end{aligned}$$

The *slope* of the aggregate supply function (7) thus varies with the fraction θ of total individual price variance, $\sigma^2 + \tau^2$, which is due to *relative* price variation.⁴ In cases where τ^2 is relatively small, so that individual price changes are virtually certain to reflect general price changes, the supply curve is nearly vertical. At the other extreme when general prices are stable (σ^2 is relatively small) the slope of the supply curve approaches the limiting value of γ .⁴

B. Completion and Solution of the Model

A central assumption in the development above is that supply behavior is based on the *correct* distribution of the unobserved current price level, P_t . To proceed, then, it is necessary to determine what this correct distribution is, a step which requires the completion of the model by inclusion of an aggregate demand side.

As suggested earlier, this will be done by postulating a demand function for goods of the form:

$$(8) \quad y_t + P_t = x_t$$

where x_t is an exogenous shift variable—equal to the observable *log* of nominal GNP. Further, let $\{\Delta x_t\}$ be a sequence of independent, normal variates with mean δ and variance σ_x^2 .⁵

⁴ This predicted relationship between a supply elasticity and the variance of a component of the price series is analogous to the link between the income elasticity of consumption demand and the variances of permanent and transitory income components which Friedman (1957) observes. As will be seen in Section II, it works in empirical testing in much the same way as well.

⁵ This particular characterization of the "shocks" to the economy is not central to the theory, but to discuss

The relevant history of the economy then consists (at most) of y_{nt} (which fixes calendar time), the demand shifts x_t, x_{t-1}, \dots , and past actual real outputs y_{t-1}, y_{t-2}, \dots . Since the model is linear in logs, it is reasonable to conjecture a price solution of the form:⁶

$$(9) \quad P_t = \pi_0 + \pi_1 x_t + \pi_2 x_{t-1} + \pi_3 x_{t-2} + \dots \\ + \eta_1 y_{t-1} + \eta_2 y_{t-2} + \dots + \xi_0 y_{nt}$$

Then \bar{P}_t will be the expectation of P_t , based on all information *except* x_t (the current demand level) or:

$$(10) \quad \bar{P}_t = \bar{P}_0 + \pi_1(x_{t-1} + \delta) + \pi_2 x_{t-1} \\ + \pi_3 x_{t-2} + \dots + \eta_1 y_{t-1} \\ + \eta_2 y_{t-2} + \dots + \xi_0 y_{nt}$$

To solve for the unknown parameters π_i, η_j and ξ_0 we first eliminate y_t between (7) and (8), or equate quantity demanded and supplied. Then inserting the right sides of (9) and (10) in place of P_t and \bar{P}_t , one obtains an identity in $\{x_t\}$, $\{y_t\}$, and y_{nt} , which is then used to obtain the parameter values. The resulting solutions for price and output are:⁷

$$P_t = \frac{\theta\gamma\delta}{1 + \theta\gamma} - \lambda\beta + \frac{1}{1 + \theta\gamma} x_t$$

rational expectations formation at all, *some* explicit stochastic description is clearly required. Independence is used here partly for simplicity, partly because it is empirically roughly accurate for most countries in the sample. The effect of autocorrelation in the shocks would, as can be easily traced out, be to add higher order lag terms to the solutions found below.

⁶ This solution method is adapted from Lucas (1972), which is in turn based on the ideas of John Muth.

⁷ If a demand function of the form $y_t = \xi P_t + x_t$ had been used, these solutions would assume the same form, with different expressions for the coefficients. If $\xi \neq 1$, however, x_t is an unobserved shock, unequal in general to observed nominal income. In this case, the model still predicts the time-series structure (moments and lagged moments) of the series y_{ct} and ΔP_t and is thus, in principle, testable. I have found empirical experimenting along these lines suggestive, but the series used are simply too short to yield results of any reliability.

$$+ \frac{\theta\gamma}{1 + \theta\gamma} x_{t-1} - \lambda y_{t-1} - (1 - \lambda) y_{nt}$$

$$y_t = - \frac{\theta\gamma\delta}{1 + \theta\gamma} + \lambda\beta + \frac{\theta\gamma}{1 + \theta\gamma} \Delta x_t$$

$$+ \lambda y_{t-1} + (1 - \lambda) y_{nt}$$

In terms of ΔP_t and y_{ct} , and letting $\pi = \theta\gamma/(1 + \theta\gamma)$, the solutions are:

$$(11) \quad y_{ct} = -\pi\delta + \pi\Delta x_t + \lambda y_{c,t-1}$$

$$(12) \quad \Delta P_t = -\beta + (1 - \pi)\Delta x_t + \pi\Delta x_{t-1} \\ - \lambda\Delta y_{c,t-1}$$

Let us review these solutions for internal consistency. Evidently, P_t is normally distributed about \bar{P}_t . The conditional variance of P_t will have the constant (as assumed) variance $1/(1 + \theta\gamma)^2 \sigma_x^2$. Thus those features of the behavior of prices which were assumed "known" by suppliers in subsection A are, in fact, true in this economy.

To review, equations (11) and (12) are the *equilibrium values* of the inflation rate and real output (as a percentage deviation from trend). They give the intersection points of an aggregate demand schedule, shifted by changes in x_t , and an aggregate supply schedule shifted by variables (lagged prices) which determine expectations. In order to avoid the introduction of an additional, spurious "expectations parameter," one cannot solve for this intersection on a period-by-period basis; accordingly, we have adopted a method which yields equilibrium "paths" of prices and output. Otherwise, the interpretation of (11) and (12) is entirely conventional.

Not surprisingly, the solution values of inflation and the cyclical component of real output are indicated by (11) and (12) to be distributed lags of current and past changes in nominal output. A change in the nominal expansion rate, Δx_t , has an immediate effect on real output, and lagged effects which decay geometrically. The

immediate effect on prices is one minus the real output effect, with the remainder of the impact coming in the succeeding period. We note in particular that this lag pattern may well produce periods of simultaneous inflation and below average real output. Though these periods arise because of supply shifts, the shifts result from lagged perception of demand changes, and *not* from autonomous changes in the cost structure of suppliers.

In addition to these features, the model does indeed assert the existence of a natural rate of output: the *average* rate of demand expansion, δ , appears in (11) with a coefficient equal in magnitude to the coefficient of the current rate, and with the opposite sign. Thus changes in the average rate of nominal income growth will have *no* effect on average real output. On the other hand, unanticipated demand shifts do have output effects, with magnitude given by the parameter π . Since this effect depends on "fooling" suppliers (in the sense of subsection A), one expects that π will be larger the smaller the variance of the demand shifts. We next develop this implication explicitly.

From the definition of π in terms of θ and γ , and the definition of θ in terms of σ^2 and τ^2 we have

$$\pi = \frac{\tau^2\gamma}{\sigma^2 + \tau^2(1 + \gamma)}$$

Combining with the expression for σ^2 obtained above, this gives

$$(13) \quad \pi = \frac{\tau^2\gamma}{(1 - \pi)^2\sigma_x^2 + \tau^2(1 + \gamma)}$$

For fixed τ^2 and γ , then, π takes the value $\gamma/(1+\gamma)$ at $\sigma_x^2=0$ and tends monotonically to zero as σ_x^2 tends to infinity.

The prediction that the average deviation of output from trend, $E(y_{ct})$, is invariant under demand policies is not, of course, subject to test: the deviations from

a *fitted* trend line must average to zero. Accordingly, we must base tests of the natural rate hypothesis (in this context) on (13): a relationship between an observable variance and a slope parameter.

II. Test Results

Testing the hypothesis advanced above involves two steps. First, within each country (11) and (12) should perform reasonably well. In particular, under the presumption that demand fluctuations are the major source of variation in ΔP_t and y_{ct} , the fits should be "good." The estimated values of π and λ should be between zero and one. Finally, since (11) and (12) involve five slope parameters but only two theoretical ones, the estimated π and λ values obtained from fitting (11) should work reasonably well in explaining variations in ΔP_t .

The main object of this study, however, is not to "explain" output and price level movements within a given country, but rather to see whether the terms of the output-inflation "tradeoff" vary across countries in the way predicted by the natural rate theory. For this purpose, we shall utilize the theoretical relationship (13) and the estimated values of π and σ_x^2 . Under the assumption that τ^2 and γ are relatively stable across countries, the estimated π values should decline as the sample variance of Δx_t increases.

Descriptive statistics for the eighteen countries in the sample are given in Table 1.⁸ As is evident, there is no association

⁸ The raw data on real and nominal GNP are from *Yearbook of National Accounts Statistics*, where series from many countries are collected and put on a uniform basis. The choice of countries is by no means random: the eighteen used are all the countries from which continuous series are available. The sample could thus be broadened considerably by use of sources from individual countries. To obtain the variables used in the tests, the logs of real and nominal output, y_t and x_t , are logs of the series in the source. The log of the price level, P_t , is the difference $x_t - y_t$; y_{ct} is the residual from the trend line $y_t = a + bt$, fit by least squares from the sample

TABLE 1—DESCRIPTIVE STATISTICS, 1952-67

Country	Mean Δy_t	Mean ΔP_t	Variance y_{ct}	Variance ΔP_t	Variance Δx_t
Argentina	.026	.220	.00096	.01998	.01555
Austria	.048	.038	.00104	.00113	.00124
Belgium	.034	.021	.00075	.00033	.00072
Canada	.043	.024	.00109	.00018	.00139
Denmark	.039	.041	.00082	.00038	.00084
West Germany	.056	.026	.00147	.00026	.00073
Guatemala	.046	.004	.00111	.00079	.00096
Honduras	.044	.012	.00042	.00084	.00109
Ireland	.025	.038	.00139	.00060	.00111
Italy	.053	.032	.00022	.00044	.00040
Netherlands	.047	.036	.00055	.00043	.00101
Norway	.038	.034	.00092	.00033	.00098
Paraguay	.054	.157	.00488	.03192	.03450
Puerto Rico	.058	.024	.00205	.00021	.00077
Sweden	.039	.036	.00030	.00043	.00041
United Kingdom	.028	.034	.00022	.00037	.00014
United States	.036	.019	.00105	.00007	.00064
Venezuela	.060	.016	.00175	.00068	.00127

between average real growth rates and average rates of inflation: this fact seems to be consistent with both the conventional and natural rate views of the tradeoff. Since our interest is in comparing real output and price behavior under different time patterns of nominal income, these statistics are somewhat disappointing. Essentially two types of nominal income behavior are observed: the highly volatile and expansive policies of Argentina and Paraguay, and the relatively smooth and moderately expansive policies of the remaining sixteen countries. But if the sample provides only two "points," they are indeed widely separated: the estimated variance of demand in the high inflation countries is on the order of 10 times that in the stable price countries.

The first three columns of Table 2 summarize the performance of equation (11) in accounting for movements in y_{ct} . The estimated values for π all lie between zero and one; with the exceptions of Argentina

period. The moments given in Table 1 are maximum likelihood estimates based on these series. The estimates reported in Table 2 are by ordinary least squares.

and Puerto Rico, so do the estimated λ values. The R^2 's indicate that for many, or perhaps most countries, important output-determining variables have been omitted from the model. The R^2 's for the inflation rate equation, (12), are given in column (4) of Table 2. In general, these tend to be lower than for equation (11), and not surprisingly the estimated coefficients from (12) (which are not shown) tend to behave erratically. Column (5) of Table 2 gives the fraction of the variance of ΔP_t explained by (12) when the coefficient estimates from (11) are imposed. (A "—" indicates a negative value.)⁹

With respect to its performance as an intracountry model of income and price determination, then, the system (11)-(12) passes the formal tests of significance. On the other hand, the goodness-of-fit statis-

⁹ The loss of explanatory power when these coefficients are imposed on (12) can be assessed formally by an approximate Chi-square test. By this measure, the loss is significant at the .05 level for Paraguay only. As Table 2 shows, however, this test is somewhat deceptive: for several countries the least squares estimates of (12) are so poor that there is little explanatory power to lose, and the test is "passed" vacuously.

TABLE 2—SUMMARY STATISTICS BY COUNTRY, 1953-67

Country	π	λ	R_y^2	$R_{\Delta P}^2$	R_ω^2
Argentina	.011 (.070)	-.126 (.258)	.018	.929	.914
Austria	.319 (.179)	.703 (.209)	.507	.518	—
Belgium	.502 (.100)	.741 (.093)	.875	.772	.661
Canada	.759 (.064)	.736 (.075)	.936	.418	—
Denmark	.571 (.118)	.679 (.110)	.812	.498	.282
West Germany	.820 (.136)	.784 (.110)	.881	.130	—
Guatemala	.674 (.301)	.695 (.274)	.356	.016	—
Honduras	.287 (.152)	.414 (.250)	.274	.521	.358
Ireland	.430 (.121)	.858 (.111)	.847	.499	.192
Italy	.622 (.134)	.042 (.183)	.746	.934	.914
Netherlands	.531 (.111)	.571 (.149)	.711	.627	.580
Norway	.530 (.088)	.841 (.096)	.893	.633	.427
Paraguay	.022 (.079)	.742 (.201)	.568	.941	.751
Puerto Rico	.689 (.121)	1.029 (.072)	.939	.419	—
Sweden	.287 (.166)	.584 (.186)	.525	.648	.405
United Kingdom	.665 (.290)	.178 (.209)	.394	.266	.115
United States	.910 (.086)	.887 (.070)	.945	.571	.464
Venezuela	.514 (.183)	.937 (.148)	.755	.425	—

tics are generally considerably poorer than we have come to expect from annual time-series models.

In contrast to these somewhat mixed results, the behavior of the estimated π values across countries is in striking conformity with the natural rate hypothesis. For the sixteen stable price countries, $\hat{\pi}$ ranges from .287 to .910; for the two volatile price countries, this estimate is smaller by a factor of 10! To illustrate this order-of-magnitude effect more sharply, let us examine the complete results for two countries: the United States and Argen-

tina. For the United States, the fitted versions of (11) and (12) are:

$$y_{ct} = - .049 + (.910)\Delta x_t + (.887)y_{c,t-1}$$

$$\begin{aligned} \Delta P_t = & - .028 + (.119)\Delta x_t + (.758)\Delta x_{t-1} \\ & - (.637)\Delta y_{c,t-1} \end{aligned}$$

The comparable results for Argentina are:

$$y_{ct} = - .006 + (.011)\Delta x_t - (.126)y_{c,t-1}$$

$$\begin{aligned} \Delta P_t = & - .047 + (1.140)\Delta x_t - (.083)\Delta x_{t-1} \\ & + (.102)\Delta y_{c,t-1} \end{aligned}$$

In a stable price country like the United States, then, policies which increase nomi-

nal income tend to have a large initial effect on real output, together with a small, positive initial effect on the rate of inflation. Thus the apparent short-term tradeoff is favorable, as long as it remains unused. In contrast, in a volatile price country like Argentina, nominal income changes are associated with equal, contemporaneous price movements with no discernible effect on real output. These results are, of course, inconsistent with the existence of even moderately stable Phillips curves. On the other hand, they follow directly from the view that inflation stimulates real output if, and only if, it succeeds in "fooling" suppliers of labor and goods into thinking *relative* prices are moving in their favor.

III. Concluding Remarks

The basic idea underlying the tests reported above is extremely simple, yet I am afraid it may have become obscured by the rather special model in which it is embodied. In this section, I shall try to restate this idea in a way which, though not quite accurate enough to form the basis for econometric work, conveys its essential feature more directly.

The propositions to be compared empirically refer to the effects of aggregate demand policies which tend to move inflation rates and output (relative to trend) in the same direction, or alternatively, unemployment and inflation in opposite directions. The conventional Phillips curve account of this observed co-movement says that the terms of the tradeoff arise from relatively stable structural features of the economy, and are thus independent of the nature of the aggregate demand policy pursued. The alternative explanation of the same observed tradeoff is that the positive association of price changes and output arises because suppliers misinterpret general price movements for relative price changes. It follows from this view,

first, that changes in average inflation rates will not increase average output, and secondly, that the higher the *variance* in average prices, the less "favorable" will be the observed tradeoff.

The most natural cross-national comparison of these propositions would seem to be a direct examination of the association of average inflation rates and average output, relative to "normal" or "full employment." Unfortunately, there seems to be no satisfactory way to measure normal output. The deviation-from-fitted-trend method I have used *defines* normal output to be average output. The use of unemployment series suffers from the same difficulty, since one must somehow select the (obviously positive) rate to be denoted full employment.

Thus although the issue revolves around the relation between *means* of inflation and output rates, it cannot be resolved by examination of sample averages. Fortunately, the existence of a stable tradeoff also implies a relationship between *variances* of inflation and output rates, as illustrated in Figure 1. With a stable tradeoff, policies which lead to wide variation in prices must also induce comparable variation in real output. If these sample variances do not tend to move together (and, as Table 1 shows, they do not) one

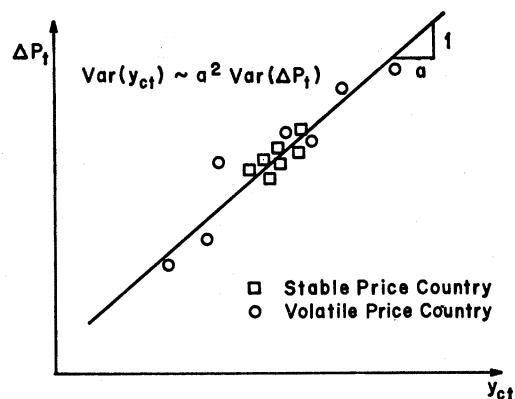


FIGURE 1

can only conclude that the tradeoff tends to fade away the more frequently it is used, or abused.

This simple argument leads to a formal test if the output-inflation association is entirely contemporaneous. In fact, however, it involves lagged effects which make a direct comparison of variances, as just suggested, difficult in short time-series. Accordingly, it has been necessary to impose a specific, simple structure on the data. As we have seen, this structure accounts for output and inflation rate movements only moderately well, but well enough to capture the main phenomenon predicted by the natural rate theory: the higher the variance of demand, the more unfavorable are the terms of the Phillips tradeoff.

REFERENCES

- M. Friedman, *A Theory of the Consumption Function*, Princeton 1957.
- _____, "The Role of Monetary Policy," *Amer. Econ. Rev.*, Mar. 1968, 58, 1-17.
- D. F. Gordon and A. Hynes, "On the Theory of Price Dynamics," in E. S. Phelps et al., *Micro-economics of Inflation and Employment Theory*, New York 1969.
- R. E. Lucas, Jr., "Expectations and the Neutrality of Money," *J. Econ. Theor.*, Apr. 1972, 4, 103-24.
- _____, "Econometric Testing of the Natural Rate Hypothesis," *Conference on the Econometrics of Price Determination*, Washington 1972, 50-59.
- _____, and L. A. Rapping, "Real Wages, Employment and the Price Level," *J. Polit. Econ.*, Sept./Oct. 1969, 77, 721-54.
- _____, and _____, "Unemployment in the Great Depression: Is There a Full Explanation?", *J. Polit. Econ.*, Jan./Feb. 1972, 80, 186-91.
- J. F. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, 29, 315-35.
- E. S. Phelps, introductory chapter in E. S. Phelps et al., *Micro-economics of Inflation and Employment Theory*, New York 1969.
- F. Raines, "Macroeconomic Demand and Supply: an Integrative Approach," Washington Univ. working paper, Apr. 1971.
- A. Rees, "On Equilibrium in Labor Markets," *J. Polit. Econ.*, Mar./Apr. 1970, 78, 306-10.
- P. Taubman and M. Wilkinson, "User Cost, Capital Utilization and Investment Theory," *Int. Econ. Rev.*, June 1970, 11, 209-15.
- United Nations, Department of Economic and Social Affairs, United Nations Statistical Office, *Yearbook of National Accounts Statistics*, 66 and 68, New York 1958.

The Political Economy of the Rent-Seeking Society

By ANNE O. KRUEGER*

In many market-oriented economies, government restrictions upon economic activity are pervasive facts of life. These restrictions give rise to rents of a variety of forms, and people often compete for the rents. Sometimes, such competition is perfectly legal. In other instances, rent seeking takes other forms, such as bribery, corruption, smuggling, and black markets.

It is the purpose of this paper to show some of the ways in which rent seeking is competitive, and to develop a simple model of competitive rent seeking for the important case when rents originate from quantitative restrictions upon international trade. In such a case 1) competitive rent seeking leads to the operation of the economy inside its transformation curve; 2) the welfare loss associated with quantitative restrictions is unequivocally greater than the loss from the tariff equivalent of those quantitative restrictions; and 3) competitive rent seeking results in a divergence between the private and social costs of certain activities. Although the analysis is general, the model has particular applicability for developing countries, where government interventions are frequently all-embracing.

A preliminary section of the paper is concerned with the competitive nature of rent seeking and the quantitative importance of rents for two countries, India and Turkey. In the second section, a formal model of rent seeking under quantitative

restrictions on trade is developed and the propositions indicated above are established. A final section outlines some other forms of rent seeking and suggests some implications of the analysis.

I. Competitive Rent Seeking

A. Means of Competition

When quantitative restrictions are imposed upon and effectively constrain imports, an import license is a valuable commodity. It is well known that under some circumstances, one can estimate the tariff equivalents of a set of quantitative restrictions and analyze the effects of those restrictions in the same manner as one would the tariff equivalents. In other circumstances, the resource-allocational effects of import licensing will vary, depending upon who receives the license.¹

It has always been recognized that there are *some* costs associated with licensing: paperwork, the time spent by entrepreneurs in obtaining their licenses, the cost of the administrative apparatus necessary to issue licenses, and so on. Here, the argument is carried one step further: in many circumstances resources are devoted to competing for those licenses.

The consequences of that rent seeking are examined below. First, however, it will be argued that rent-seeking activities are often competitive and resources are devoted to competing for rents. It is difficult, if not impossible, to find empirically observable measures of the degree to which rent seeking is competitive. Instead, some

* Professor of economics, University of Minnesota. I am indebted to James M. Henderson for invaluable advice and discussion on successive drafts. Jagdish Bhagwati and John C. Hause made helpful comments on earlier drafts of this paper.

¹ This phenomenon is explored in detail in Bhagwati and Krueger.

mechanisms under which rent seeking is almost certain to be competitive are examined. Then other cases are considered in which it is less obvious, but perhaps equally plausible, that competition results.

Consider first the results of an import-licensing mechanism when licenses for imports of intermediate goods are allocated in proportion to firms' capacities. That system is frequently used, and has been analyzed for the Indian case by Jagdish Bhagwati and Padma Desai. When licenses are allocated in proportion to firms' capacities, investment in additional physical plant confers upon the investor a higher expected receipt of import licenses. Even with initial excess capacity (due to quantitative restrictions upon imports of intermediate goods), a rational entrepreneur may still expand his plant if the expected gains from the additional import licenses he will receive, divided by the cost of the investment, equal the returns on investment in other activities.² This behavior could be perfectly rational even if, for all entrepreneurs, the total number of import licenses will remain fixed. In fact, if imports are held constant as domestic income grows, one would expect the domestic value of a constant quantity of imports to increase over time, and hence installed capacity would increase while output remained constant. By investing in additional capacity, entrepreneurs devote resources to compete for import licenses.

A second sort of licensing mechanism frequently found in developing countries is used for imports of consumer goods. There, licenses are allocated *pro rata* in proportion to the applications for those licenses from importers-wholesalers. Entry

² Note that: 1) one would expect to find greater excess capacity in those industries where rents are higher; and 2) within an industry, more efficient firms will have greater excess capacity than less efficient firms, since the return on a given amount of investment will be higher with greater efficiency.

is generally free into importing-wholesaling, and firms usually have U-shaped cost curves. The result is a larger-than-optimal number of firms, operating on the downward sloping portion of their cost curves, yet earning a "normal" rate of return. Each importer-wholesaler receives fewer imports than he would buy at existing prices in the absence of licensing, but realizes a sufficient return on those licenses he does receive to make it profitable to stay in business. In this case, competition for rents occurs through entry into the industry with smaller-than-optimally sized firms, and resources are used in that the same volume of imports could be efficiently distributed with fewer inputs if firms were of optimal size.

A third sort of licensing mechanism is less systematic in that government officials decide on license allocations. Competition occurs to some extent through both mechanisms already mentioned as businessmen base their decisions on expected values. But, in addition, competition can also occur through allocating resources to influencing the probability, or expected size, of license allocations. Some means of influencing the expected allocation—trips to the capital city, locating the firm in the capital, and so on—are straightforward. Others, including bribery, hiring relatives of officials or employing the officials themselves upon retirement, are less so. In the former case, competition occurs through choice of location, expenditure of resources upon travel, and so on. In the latter case, government officials themselves receive part of the rents.

Bribery has often been treated as a transfer payment. However, there is competition for government jobs and it is reasonable to believe that expected total remuneration is the relevant decision variable for persons deciding upon careers. Generally, entry into government service requires above-average educational at-

tainments. The human capital literature provides evidence that choices as to how much to invest in human capital are strongly influenced by rates of return upon the investment. For a given level of educational attainment, one would expect the rate of return to be approximately equated among various lines of endeavor. Thus, if there appear to be high official-plus-unofficial incomes accruing to government officials and higher education is a prerequisite for seeking a government job, more individuals will invest in higher education. It is not necessary that government officials earn the same total income as other college graduates. All that is necessary is that there is an excess supply of persons seeking government employment, or that highly educated persons make sustained efforts to enter government services. Competition takes place through attaining the appropriate credentials for entry into government service and through accepting unemployment while making efforts to obtain appointments. Efforts to influence those in charge of making appointments, of course, just carry the argument one step further back.

To argue that competition for entry into government service is, in part, a competition for rents does not imply that all government servants accept bribes nor that they would leave government service in their absence. Successful competitors for government jobs might experience large windfall gains even at their official salaries. However, if the possibility of those gains induces others to expend time, energy, and resources in seeking entry into government services, the activity is competitive for present purposes.

In all these license-allocation cases, there are means, legal and illegal, for competing for rents. If individuals choose their activities on the basis of expected returns, rates of return on alternative activities will be equated and, in that sense, markets

will be competitive.³ In most cases, people do not perceive themselves to be rent seekers and, generally speaking, individuals and firms do not specialize in rent seeking. Rather, rent seeking is one part of an economic activity, such as distribution or production, and part of the firm's resources are devoted to the activity (including, of course, the hiring of expediters). The fact that rent seeking and other economic activities are not generally conducted by separate economic entities provides the motivation for the form of the model developed below.

B. Are Rents Quantitatively Important?

Granted that rent seeking may be highly competitive, the question remains whether rents are important. Data from two countries, India and Turkey, suggest that they are. Gunnar Myrdal believes India may "... on the balance, be judged to have somewhat less corruption than any other country in South Asia" (p. 943). Nonetheless, it is generally believed that "corruption" has been increasing, and that much of the blame lies with the proliferation of economic controls following independence.⁴

Table 1 presents crude estimates, based on fairly conservative assumptions of the value of rents of all sorts in 1964. One important source of rents—investment licensing—is not included for lack of any valid basis on which to estimate its value. Many smaller controls are also excluded. Nonetheless, it is apparent from Table 1 that

³ It may be objected that illegal means of competition may be sufficiently distasteful that perfect competition will not result. Three comments are called for. First, it requires only that enough people at the margin do not incur disutility from engaging in these activities. Second, most lines of economic activity in many countries cannot be entered without some rent-seeking activity. Third, risks of detection (especially when bribery is expected) and the value judgments associated with illegal activities differ from society to society. See Ronald Wraith and Edgar Simpkins.

⁴ Santhanam Committee, pp. 7-8.

TABLE 1—ESTIMATES OF VALUE OF RENTS: INDIA, 1964

Source of Rent	Amount of Rent (Rs. million)
Public investment	365
Imports	10,271
Controlled commodities	3,000
Credit rationing	407
Railways	602
Total	14,645

Sources:

1) Public investment: The Santhanam Committee, pp. 11-12, placed the loss in public investment at *at least* 5 percent of investment. That figure was multiplied by the average annual public investment in the *Third Five Year Plan*.

2) Imports: The Santhanam Committee, p. 18, stated that import licenses were worth 100 to 500 percent of their face value. Seventy-five percent of the value of 1964 imports was used here as a conservative estimate.

3) Controlled commodities: These commodities include steel, cement, coal, passenger cars, scooters, food, and other price—and/or distribution-controlled commodities, as well as foreign exchange used for illegal imports and other unrecorded transactions. The figure is the lower bound estimate given by John Monteiro, p. 60. Monteiro puts the upper bound estimate at Rs. 30,000 billion, although he rejects the figure on the (dubious) ground that notes in circulation are less than that sum.

4) Credit rationing: The bank rate in 1964 was 6 percent; Rs. 20.3 billion of loans were outstanding. It is assumed that *at least* an 8 percent interest rate would have been required to clear the market, and that 3 percent of bank loans outstanding would be equivalent to the present value of new loans at 5 percent. Data source: Reserve Bank of India, Tables 534 and 554.

5) Railways: Monteiro, p. 45, cites commissions of 20 percent on railway purchases, and extra-official fees of Rs. 0.15 per wagon and Rs. 1.4 per 100 maunds loaded. These figures were multiplied by the 1964 traffic volume; 203 million tons of revenue-paying traffic originated in that year. Third plan expenditure on railroads was Rs. 13,260 million. There were 350,000 railroad goods wagons in 1964-65. If a wagon was loaded once a week, there were 17,500,000 wagons of freight. At Rs. 0.15 per load, this would be Rs. 2.6 million; 100 maunds equal 8,228 pounds so at 1.4 Rs. per 100 maunds, Rs. 69 million changed hands; if one-fifth of railroad expenditures were made in 1964-65, Rs. 2652 million was spent in 1964; at 20 percent, this would be Rs. 530 million, for a total of Rs. 602 million.

import licenses provided the largest source of rents. The total value of rents of Rs. 14.6 billion contrasts with Indian national

income of Rs. 201 billion in 1964. At 7.3 percent of national income, rents must be judged large relative to India's problems in attempting to raise her savings rate.

For Turkey, excellent detailed estimates of the value of import licenses in 1968 are available.⁵ Data on the c.i.f. prices of individual imports, their landed cost (c.i.f. price plus all duties, taxes, and landing charges), and wholesale prices were collected for a sizeable sample of commodities representing about 10 percent of total imports in 1968. The c.i.f. value of imports in the sample was TL 547 million and the landed cost of the imports was TL 1,443 million. The value at the wholesale level of these same imports was TL 3,568 million. Of course, wholesalers incur some handling, storage, and transport costs. The question, therefore, is the amount that can be attributed to normal wholesaling costs. If one assumes that a 50 percent markup would be adequate, then the value of import licenses was TL 1,404 million, or almost three times the c.i.f. value of imports. Imports in 1968 were recorded (c.i.f.) as 6 percent of national income. On the basis of Aker's data, this would imply that rents from import licenses in Turkey in 1968 were about 15 percent of GNP.

Both the Indian and the Turkish estimates are necessarily somewhat rough. But they clearly indicate that the value of import licenses to the recipients was sizeable. Since means were available of competing for the licenses, it would be surprising if competition did not occur for prizes that large. We turn, therefore, to an examination of the consequences of competitive rent seeking.

⁵ I am indebted to Ahmet Aker of Robert College who kindly made his data available to me. Details and a description of the data can be found in my forthcoming book.

II. The Effects of Competitive Rent Seeking

The major proposition of this paper is that competitive rent seeking for import licenses entails a welfare cost in addition to the welfare cost that would be incurred if the same level of imports were achieved through tariffs. The effects of tariffs upon production, trade, and welfare are well known, and attention is focussed here upon the additional cost of competitive rent seeking. A simple model is used to develop the argument. Initially, free trade is assumed. Then, a tariff or equivalent import restriction is introduced. Finally, an equal import restriction with competitive rent seeking is examined.

A. The Basic Model

Two commodities are consumed by the country under investigation: food and consumption goods. Food is produced domestically and exported. Consumption goods are imported. Distribution is a productive activity whereby food is purchased from the agricultural sector, exported, and the proceeds are used to import consumption goods which are sold in the domestic market. Labor is assumed to be the only domestic factor of production.⁶ It is assumed that the country under consideration is small and cannot affect its international terms of trade. Physical units are selected so that the fixed international prices of both goods are unity.

The agricultural production function is

$$(1) \quad A = A(L_A) \quad A' > 0, \quad A'' < 0$$

where A is the output of food and L_A is the quantity of labor employed in agriculture. The sign of the second derivative reflects a diminishing marginal physical

⁶ Labor could be regarded as a composite domestic factor of production. Extensions to two or more factors would complicate the analysis, but would not alter its basic results.

product of labor in agriculture, due, presumably, to fixity in the supply of land.

The level of distribution output, D , is defined to equal the level of consumption-goods imports, M :

$$(2) \quad D = M$$

One unit of distributive services entails exchanging one unit of imports for food with the agricultural sector at the domestic terms of trade, and exporting the food in exchange for imports at the international terms of trade. Constant returns to scale are assumed for the distribution activity; one unit of distribution requires k units of labor. Total labor employed in distribution, L_D , is

$$(3) \quad L_D = kD$$

A distribution charge of p_D per unit is added to the international price of imports:

$$(4) \quad p_M = 1 + p_D$$

where p_M is the domestic price of imports. The domestic price of food is assumed to equal its unit international price.⁷

Society's demand for imports depends upon the domestic price of imports and total income generated in agriculture:⁸

$$(5) \quad M = M(p_M, A)$$

where $\partial M / \partial p_M < 0$ and $\partial M / \partial A > 0$. Demand decreases with increases in the price of imports, and increases with increases in agricultural output (income). Equation (5) is derived from micro utility maximization with the assumption that farmers, distributors, and rent seekers all have the same consumption behavior. Domestic

⁷ These assumptions establish a domestic numeraire. The real analysis would be unaffected by proportional changes in the domestic prices.

⁸ Food and imports are consumed. But, by choice of food as the numeraire (see equation (6)) and the assumed constancy of international prices, agricultural output serves as a measure of income.

food consumption, F , is simply the quantity not exported:

$$(6) \quad F = A - M$$

Since the fixed international terms of trade equal unity, food exports equal consumption goods imports.

Finally, it is assumed that the economy under consideration has a fixed labor supply, \bar{L} :

$$(7) \quad \bar{L} = L_A + L_D + L_R$$

where L_R is the quantity of labor engaged in rent seeking.

B. Free Trade

Under free trade, there is free entry into both agriculture and distribution and competition equates the wage in the two activities:

$$(8) \quad A' = p_D/k$$

Equations (1) to (8) constitute the free-trade system. These eight equations contain the eight variables A , M , D , F , L_A , L_D , p_M , and p_D . Since there is no rent seeking under free trade, $L_R \equiv 0$.

It is easily established that free trade is optimal in the sense that the domestic price ratio under free trade equals the marginal rate of transformation between food consumption and imports. The consumption possibility locus is obtained by substituting into (6) from (1) and (7)

$$F = A(\bar{L} - kM) - M$$

The locus has a marginal rate of transformation greater than one:

$$(9) \quad \frac{-dF}{dM} = kA' + 1 > 1$$

which reflects the positive distribution cost of substituting imports for food consumption. The locus is concave:

$$\frac{d^2F}{dM^2} = k^2A'' < 0$$

since $A'' < 0$, which follows from diminishing returns in food production. Substituting from (8) into (9),

$$\frac{-dF}{dM} = 1 + p_D$$

which establishes the aforementioned equality.

A free-trade solution is depicted in Figure 1. Domestic food consumption and import consumption are measured along OF and OM , respectively. The consumption possibility locus is $\hat{F}\hat{M}$. At the point \hat{F} no imports are consumed and hence there is no distribution. If distribution were costless, society could choose its consumption point from the line $\hat{F}A$. However, to consume one unit of import requires exchanging one unit of food and withdrawing k workers from agriculture to provide the requisite distributive services. With diminishing marginal product of labor in agriculture, the cost of additional imports in terms of foregone food production rises. Thus, the price of distribution, and hence the domestic price of imports, increases in moving northwest from \hat{F} . The consump-

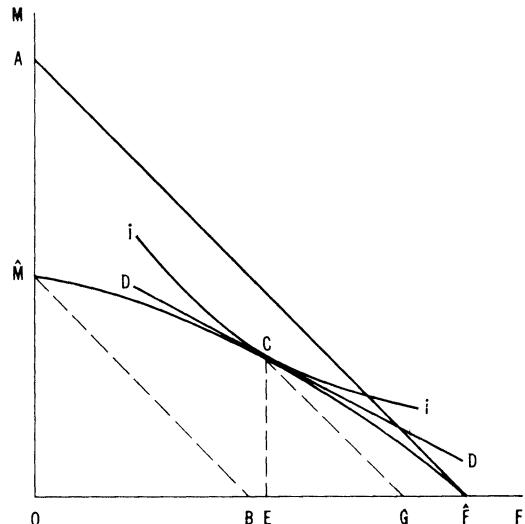


FIGURE 1. FREE TRADE

tion point \hat{M} has OB food exchanged for $O\hat{M}$ of imports. The distance $\hat{F}B$ is the agricultural output foregone to distribute $O\hat{M}$ imports.

If society's preferences are given by the indifference curve ii' , point C is optimal. The price of distribution is reflected in the difference between the slope of $\hat{F}A$ and the slope of DD' at C . At the point C , OG food would be produced, with EG ($= EC$) exported, and the rest domestically consumed.

C. A Tariff or an Import Restriction Without Rent Seeking

Consider now a case in which there is a restriction upon the quantity of imports

$$(10) \quad M = \bar{M}$$

where \bar{M} is less than the import quantity that would be realized under free trade. Since entry into distribution is now limited, the competitive wage equality (8) will no longer hold. The relevant system contains (1) to (7) and (10). The variables are the same as in the free-trade case and again $L_R=0$. The system may be solved sequentially: given (10), D follows from (2), L_D from (3), L_A from (7), A from (1), F from (6), p_M from (5), and p_D from (4). Since equations (1), (6), and (7) remain intact, the solution for this case is also on the consumption possibility locus.

It is useful to establish the directions of change for the variables following a switch from free trade to import restriction. The reduced import level will reduce the labor employed in distribution and increase the labor force in agriculture. Diminishing returns will reduce the agricultural wage. The domestic price of imports, the distributive margin, and the wage of distributors will increase. Distributors will earn a rent in the sense that their wage will exceed the wage of those engaged in agriculture.

In the absence of rent seeking, a tariff

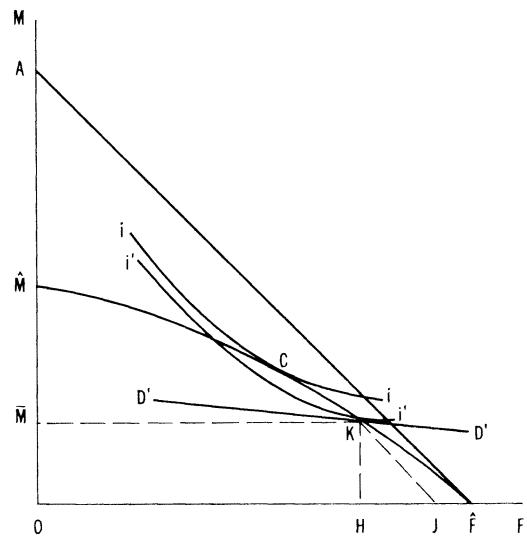


FIGURE 2. IMPORT RESTRICTION
WITHOUT RENT SEEKING

and a quantitative restriction are equivalent⁹ aside from the resultant income distribution. Under a quantitative restriction the distributive wage is higher than the agricultural. If instead there were an equivalent tariff with redistribution of the proceeds, the marginal product of labor in agriculture would be unchanged, but agricultural workers would benefit by the amount of tariff proceeds redistributed to them whereas traders' income would be lower. Since the allocation of labor under a tariff and quantitative restriction without rent seeking is the same and domestic prices are the same, the only difference between the two situations lies in income distribution.

The solution under a quantitative restriction is illustrated in Figure 2, where $\hat{F}\hat{M}$ is again the consumption possibility locus and C the free-trade solution. With a quantitative restriction on imports in the amount $O\bar{M}$, the domestic prices of

⁹ The change in the price of the import from the free-trade solution is the tariff equivalent of the quantitative restriction described here.

imports, and hence of distribution, rise from free trade to import restriction. Food output (OJ) and domestic consumption of food increase, and exports decline to HJ ($=OM$). The indifference curve $i'i'$ lies below ii (and the point C), and the welfare loss may be described by the consumption and production cost measure given by Harry Johnson.

The wage rate in distribution unequivocally rises for a movement from free trade to a quantitative restriction. The total income of distributors will increase, decrease, or remain unchanged depending upon whether the proportionate increase in p_D is greater than, less than, or equal to the absolute value of the proportionate decrease of imports. For the moment, let p_D , p_M , and M represent free-trade solution values, and let p_D^* , p_M^* , and \bar{M} represent import-restriction solution values. The total arc elasticity of demand for imports for the interval under consideration, η , is

$$(11) \quad \eta = \frac{-(\bar{M} - M)}{\bar{M} + M} \cdot \frac{p_M^* + p_M}{p_M^* - p_M}$$

Total expenditures on imports will increase, decrease, or remain unchanged as η is less than one, greater than one, or equal to one. The total income of distributors will increase if

$$p_D^* \bar{M} > p_D M$$

Multiplying both sides of this inequality by $(p_M^* + p_M)/(p_M^* - p_M)$, substituting from (11), and using (4),

$$(12) \quad 1 + 2/(p_D^* + p_D) > \eta$$

Hence, distributors' total income can increase even if the demand for imports is price elastic.¹⁰ The smaller is the free-trade

¹⁰ Proof of (12) uses the step that $p_D^* \bar{M} > p_D M$ implies $(p_D^* - p_D)/(p_D^* + p_D) > -(\bar{M} - M)/(\bar{M} + M)$. Note that in the continuous case, (12) reduces to $1 + 1/p_D > \eta$.

distributive markup, the more likely it is that the distributors' total income will increase with a curtailment of imports. The reason is that an increase in the domestic price of imports results in a proportionately greater increase in the price of distribution.

D. An Import Restriction with Competitive Rent Seeking

In the import-restriction model just presented, the wage in distribution p_D/k exceeds the wage in agriculture A' . Under this circumstance, it would be surprising if people did not endeavor to enter distribution in response to its higher return. Resources can be devoted to rent seeking in all the ways indicated in Section IA. This rent-seeking activity can be specified in a number of different ways. A simple and intuitively plausible specification is that people will seek distributive rents until the average wage in distribution and rent seeking equals the agricultural wage:¹¹

$$(13) \quad A' = \frac{p_D \bar{M}}{L_D + L_R}$$

One can regard all distributors and rent seekers as being partially engaged in each activity or one can think of rent seekers as entering in the expectation of receiving import licenses. In the latter case, the final solution classifies the successful seekers in L_D and the unsuccessful ones in L_R . Equation (13) implies risk neutrality in this circumstance.

The model for import restriction with rent seeking contains the same equations,

¹¹ As an alternative, the distributive production function (3) can be altered to treat all persons competing for import licenses as distributors so that L_D also encompasses L_R and $A' = p_D \bar{M}/L_D$. Another alternative is to introduce a rent-seeking activity distinct from distribution with a wage determined from total rents $(p_D - A'k)\bar{M}/L_R$, and require that this wage equal the wages in distribution and agriculture. These specifications give results equivalent to those that follow from (13).

(1) to (7) and (10), and the same variables as the model for import restrictions without rent seeking. In addition, the new model contains (13) and the introduction of L_R as a variable. The essential factor of rent seeking is that L_R becomes positive.

Let us start with a solution for an import restriction without rent seeking and ask what happens to the values of the variables when rent seeking is introduced. By assumption $M = \bar{M}$ is unchanged, so that L_D is unchanged. Therefore, $dL_A = -dL_R$, because the labor that enters rent seeking can only come from agriculture. Substituting into the total differential of (1) and using (6),

$$(14) \quad dF = dA = -A'dL_R < 0$$

Agricultural production and food consumption are reduced by the introduction of rent seeking. Since the import level remains unchanged, rent seeking entails a welfare loss beyond that for an import restriction without rent seeking. The concavity of the agricultural production function results in a food loss that is less than proportional to decrements in L_A . Differentiating (5) totally,

$$(15) \quad 0 = M_1 dp_M + M_2 dA$$

where M_1 and M_2 are the partial derivatives of (5) with respect to p_M and A , respectively. Solving (15) for dp_M , and substituting from (4) and (14),

$$(16) \quad dp_D = dp_M = \frac{M_2}{M_1} A'dL_R < 0$$

since $M_1 < 0$ and $M_2 > 0$. The domestic cost of imports will be lower under rent-seeking competition. This follows from the decrease in the consumption of food relative to imports.

The results of (14) and (16) are not dependent upon the particular form of the equilibrium of the labor market. They hold for any specification of competitive

rent seeking. Equation (13) serves to determine particular values for L_R and other variables of the system. The mere existence of competitive rent seeking is enough to determine the directions of change of the variables.

The above results are sufficient to indicate that, for any given level of import restrictions, competition among rent seekers is clearly inferior to the tariff equivalent of the restrictions, in that there could be more food consumed with no fewer imports under the latter case than the former. To the extent that rent seeking is competitive, the welfare cost of import restrictions is equal to the welfare cost of the tariff equivalent *plus the additional cost of rent-seeking activities*. Measurement of that excess cost is considered below.

The tariff-equivalent and rent-seeking equilibria are contrasted in Figure 3. Equilibrium under rent seeking will be at some point such as L , with the same consumption of imports, but smaller production and consumption of food than occurs under a tariff. The points K and C are the tariff-equivalent and free-trade equilibria, respectively. The line $D'D'$ cor-

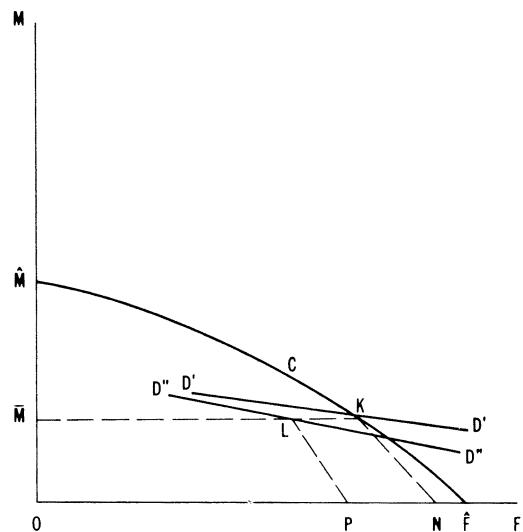


FIGURE 3. RENT-SEEKING IMPORT RESTRICTION

responds to the domestic price of imports in Figure 2, and the steeper line $D''D''$ corresponds to the lower domestic price of imports under competitive rent seeking.

So far, it has been shown that for any given level of import restriction, a tariff is Pareto-superior to competitive rent seeking, and the properties of rent-seeking equilibrium have been contrasted with those of the tariff-equivalent case in the absence of competition for the rents. A natural question is whether anything can be said about the properties of rent-seeking equilibrium in contrast to those of a free-trade equilibrium, which is, after all, the optimal solution. It has been seen that the number of persons engaged in distribution declines from free trade to import restriction without rent seeking, and increases as one goes from that situation to competition for import licenses. Likewise, agricultural output increases between free trade and the tariff-equivalent case, and declines between that and rent seeking. The question is whether any unambiguous signs can be placed on the direction of these changes between free trade and rent seeking and, in particular, is it possible that society might produce and consume less of both goods under rent seeking than under free trade?

The answer is that if inequality (12) is satisfied, the absolute number of persons ($L_D + L_R$) in distribution will increase going from a free-trade to a rent-seeking equilibrium. If import demand is more elastic, the number of persons in distribution will decline. Contrasted with a free-trade equilibrium, there would be less agricultural output and fewer imports when inequality (12) holds. If, with import restriction, the income from distribution $p_D^* \bar{M}$ is greater than distributors' income at free trade, more persons will be employed in distribution-cum-rent seeking with import restriction than are employed under free trade.

E. Measuring the Welfare Loss from Rent Seeking

A tariff has both production and consumption costs, and it has already been shown that rent seeking entails costs in addition to those of a tariff. Many forms of competition for rents, however, are by their nature difficult to observe and quantify and one might therefore question the empirical content of the result so far obtained.

Fortunately, there is a way to estimate the production cost of rent seeking. That cost, in fact, is equal to the value of the rents. This can be shown as follows. The rent per import license, r , is:

$$(17) \quad r = p_D - kA'$$

This follows because the labor required to distribute one unit of imports is k , which could be used in agriculture with a return A' . Note that at free trade r equals zero. A distributor could efficiently distribute an import and earn his opportunity cost in agriculture with zero rent. The total value of rents, R , with competitive rent seeking is thus the rent per unit of imports times the amount imported.

$$(18) \quad R = r\bar{M} = (p_D - kA')\bar{M}$$

Using (3) and (13),

$$(19) \quad \begin{aligned} R &= \left(p_D - \frac{k p_D \bar{M}}{L_D + L_R} \right) \bar{M} \\ &= p_D \left(1 - \frac{L_D}{L_D + L_R} \right) \bar{M} \\ &= \frac{p_D \bar{M} L_R}{L_D + L_R} \end{aligned}$$

Thus the total value of rents reflects the agricultural wage (A') times the number of rent seekers.

The value of rents reflects the value (at current prices) of the domestic factors of production which could be extracted from the economy with no change in the final

goods and services available for society's utilization. Thus, if the value of rents is known, it indicates the volume of resources that could be transferred out of distribution and into other activities, with no loss of distributive services from an initial position of rent-seeking activity. The estimates of rents in India and Turkey, therefore, may be interpreted as the deadweight loss from quantitative restrictions in addition to the welfare cost of their associated tariff equivalents if one believes that there is competition for the rents.

The value of the rents overstates the increase in food output and consumption that could be attained with a tariff to the extent that the marginal product of labor in agriculture is diminishing, since the equilibrium wage will rise between the tariff and the competitive rent-seeking situation. In the case of a constant marginal product of labor in alternative uses, the value of rents will exactly measure foregone output.

F. The Implications of Rent Seeking for Trade Theory

Recognition of the fact of rent seeking alters a variety of conclusions normally obtained in the trade literature and examination of such cases is well beyond the scope of this paper. A few immediately derivable results are worth brief mention, however.

First, an import prohibition might be preferable to a nonprohibitive quota if there is competition for licenses under the quota. This follows immediately from the fact that a prohibition would release resources from rent seeking and the excess cost of domestic production might be less than the value of the rents. Second, one could not, in general, rank the tariff-equivalents of two (or more) quotas, since the value of rents is a function of both the amount of rent per unit (the tariff equiva-

lent) and the volume of imports of each item.¹² Third, it has generally been accepted that the more inelastic domestic demand the less is likely to be the welfare cost of a given tariff. For the quota-cum-rents case, the opposite is true: the more price inelastic is demand, the greater will be the value of rents and the greater, therefore, the deadweight loss associated with rent seeking. Fourth, it is usually believed that competition among importers will result in a better allocation of resources than will a monopoly. If rent seeking is a possibility, however, creating a monopoly position for one importer will generally result in a higher real income if not in a preferable income distribution for society. Finally, devaluation under quantitative restrictions may have important allocation effects because it diminishes the value of import licenses, and hence the amount of rent-seeking activity, in addition to its effects upon exports.

III. Conclusions and Implications

In this paper, focus has been on the effects of competition for import licenses under a quantitative restriction of imports. Empirical evidence suggests that the value of rents associated with import licenses can be relatively large, and it has been shown that the welfare cost of quantitative restrictions equals that of their tariff equivalents plus the value of the rents.

While import licenses constitute a large and visible rent resulting from government intervention, the phenomenon of rent seeking is far more general. Fair trade laws result in firms of less-than-optimal size. Minimum wage legislation generates equilibrium levels of unemployment above the optimum with associated deadweight losses, as shown by John Harris and

¹² I am indebted to Bhagwati for pointing out this implication.

Michael Todaro, and Todaro. Ceilings on interest rates and consequent credit rationing lead to competition for loans and deposits and/or high-cost banking operations. Regulating taxi fares affects the average waiting time for a taxi and the percent of time taxis are idle, but probably not their owners' incomes, unless taxis are also licensed. Capital gains tax treatment results in overbuilding of apartments and uneconomic oil exploration. And so on.

Each of these and other interventions lead people to compete for the rents although the competitors often do not perceive themselves as such. In each case there is a deadweight loss associated with that competition over and above the traditional triangle. In general, prevention of that loss can be achieved only by restricting entry into the activity for which a rent has been created.

That, in turn, has political implications. First, even if they *can* limit competition for the rents, governments which consider they must impose restrictions are caught on the horns of a dilemma: if they do restrict entry, they are clearly "showing favoritism" to one group in society and are choosing an unequal distribution of income. If, instead, competition for the rents is allowed (or cannot be prevented), income distribution may be less unequal and certainly there will be less appearance of favoring special groups, although the economic costs associated with quantitative restrictions will be higher.

Second, the existence of rent seeking surely affects people's perception of the economic system. If income distribution is viewed as the outcome of a lottery where wealthy individuals are successful (or lucky) rent seekers, whereas the poor are those precluded from or unsuccessful in rent seeking, the market mechanism is bound to be suspect. In the United States, rightly or wrongly, societal consensus has

been that high incomes reflect—at least to some degree—high social product. As such, the high American per capita income is seen as a result of a relatively free market mechanism and an unequal distribution is tolerated as a by-product. If, instead, it is believed that few businesses would survive without exerting "influence," even if only to bribe government officials to do what they ought in any event to do, it is difficult to associate pecuniary rewards with social product. The perception of the price system as a mechanism rewarding the rich and well-connected may also be important in influencing political decisions about economic policy. If the market mechanism is suspect, the inevitable temptation is to resort to greater and greater intervention, thereby increasing the amount of economic activity devoted to rent seeking. As such, a political "vicious circle" may develop. People perceive that the market mechanism does not function in a way compatible with socially approved goals because of competitive rent seeking. A political consensus therefore emerges to intervene further in the market, rent seeking increases, and further intervention results. While it is beyond the competence of an economist to evaluate the political impact of rent seeking, the suspicion of the market mechanism so frequently voiced in some developing countries may result from it.

Finally, all market economies have some rent-generating restrictions. One can conceive of a continuum between a system of no restrictions and a perfectly restricted system. With no restrictions, entrepreneurs would seek to achieve windfall gains by adopting new technology, anticipating market shifts correctly, and so on. With perfect restrictions, regulations would be so all-pervasive that rent seeking would be the only route to gain. In such a system, entrepreneurs would devote all their time and resources to capturing windfall rents.

While neither of these extreme types could ever exist, one can perhaps ask whether there might be some point along the continuum beyond which the market fails to perform its allocative function to any satisfactory degree. It will remain for further work to formalize these conjectures and to test their significance. It is hoped, however, that enough has been said to stimulate interest and research on the subject.

REFERENCES

- J. Bhagwati, "On the Equivalence of Tariffs and Quotas," in his *Trade, Tariffs and Growth*, London 1969.
- and P. Desai, *Planning for Industrialization: A Study of India's Trade and Industrial Policies Since 1950*, Cambridge 1970.
- and A. Krueger, *Foreign Trade Regimes and Economic Development: Experience and Analysis*, New York forthcoming.
- J. R. Harris and M. P. Todaro, "Migration, Unemployment, and Development: A Two-Sector Analysis," *Amer. Econ. Rev.*, Mar. 1970, 60, 126-42.
- H. G. Johnson, "The Cost of Protection and the Scientific Tariff," *J. Polit. Econ.*, Aug. 1960, 68, 327-45.
- A. Krueger, *Foreign Trade Regimes and Economic Development: Turkey*, New York 1974.
- J. B. Monteiro, *Corruption*, Bombay 1966.
- G. Myrdal, *Asian Drama*, Vol. III, New York 1968.
- M. P. Todaro, "A Model of Labor Migration and Urban Employment in Less Developed Countries," *Amer. Econ. Rev.*, Mar. 1969, 59, 138-48.
- R. Wraith and E. Simpkins, *Corruption in Developing Countries*, London 1963.
- Government of India, Planning Commission, *Third Five Year Plan*, New Delhi, Aug. 1961.
- Reserve Bank of India, *Report on Currency and Finance*, 1967-68.
- Santhanam Committee, *Report on the Committee on Prevention of Corruption*, Government of India, Ministry of Home Affairs, New Delhi 1964.

Monopolistic Competition and Optimum Product Diversity

By AVINASH K. DIXIT AND JOSEPH E. STIGLITZ*

The basic issue concerning production in welfare economics is whether a market solution will yield the socially optimum kinds and quantities of commodities. It is well known that problems can arise for three broad reasons: distributive justice; external effects; and scale economies. This paper is concerned with the last of these.

The basic principle is easily stated.¹ A commodity should be produced if the costs can be covered by the sum of revenues and a properly defined measure of consumer's surplus. The optimum amount is then found by equating the demand price and the marginal cost. Such an optimum can be realized in a market if perfectly discriminatory pricing is possible. Otherwise we face conflicting problems. A competitive market fulfilling the marginal condition would be unsustainable because total profits would be negative. An element of monopoly would allow positive profits, but would violate the marginal condition.² Thus we expect a market solution to be suboptimal. However, a much more precise structure must be put on the problem if we are to understand the nature of the bias involved.

It is useful to think of the question as one of quantity versus diversity. With scale economies, resources can be saved by producing fewer goods and larger quantities of each. However, this leaves less variety, which entails some welfare loss. It is easy and probably not too unrealistic to model scale economies by supposing that each

potential commodity involves some fixed set-up cost and has a constant marginal cost. Modeling the desirability of variety has been thought to be difficult, and several indirect approaches have been adopted. The Hotelling spatial model, Lancaster's product characteristics approach, and the mean-variance portfolio selection model have all been put to use.³ These lead to results involving transport costs or correlations among commodities or securities, and are hard to interpret in general terms. We therefore take a direct route, noting that the convexity of indifference surfaces of a conventional utility function defined over the quantities of all potential commodities already embodies the desirability of variety. Thus, a consumer who is indifferent between the quantities (1,0) and (0,1) of two commodities prefers the mix (1/2,1/2) to either extreme. The advantage of this view is that the results involve the familiar own- and cross-elasticities of demand functions, and are therefore easier to comprehend.

There is one case of particular interest on which we concentrate. This is where potential commodities in a group or sector or industry are good substitutes among themselves, but poor substitutes for the other commodities in the economy. Then we are led to examining the market solution in relation to an optimum, both as regards biases within the group, and between the group and the rest of the economy. We expect the answer to depend on the intra- and intersector elasticities of substitution. To demonstrate the point as simply as possible, we shall aggregate the rest of the economy into one good labeled 0, chosen as the numeraire. The economy's endowment of it is normalized at unity; it can be thought of as the time at the disposal of the consumers.

*Professors of economics, University of Warwick and Stanford University, respectively. Stiglitz's research was supported in part by NSF Grant SOC74-22182 at the Institute for Mathematical Studies in the Social Sciences, Stanford. We are indebted to Michael Spence, to a referee, and the managing editor for comments and suggestions on earlier drafts.

¹See also the exposition by Michael Spence.

²A simple exposition is given by Peter Diamond and Daniel McFadden.

³See the articles by Harold Hotelling, Nicholas Stern, Kelvin Lancaster, and Stiglitz.

The potential range of related products is labeled 1,2,3,... Writing the amounts of the various commodities as x_0 and $x = (x_1, x_2, x_3, \dots)$, we assume a separable utility function with convex indifference surfaces:

$$(1) \quad u = U(x_0, V(x_1, x_2, x_3, \dots))$$

In Sections I and II we simplify further by assuming that V is a symmetric function, and that all commodities in the group have equal fixed and marginal costs. Then the actual labels given to commodities are immaterial, even though the total number n being produced is relevant. We can thus label these commodities 1,2, ..., n , where the potential products $(n+1), (n+2), \dots$ are not being produced. This is a restrictive assumption, for in such problems we often have a natural asymmetry owing to graduated physical differences in commodities, with a pair close together being better mutual substitutes than a pair farther apart. However, even the symmetric case yields some interesting results. In Section III, we consider some aspects of asymmetry.

We also assume that all commodities have unit income elasticities. This differs from a similar recent formulation by Michael Spence, who assumes U linear in x_0 , so that the industry is amenable to partial equilibrium analysis. Our approach allows a better treatment of the intersectoral substitution, but the other results are very similar to those of Spence.

We consider two special cases of (1). In Section I, V is given a CES form, but U is allowed to be arbitrary. In Section II, U is taken to be Cobb-Douglas, but V has a more general additive form. Thus the former allows more general intersector relations, and the latter more general intra-sector substitution, highlighting different results.

Income distribution problems are neglected. Thus U can be regarded as representing Samuelsonian social indifference curves, or (assuming the appropriate aggregation conditions to be fulfilled) as a multiple of a representative consumer's utility. Product diversity can then be interpreted either as different consumers using different

varieties, or as diversification on the part of each consumer.

I. Constant-Elasticity Case

A. Demand Functions

The utility function in this section is

$$(2) \quad u = U\left(x_0, \left\{\sum_i x_i^\rho\right\}^{1/\rho}\right)$$

For concavity, we need $\rho < 1$. Further, since we want to allow a situation where several of the x_i are zero, we need $\rho > 0$. We also assume U homothetic in its arguments.

The budget constraint is

$$(3) \quad x_0 + \sum_{i=1}^n p_i x_i = I$$

where p_i are prices of the goods being produced, and I is income in terms of the numeraire, i.e., the endowment which has been set at 1 plus the profits of the firms distributed to the consumers, or minus the lump sum deductions to cover the losses, as the case may be.

In this case, a two-stage budgeting procedure is valid.⁴ Thus we define dual quantity and price indices

$$(4) \quad y = \left\{ \sum_{i=1}^n x_i^\rho \right\}^{1/\rho} \quad q = \left\{ \sum_{i=1}^n p_i^{-1/\beta} \right\}^{-\beta}$$

where $\beta = (1 - \rho)/\rho$, which is positive since $0 < \rho < 1$. Then it can be shown⁵ that in the first stage,

$$(5) \quad y = I \frac{s(q)}{q} \quad x_0 = I(1 - s(q))$$

for a function s which depends on the form of U . Writing $\sigma(q)$ for the elasticity of substitution between x_0 and y , we define $\theta(q)$ as the elasticity of the function s , i.e., $qs'(q)/s(q)$. Then we find

$$(6) \quad \theta(q) = \{1 - \sigma(q)\} \{1 - s(q)\} < 1$$

but $\theta(q)$ can be negative as $\sigma(q)$ can exceed 1.

⁴See p. 21 of John Green.

⁵These details and several others are omitted to save space, but can be found in the working paper by the authors, cited in the references.

Turning to the second stage of the problem, it is easy to show that for each i ,

$$(7) \quad x_i = y \left[\frac{q}{p_i} \right]^{1/(1-\rho)}$$

where y is defined by (4). Consider the effect of a change in p_i alone. This affects x_i directly, and also through q ; thence through y as well. Now from (4) we have the elasticity

$$(8) \quad \frac{\partial \log q}{\partial \log p_i} = \left(\frac{q}{p_i} \right)^{1/\beta}$$

So long as the prices of the products in the group are not of different orders of magnitude, this is of the order $(1/n)$. We shall assume that n is reasonably large, and accordingly neglect the effect of each p_i on q ; thus the indirect effects on x_i . This leaves us with the elasticity

$$(9) \quad \frac{\partial \log x_i}{\partial \log p_i} = \frac{-1}{(1-\rho)} = \frac{-(1+\beta)}{\beta}$$

In the Chamberlinian terminology, this is the elasticity of the *dd* curve, i.e., the curve relating the demand for each product type to its own price with all other prices held constant.

In our large group case, we also see that for $i \neq j$, the cross elasticity $\partial \log x_i / \partial \log p_j$ is negligible. However, if all prices in the group move together, the individually small effects add to a significant amount. This corresponds to the Chamberlinian *DD* curve. Consider a symmetric situation where $x_i = x$ and $p_i = p$ for all i from 1 to n . We have

$$(10) \quad \begin{aligned} y &= xn^{1/\rho} = xn^{1+\beta} \\ q &= pn^{-\beta} = pn^{-(1-\rho)/\rho} \end{aligned}$$

and then from (5) and (7),

$$(11) \quad x = \frac{Is(q)}{pn}$$

The elasticity of this is easy to calculate; we find

$$(12) \quad \frac{\partial \log x}{\partial \log p} = -[1 - \theta(q)]$$

Then (6) shows that the *DD* curve slopes

downward. The conventional condition that the *dd* curve be more elastic is seen from (9) and (12) to be

$$(13) \quad \frac{1}{\beta} + \theta(q) > 0$$

Finally, we observe that for $i \neq j$,

$$(14) \quad \frac{x_i}{x_j} = \left[\frac{p_j}{p_i} \right]^{1/(1-\rho)}$$

Thus $1/(1-\rho)$ is the elasticity of substitution between any two products within the group.

B. Market Equilibrium

It can be shown that each commodity is produced by one firm. Each firm attempts to maximize its profit, and entry occurs until the marginal firm can only just break even. Thus our market equilibrium is the familiar case of Chamberlinian monopolistic competition, where the question of quantity versus diversity has often been raised.⁶ Previous analyses have failed to consider the desirability of variety in an explicit form, and have neglected various intra- and intersector interactions in demand. As a result, much vague presumption that such an equilibrium involves excessive diversity has built up at the back of the minds of many economists. Our analysis will challenge several of these ideas.

The profit-maximization condition for each firm acting on its own is the familiar equality of marginal revenue and marginal cost. Writing c for the common marginal cost, and noting that the elasticity of demand for each firm is $(1+\beta)/\beta$, we have for each active firm:

$$p_i \left(1 - \frac{\beta}{1+\beta} \right) = c$$

Writing p_e for the common equilibrium price for each variety being produced, we have

$$(15) \quad p_e = c(1+\beta) = \frac{c}{\rho}$$

⁶See Edwin Chamberlin, Nicholas Kaldor, and Robert Bishop.

The second condition for equilibrium is that firms enter until the next potential entrant would make a loss. If n is large enough so that 1 is a small increment, we can assume that the marginal firm is exactly breaking even, i.e., $(p_n - c)x_n = a$, where x_n is obtained from the demand function and a is the fixed cost. With symmetry, this implies zero profit for all intramarginal firms as well. Then $I = 1$, and using (11) and (15) we can write the condition so as to yield the number n_e of active firms:

$$(16) \quad \frac{s(p_e n_e^{-\beta})}{p_e n_e} = \frac{a}{\beta c}$$

Equilibrium is unique provided $s(p_e n_e^{-\beta})/p_e n_e$ is a monotonic function of n . This relates to our earlier discussion about the two demand curves. From (11) we see that the behavior of $s(pn^{-\beta})/pn$ as n increases tells us how the demand curve DD for each firm shifts as the number of firms increases. It is natural to assume that it shifts to the left, i.e., the function above decreases as n increases for each fixed p . The condition for this in elasticity form is easily seen to be

$$(17) \quad 1 + \beta \theta(q) > 0$$

This is exactly the same as (13), the condition for the dd curve to be more elastic than the DD curve, and we shall assume that it holds.

The condition can be violated if $\sigma(q)$ is sufficiently higher than one. In this case, an increase in n lowers q , and shifts demand towards the monopolistic sector to such an extent that the demand curve for each firm shifts to the right. However, this is rather implausible.

Conventional Chamberlinian analysis assumes a fixed demand curve for the group as a whole. This amounts to assuming that $n \cdot x$ is independent of n , i.e., that $s(pn^{-\beta})$ is independent of n . This will be so if $\beta = 0$, or if $\sigma(q) = 1$ for all q . The former is equivalent to assuming that $\rho = 1$, when all products in the group are perfect substitutes, i.e., diversity is not valued at all. That would be contrary to the intent of the whole analysis. Thus, implicitly, conventional analysis assumes $\sigma(q) = 1$. This gives a con-

stant budget share for the monopolistically competitive sector. Note that in our parametric formulation, this implies a unit-elastic DD curve, (17) holds, and so equilibrium is unique.

Finally, using (7), (11), and (16), we can calculate the equilibrium output for each active firm:

$$(18) \quad x_e = \frac{a}{\beta c}$$

We can also write an expression for the budget share of the group as a whole:

$$(19) \quad s_e = s(q_e)$$

where $q_e = p_e n_e^{-\beta}$

These will be useful for subsequent comparisons.

C. Constrained Optimum

The next task is to compare the equilibrium with a social optimum. With economies of scale, the first best or unconstrained (really constrained only by technology and resource availability) optimum requires pricing below average cost, and therefore lump sum transfers to firms to cover losses. The conceptual and practical difficulties of doing so are clearly formidable. It would therefore appear that a more appropriate notion of optimality is a constrained one, where each firm must have nonnegative profits. This may be achieved by regulation, or by excise or franchise taxes or subsidies. The important restriction is that lump sum subsidies are not available.

We begin with such a constrained optimum. The aim is to choose n , p_i , and x_i so as to maximize utility, satisfying the demand functions and keeping the profit for each firm nonnegative. The problem is somewhat simplified by the result that all active firms should have the same output levels and prices, and should make exactly zero profit. We omit the proof. Then we can set $I = 1$, and use (5) to express utility as a function of q alone. This is of course a decreasing function. Thus the problem of maximizing u becomes that of minimizing q , i.e.,

$$\min_{n,p} pn^{-\beta}$$

subject to

$$(20) \quad (p - c) \frac{s(pn^{-\beta})}{pn} = a$$

To solve this, we calculate the logarithmic marginal rate of substitution along a level curve of the objective, the similar rate of transformation along the constraint, and equate the two. This yields the condition

$$(21) \quad \frac{\frac{c}{p-c} + \theta(q)}{1 + \beta\theta(q)} = \frac{1}{\beta}$$

The second-order condition can be shown to hold, and (21) simplifies to yield the price for each commodity produced in the constrained optimum, p_c , as

$$(22) \quad p_c = c(1 + \beta)$$

Comparing (15) and (22), we see that the two solutions have the same price. Since they face the same break-even constraint, they have the same number of firms as well, and the values for all other variables can be calculated from these two. Thus we have a rather surprising case where the monopolistic competition equilibrium is identical with the optimum constrained by the lack of lump sum subsidies. Chamberlin once suggested that such an equilibrium was "a sort of ideal"; our analysis shows when and in what sense this can be true.

D. Unconstrained Optimum

These solutions can in turn be compared to the unconstrained or first best optimum. Considerations of convexity again establish that all active firms should produce the same output. Thus we are to choose n firms each producing output x in order to maximize

$$(23) \quad u = U(1 - n(a + cx), xn^{1+\beta})$$

where we have used the economy's resource balance condition and (10). The first-order conditions are

$$(24) \quad -ncU_0 + n^{1+\beta}U_y = 0$$

$$(25) \quad -(a + cx)U_0 + (1 + \beta)xn^\beta U_y = 0$$

From the first stage of the budgeting problem, we know that $q = U_y/U_0$. Using (24) and (10), we find the price charged by each active firm in the unconstrained optimum, p_u , equal to marginal cost

$$(26) \quad p_u = c$$

This, of course, is no surprise. Also from the first-order conditions, we have

$$(27) \quad x_u = \frac{a}{c\beta}$$

Finally, with (26), each active firm covers its variable cost exactly. The lump sum transfers to firms then equal an , and therefore $I = 1 - an$, and

$$x = (1 - an) \frac{s(pn^{-\beta})}{pn}$$

The number of firms n_u is then defined by

$$(28) \quad \frac{s(cn_u^{-\beta})}{n_u} = \frac{a/\beta}{1 - an_u}$$

We can now compare these magnitudes with the corresponding ones in the equilibrium or the constrained optimum. The most remarkable result is that the output of each active firm is the same in the two situations. The fact that in a Chamberlinian equilibrium each firm operates to the left of the point of minimum average cost has been conventionally described by saying that there is excess capacity. However, when variety is desirable, i.e., when the different products are not perfect substitutes, it is not in general optimum to push the output of each firm to the point where all economies of scale are exhausted.⁷ We have shown in one case that is not an extreme one, that the first best optimum does not exploit economies of scale beyond the extent achieved in the equilibrium. We can then easily conceive of cases where the equilibrium exploits economies of scale too far from the point of view of social optimality. Thus our results undermine the validity of the folklore of excess capacity, from the point of view of the

⁷See David Starrett.

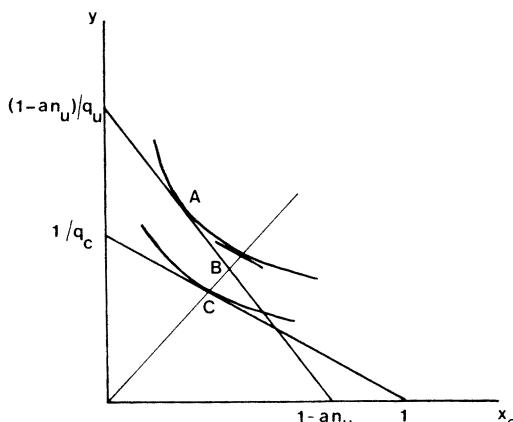


FIGURE 1

unconstrained optimum as well as the constrained one.

A direct comparison of the numbers of firms from (16) and (28) would be difficult, but an indirect argument turns out to be simple. It is clear that the unconstrained optimum has higher utility than the constrained optimum. Also, the level of lump sum income in it is less than that in the latter. It must therefore be the case that

$$(29) \quad q_u < q_c = q_e$$

Further, the difference must be large enough that the budget constraint for x_0 and the quantity index y in the unconstrained case must lie outside that in the constrained case in the relevant region, as shown in Figure 1. Let C be the constrained optimum, A the unconstrained optimum, and let B be the point where the line joining the origin to C meets the indifference curve in the unconstrained case. By homotheticity the indifference curve at B is parallel to that at C , so each of the moves from C to B and from B to A increases the value of y . Since the value of x is the same in the two optima, we must have

$$(30) \quad n_u > n_c = n_e$$

Thus the unconstrained optimum actually allows more variety than the constrained optimum and the equilibrium; this is another point contradicting the folklore on excessive diversity.

Using (29) we can easily compare the budget shares. In the notation we have been using, we find $s_u \geq s_c$ as $\theta(q) \geq 0$, i.e., as $\sigma(q) \geq 1$ providing these hold over the entire relevant range of q .

It is not possible to have a general result concerning the relative magnitudes of x_0 in the two situations; an inspection of Figure 1 shows this. However, we have a sufficient condition:

$$\begin{aligned} x_{0u} &= (1 - an_u)(1 - s_u) < 1 - s_u \leq 1 - s_c \\ &= x_{0c} \text{ if } \sigma(q) \geq 1 \end{aligned}$$

In this case the equilibrium or the constrained optimum use more of the numeraire resource than the unconstrained optimum. On the other hand, if $\sigma(q) = 0$ we have L-shaped isoquants, and in Figure 1, points A and B coincide giving the opposite conclusion.

In this section we have seen that with a constant intrasector elasticity of substitution, the market equilibrium coincides with the constrained optimum. We have also shown that the unconstrained optimum has a greater number of firms, each of the same size. Finally, the resource allocation between the sectors is shown to depend on the intersector elasticity of substitution. This elasticity also governs conditions for uniqueness of equilibrium and the second-order conditions for an optimum.

Henceforth we will achieve some analytic simplicity by making a particular assumption about intersector substitution. In return, we will allow a more general form of intrasector substitution.

II. Variable Elasticity Case

The utility function is now

$$(31) \quad u = x_0^{1-\gamma} \left\{ \sum_i v(x_i) \right\}^\gamma$$

with v increasing and concave, $0 < \gamma < 1$. This is somewhat like assuming a unit intersector elasticity of substitution. However, this is not rigorous since the group utility $V(\underline{x}) = \sum_i v(x_i)$ is not homothetic and therefore two-stage budgeting is not applicable.

It can be shown that the elasticity of the dd curve in the large group case is

$$(32) \quad -\frac{\partial \log x_i}{\partial \log p_i} = -\frac{v'(x_i)}{x_i v''(x_i)} \quad \text{for any } i$$

This differs from the case of Section I in being a function of x_i . To highlight the similarities and the differences, we define $\beta(x)$ by

$$(33) \quad \frac{1 + \beta(x)}{\beta(x)} = -\frac{v'(x)}{x v''(x)}$$

Next, setting $x_i = x$ and $p_i = p$ for $i = 1, 2, \dots, n$, we can write the *DD* curve and the demand for the numeraire as

$$(34) \quad x = \frac{I}{np} \omega(x), \quad x_0 = I[1 - \omega(x)]$$

where

$$(35) \quad \begin{aligned} \omega(x) &= \frac{\gamma \rho(x)}{[\gamma \rho(x) + (1 - \gamma)]} \\ \rho(x) &= \frac{x v'(x)}{v(x)} \end{aligned}$$

We assume that $0 < \rho(x) < 1$, and therefore have $0 < \omega(x) < 1$.

Now consider the Chamberlinian equilibrium. The profit-maximization condition for each active firm yields the common equilibrium price p_e in terms of the common equilibrium output x_e as

$$(36) \quad p_e = c[1 + \beta(x_e)]$$

Note the analogy with (15). Substituting (36) in the zero pure profit condition, we have x_e defined by

$$(37) \quad \frac{c x_e}{a + c x_e} = \frac{1}{1 + \beta(x_e)}$$

Finally, the number of firms can be calculated using the *DD* curve and the break-even condition, as

$$(38) \quad n_e = \frac{\omega(x_e)}{a + c x_e}$$

For uniqueness of equilibrium we once again use the conditions that the *dd* curve is more elastic than the *DD* curve, and that entry shifts the *DD* curve to the left. However, these conditions are rather involved and opaque, so we omit them.

Let us turn to the constrained optimum.

We wish to choose n and x to maximize u , subject to (34) and the break-even condition $p x = a + cx$. Substituting, we can express u as a function of x alone:

$$(39) \quad u = \gamma^\gamma (1 - \gamma)^{(1-\gamma)} \frac{\left[\frac{\rho(x)v(x)}{a + cx} \right]^\gamma}{\gamma \rho(x) + (1 - \gamma)}$$

The first-order condition defines x_c :

$$(40) \quad \frac{cx_c}{a + cx_c} = \frac{1}{1 + \beta(x_c)} - \frac{\omega(x_c)x_c\rho'(x_c)}{\gamma \rho(x_c)}$$

Comparing this with (37) and using the second-order condition, it can be shown that provided $\rho'(x)$ is one-signed for all x ,

$$(41) \quad x_c \gtrless x_e \text{ according as } \rho'(x) \lessgtr 0$$

With zero pure profit in each case, the points (x_e, p_e) and (x_c, p_c) lie on the same declining average cost curve, and therefore

$$(42) \quad p_c \lessgtr p_e \text{ according as } x_c \gtrless x_e$$

Next we note that the *dd* curve is tangent to the average cost curve at (x_e, p_e) and the *DD* curve is steeper. Consider the case $x_c > x_e$. Now the point (x_c, p_c) must lie on a *DD* curve further to the right than (x_e, p_e) , and therefore must correspond to a smaller number of firms. The opposite happens if $x_c < x_e$. Thus,

$$(43) \quad n_c \lessgtr n_e \text{ according as } x_c \gtrless x_e$$

Finally, (41) shows that in both cases that arise there, $\rho(x_c) < \rho(x_e)$. Then $\omega(x_c) < \omega(x_e)$, and from (34),

$$(44) \quad x_{0c} > x_{0e}$$

A smaller degree of intersectoral substitution could have reversed the result, as in Section I.

An intuitive reason for these results can be given as follows. With our large group assumptions, the revenue of each firm is proportional to $xv'(x)$. However, the contribution of its output to group utility is $v(x)$. The ratio of the two is $\rho(x)$. Therefore, if $\rho'(x) > 0$, then at the margin each firm finds it more profitable to expand than what would be socially desirable, so $x_e > x_c$.

Given the break-even constraint, this leads to there being fewer firms.

Note that the relevant magnitude is the elasticity of utility, and not the elasticity of demand. The two are related, since

$$(45) \quad x \frac{\rho'(x)}{\rho(x)} = \frac{1}{1 + \beta(x)} - \rho(x)$$

Thus, if $\rho(x)$ is constant over an interval, so is $\beta(x)$ and we have $1/(1 + \beta) = \rho$, which is the case of Section I. However, if $\rho(x)$ varies, we cannot infer a relation between the signs of $\rho'(x)$ and $\beta'(x)$. Thus the variation in the elasticity of demand is not in general the relevant consideration. However, for important families of utility functions there is a relationship. For example, for $v(x) = (k + mx)^j$, with $m > 0$ and $0 < j < 1$, we find that $-xv''/v'$ and xv'/v are positively related. Now we would normally expect that as the number of commodities produced increases, the elasticity of substitution between any pair of them should increase. In the symmetric equilibrium, this is just the inverse of the elasticity of marginal utility. Then a higher x would correspond to a lower n , and therefore a lower elasticity of substitution, higher $-xv''/v'$ and higher xv'/v . Thus we are led to expect that $\rho'(x) > 0$, i.e., that the equilibrium involves fewer and bigger firms than the constrained optimum. Once again the common view concerning excess capacity and excessive diversity in monopolistic competition is called into question.

The unconstrained optimum problem is to choose n and x to maximize

$$(46) \quad u = [nv(x)]^\gamma [1 - n(a + cx)]^{1-\gamma}$$

It is easy to show that the solution has

$$(47) \quad p_u = c$$

$$(48) \quad \frac{cx_u}{a + cx_u} = \rho(x_u)$$

$$(49) \quad n_u = \frac{\gamma}{a + cx_u}$$

Then we can use the second-order condition to show that

$$(50) \quad x_u \leq x_c \text{ according as } \rho'(x) \geq 0$$

This is in each case transitive with (41), and therefore yields similar output comparisons between the equilibrium and the unconstrained optimum.

The price in the unconstrained optimum is of course the lowest of the three. As to the number of firms, we note

$$n_c = \frac{\omega(x_c)}{a + cx_c} < \frac{\gamma}{a + cx_c}$$

and therefore we have a one-way comparison:

$$(51) \quad \text{If } x_u < x_c, \text{ then } n_u > n_c$$

Similarly for the equilibrium. These leave open the possibility that the unconstrained optimum has both bigger and more firms. That is not unreasonable; after all the unconstrained optimum uses resources more efficiently.

III. ASYMMETRIC CASES

The discussion so far imposed symmetry within the group. Thus the number of varieties being produced was relevant, but any group of n was just as good as any other group of n . The next important modification is to remove this restriction. It is easy to see how interrelations within the group of commodities can lead to biases. Thus, if no sugar is being produced, the demand for coffee may be so low as to make its production unprofitable when there are set-up costs. However, this is open to the objection that with complementary commodities, there is an incentive for one entrant to produce both. However, problems exist even when all the commodities are substitutes. We illustrate this by considering an industry which will produce commodities from one of two groups, and examine whether the choice of the wrong group is possible.⁸

Suppose there are two sets of commodities beside the numeraire, the two being perfect substitutes for each other and each having a constant elasticity subutility function. Further, we assume a constant budget share

⁸For an alternative approach using partial equilibrium methods, see Spence.

for the numeraire. Thus the utility function is

$$(52) \quad u = x_0^{1-s} \left\{ \left[\sum_{i_1=1}^n x_{i_1}^{\rho_1} \right]^{1/\rho_1} + \left[\sum_{i_2=1}^{n_2} x_i^{\rho_2} \right]^{1/\rho_2} \right\}^s$$

We assume that each firm in group i has a fixed cost a_i and a constant marginal cost c_i .

Consider two types of equilibria, only one commodity group being produced in each. These are given by

$$(53a) \quad \bar{x}_1 = \frac{a_1}{c_1 \beta_1}, \bar{x}_2 = 0 \\ \bar{p}_1 = c_1(1 + \beta_1) \\ \bar{n}_1 = \frac{s \beta_1}{a_1(1 + \beta_1)} \\ \bar{q}_1 = \bar{p}_1 \bar{n}_1^{-\beta_1} = c_1(1 + \beta_1)^{1+\beta_1} \left(\frac{a_1}{s} \right)^{\beta_1} \\ \bar{u}_1 = s^s (1 - s)^{1-s} \bar{q}_1^{-s}$$

$$(53b) \quad \bar{x}_2 = \frac{a_2}{c_2 \beta_2}, \bar{x}_1 = 0 \\ \bar{p}_2 = c_2(1 + \beta_2) \\ \bar{n}_2 = \frac{s \beta_2}{a_2(1 + \beta_2)} \\ \bar{q}_2 = \bar{p}_2 \bar{n}_2^{-\beta_2} = c_2(1 + \beta_2)^{1+\beta_2} \left(\frac{a_2}{s} \right)^{\beta_2} \\ \bar{u}_2 = s^s (1 - s)^{1-s} \bar{q}_2^{-s}$$

Equation (53a) is a Nash equilibrium if and only if it does not pay a firm to produce a commodity of the second group. The demand for such a commodity is

$$x_2 = \begin{cases} 0 & \text{for } p_2 \geq \bar{q}_1 \\ s/p_2 & \text{for } p_2 < \bar{q}_1 \end{cases}$$

Hence we require

$$\max_{p_2} (p_2 - c_2)x_2 = s \left(1 - \frac{c_2}{\bar{q}_1} \right) < a_2$$

or

$$(54) \quad \bar{q}_1 < \frac{s c_2}{s - a_2}$$

Similarly, (53b) is a Nash equilibrium if and

only if

$$(55) \quad \bar{q}_2 < \frac{s c_1}{s - a_1}$$

Now consider the optimum. Both the objective and the constraint are such as to lead the optimum to the production of commodities from only one group. Thus, suppose n_i commodities from group i are being produced at levels x_i each, and offered at prices p_i . The utility level is given by

$$(56) \quad u = x_0^{1-s} \{x_1 n_1^{1+\beta_1} + x_2 n_2^{1+\beta_2}\}^s$$

and the resource availability constraint is

$$(57) \quad x_0 + n_1(a_1 + c_1 x_1) + n_2(a_2 + c_2 x_2) = 1$$

Given the values of the other variables, the level curves of u in (n_1, n_2) space are concave to the origin, while the constraint is linear. We must therefore have a corner optimum. (As for the break-even constraint, unless the two $q_i = p_i n_i^{-\beta_i}$ are equal, the demand for commodities in one group is zero, and there is no possibility of avoiding a loss there.)

Note that we have structured our example so that if the correct group is chosen, the equilibrium will not introduce any further biases in relation to the constrained optimum. Therefore, to find the constrained optimum, we only have to look at the values of \bar{u}_i in (53a) and (53b) and see which is the greater. In other words, we have to see which \bar{q}_i is the smaller, and choose the situation (which may or may not be a Nash equilibrium) defined in (53a) and (53b) corresponding to it.

Figure 2 is drawn to depict the possible equilibria and optima. Given all the relevant parameters, we calculate (\bar{q}_1, \bar{q}_2) from (53a) and (53b). Then (54) and (55) tell us whether either or both of the situations are possible equilibria, while a simple comparison of the magnitudes of \bar{q}_1 and \bar{q}_2 tells us which is the constrained optimum. In the figure, the nonnegative quadrant is split into regions in each of which we have one combination of equilibria and optima. We only have to locate the point (\bar{q}_1, \bar{q}_2) in this space to know the result for the given

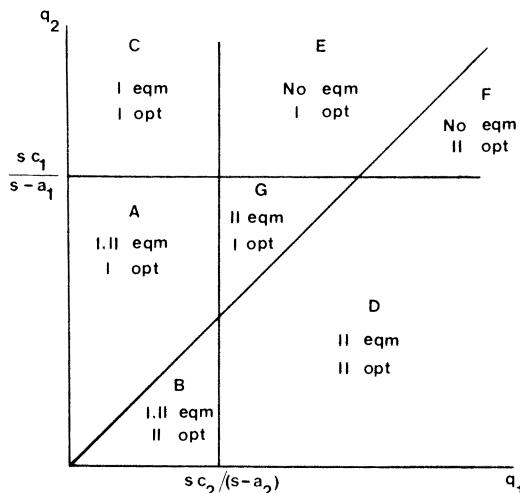


FIGURE 2. SOLUTIONS LABELED I REFER TO EQUATION (53a); SOLUTIONS LABELED II REFER TO EQUATION (53b)

parameter values. Moreover, we can compare the location of the points corresponding to different parameter values and thus do some comparative statics.

To understand the results, we must examine how \bar{q}_i depends on the relevant parameters. It is easy to see that each is an increasing function of a_i and c_i . We also find

$$(58) \quad \frac{\partial \log \bar{q}_i}{\partial \beta_i} = -\log \bar{n}_i$$

and we expect this to be large and negative. Further, we see from (9) that a higher β_i corresponds to a lower own-price elasticity of demand for each commodity in that group. Thus \bar{q}_i is an increasing function of this elasticity.

Consider initially a symmetric situation, with $sc_1/(s - a_1) = sc_2/(s - a_2)$, $\beta_1 = \beta_2$ (the region G vanishes then), and suppose the point (\bar{q}_1, \bar{q}_2) is on the boundary between regions A and B . Now consider a change in one parameter, say, a higher own-elasticity for commodities in group 2. This raises \bar{q}_2 , moving the point into region A , and it becomes optimal to produce commodities from group 1 alone. However, both (53a) and (53b) are possible Nash

equilibria, and it is therefore possible that *the high elasticity group is produced in equilibrium when the low elasticity one should have been*. If the difference in elasticities is large enough, the point moves into region C , where (53b) is no longer a Nash equilibrium. But, owing to the existence of a fixed cost, a significant difference in elasticities is necessary before entry from group 1 commodities threatens to destroy the "wrong" equilibrium. Similar remarks apply to regions B and D .

Next, begin with symmetry once again, and consider a higher c_1 or a_1 . This increases \bar{q}_1 and moves the point into region B , making it optimal to produce the low-cost group alone while leaving both (53a) and (53b) as possible equilibria, until the difference in costs is large enough to take the point to region D . The change also moves the boundary between A and C upward, opening up a larger region G , but that is not of significance here.

If both \bar{q}_1 and \bar{q}_2 are large, each group is threatened by profitable entry from the other, and no Nash equilibrium exists, as in regions E and F . However, the criterion of constrained optimality remains as before. Thus we have a case where it may be necessary to prohibit entry in order to sustain the constrained optimum.

If we combine a case where $c_1 > c_2$ (or $a_1 > a_2$) and $\beta_1 > \beta_2$, i.e., where commodities in group 2 are more elastic and have lower costs, we face a still worse possibility. For the point (\bar{q}_1, \bar{q}_2) may then lie in region G , where only (53b) is a possible equilibrium and only (53a) is constrained optimum, i.e., the market can produce only a low cost, high demand elasticity group of commodities when a high cost, low demand elasticity group should have been produced.

Very roughly, the point is that although commodities in inelastic demand have the potential for earning revenues in excess of variable costs, they also have significant consumers' surpluses associated with them. Thus it is not immediately obvious whether the market will be biased in favor of them or against them as compared with an optimum. Here we find the latter, and independent findings of Michael Spence in other

contexts confirm this. Similar remarks apply to differences in marginal costs.

In the interpretation of the model with heterogeneous consumers and social indifference curves, inelastically demanded commodities will be the ones which are intensively desired by a few consumers. Thus we have an "economic" reason why the market will lead to a bias against opera relative to football matches, and a justification for subsidization of the former and a tax on the latter, provided the distribution of income is optimum.

Even when cross elasticities are zero, there may be an incorrect choice of commodities to be produced (relative either to an unconstrained or constrained optimum) as Figure 3 illustrates. Figure 3 illustrates a case where commodity *A* has a more elastic demand curve than commodity *B*; *A* is produced in monopolistically competitive equilibrium, while *B* is not. But clearly, it is socially desirable to produce *B*, since ignoring consumer's surplus it is just marginal. Thus, the commodities that are not produced but ought to be are those with inelastic demands. Indeed, if, as in the usual analysis of monopolistic competition, eliminating one firm shifts the demand curve for the other firms to the right (i.e., increases the demand for other firms), if the con-

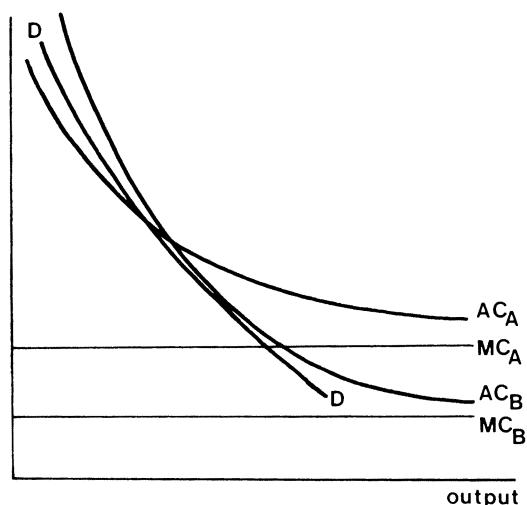


FIGURE 4

sumer surplus from *A* (at its equilibrium level of output) is less than that from *B* (i.e., the cross hatched area exceeds the striped area), then constrained Pareto optimality entails restricting the production of the commodity with the more elastic demand.

A similar analysis applies to commodities with the same demand curves but different cost structures. Commodity *A* is assumed to have the lower fixed cost but the higher marginal cost. Thus, the average cost curves cross but once, as in Figure 4. Commodity *A* is produced in monopolistically competitive equilibrium, commodity *B* is not (although it is just at the margin of being produced). But again, observe that *B* should be produced, since there is a large consumer's surplus; indeed, since were it to be produced, *B* would produce at a much higher level than *A*, there is a much larger consumer's surplus. Thus if the government were to forbid the production of *A*, *B* would be viable, and social welfare would increase.

In the comparison between constrained Pareto optimality and the monopolistically competitive equilibrium, we have observed that in the former, we replace some low fixed cost-high marginal cost commodities with high fixed cost-low marginal cost commodities, and we replace some commodities

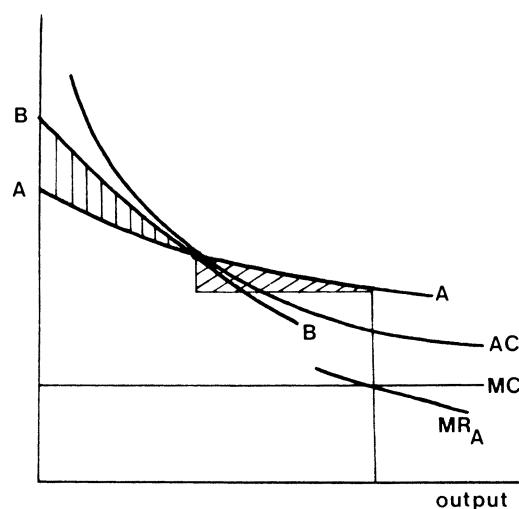


FIGURE 3

with elastic demands with commodities with inelastic demands.

IV. Concluding Remarks

We have constructed in this paper some models to study various aspects of the relationship between market and optimal resource allocation in the presence of some nonconvexities. The following general conclusions seem worth pointing out.

The monopoly power, which is a necessary ingredient of markets with nonconvexities, is usually considered to distort resources away from the sector concerned. However, in our analysis monopoly power enables firms to pay fixed costs, and entry cannot be prevented, so the relationship between monopoly power and the direction of market distortion is no longer obvious.

In the central case of a constant elasticity utility function, the market solution was constrained Pareto optimal, regardless of the value of that elasticity (and thus the implied elasticity of the demand functions). With variable elasticities, the bias could go either way, and the direction of the bias depended not on how the elasticity of demand changed, but on how the elasticity of utility changed. We suggested that there was some presumption that the market solution would be characterized by too few firms in the monopolistically competitive sector.

With asymmetric demand and cost conditions we also observed a bias against commodities with inelastic demands and high costs.

The general principle behind these results is that a market solution considers profit at the appropriate margin, while a social optimum takes into account the consumer's surplus. However, applications of this principle come to depend on details of cost and demand functions. We hope that the cases

presented here, in conjunction with other studies cited, offer some useful and new insights.

REFERENCES

- R. L. Bishop**, "Monopolistic Competition and Welfare Economics," in Robert Kuenne, ed., *Monopolistic Competition Theory*, New York 1967.
- E. Chamberlin**, "Product Heterogeneity and Public Policy," *Amer. Econ. Rev. Proc.*, May 1950, 40, 85-92.
- P. A. Diamond and D. L. McFadden**, "Some Uses of the Expenditure Function In Public Finance," *J. Publ. Econ.*, Feb. 1974, 82, 1-23.
- A. K. Dixit and J. E. Stiglitz**, "Monopolistic Competition and Optimum Product Diversity," econ. res. pap. no. 64, Univ. Warwick, England 1975.
- H. A. John Green**, *Aggregation in Economic Analysis*, Princeton 1964.
- H. Hotelling**, "Stability in Competition," *Econ. J.*, Mar. 1929, 39, 41-57.
- N. Kaldor**, "Market Imperfection and Excess Capacity," *Economica*, Feb. 1934, 2, 33-50.
- K. Lancaster**, "Socially Optimal Product Differentiation," *Amer. Econ. Rev.*, Sept. 1975, 65, 567-85.
- A. M. Spence**, "Product Selection, Fixed Costs, and Monopolistic Competition," *Rev. Econ. Stud.*, June 1976, 43, 217-35.
- D. A. Starrett**, "Principles of Optimal Location in a Large Homogeneous Area," *J. Econ. Theory*, Dec. 1974, 9, 418-48.
- N. H. Stern**, "The Optimal Size of Market Areas," *J. Econ. Theory*, Apr. 1972, 4, 159-73.
- J. E. Stiglitz**, "Monopolistic Competition in the Capital Market," tech. rep. no. 161, IMSS, Stanford Univ., Feb. 1975.

An Almost Ideal Demand System

By ANGUS DEATON AND JOHN MUELLBAUER*

Ever since Richard Stone (1954) first estimated a system of demand equations derived explicitly from consumer theory, there has been a continuing search for alternative specifications and functional forms. Many models have been proposed, but perhaps the most important in current use, apart from the original linear expenditure system, are the Rotterdam model (see Henri Theil, 1965, 1976; Anton Barten) and the translog model (see Laurits Christensen, Dale Jorgenson, and Lawrence Lau; Jorgenson and Lau). Both of these models have been extensively estimated and have, in addition, been used to test the homogeneity and symmetry restrictions of demand theory. In this paper, we propose and estimate a new model which is of comparable generality to the Rotterdam and translog models but which has considerable advantages over both. Our model, which we call the Almost Ideal Demand System (*AIDS*), gives an arbitrary first-order approximation to any demand system; it satisfies the axioms of choice exactly; it aggregates perfectly over consumers without invoking parallel linear Engel curves; it has a functional form which is consistent with known household-budget data; it is simple to estimate, largely avoiding the need for non-linear estimation; and it can be used to test the restrictions of homogeneity and symmetry through linear restrictions on fixed parameters. Although many of these desirable properties are possessed by one or other of the Rotterdam or translog models, neither possesses all of them simultaneously.

In Section I of the paper, we discuss the theoretical specification of the *AIDS* and justify the claims in the previous paragraph.

In Section II, the model is estimated on postwar British data and we use our results to test the homogeneity and symmetry restrictions. Our results are consistent with earlier findings in that both sets of restrictions are decisively rejected. We also find that imposition of homogeneity generates positive serial correlation in the errors of those equations which reject the restrictions most strongly; this suggests that the now standard rejection of homogeneity in demand analysis may be due to insufficient attention to the dynamic aspects of consumer behavior. Finally, in Section III, we offer a summary and conclusions. We believe that the results of this paper suggest that the *AIDS* is to be recommended as a vehicle for testing, extending, and improving conventional demand analysis. This does not imply that the system, particularly in its simple static form, is to be regarded as a fully satisfactory explanation of consumers' behavior. Indeed, by proposing a demand system which is superior to its predecessors, we hope to be able to reveal more clearly the problems and potential solutions associated with the usual approach.

I. Specification of the *AIDS*

In much of the recent literature on systems of demand equations, the starting point has been the specification of a function which is general enough to act as a second-order approximation to any arbitrary direct or indirect utility function or, more rarely, a cost function. For examples, see Christensen, Jorgenson, and Lau; W. Erwin Diewert (1971); or Ernst Berndt, Masako Darrough, and Diewert. Alternatively, it is possible to use a first-order approximation to the demand functions themselves as in the Rotterdam model, see Theil (1965, 1976); Barten. We shall follow these approaches in terms of generality but we start, not from some arbitrary preference

*University of Bristol, and Birkbeck College, London, respectively. We are grateful to David Mitchell for help with the calculations and to Anton Barten, David Hendry, Claus Leser, Louis Philips, and a referee for helpful comments on an earlier version.

ordering, but from a specific class of preferences, which by the theorems of Muellbauer (1975, 1976) permit exact aggregation over consumers: the representation of market demands as if they were the outcome of decisions by a rational representative consumer. These preferences, known as the *PIGLOG* class, are represented via the *cost* or *expenditure function* which defines the minimum expenditure necessary to attain a specific utility level at given prices. We denote this function $c(u, p)$ for utility u and price vector p , and define the *PIGLOG* class by

(1)

$$\log c(u, p) = (1 - u) \log \{a(p)\} + u \log \{b(p)\}$$

With some exceptions (see the Appendix), u lies between 0 (subsistence) and 1 (bliss) so that the positive linearly homogeneous functions $a(p)$ and $b(p)$ can be regarded as the costs of subsistence and bliss, respectively. The Appendix further discusses this general model as well as the implications of the underlying aggregation theory.

Next we take specific functional forms for $\log a(p)$ and $\log b(p)$. For the resulting cost function to be a flexible functional form, it must possess enough parameters so that at any single point its derivatives $\partial c / \partial p_i$, $\partial c / \partial u$, $\partial^2 c / \partial p_i \partial p_j$, $\partial^2 c / \partial u \partial p_i$, and $\partial^2 c / \partial u^2$ can be set equal to those of an arbitrary cost function. We take

$$(2) \quad \log a(p) = a_0 + \sum_k \alpha_k \log p_k$$

$$+ \frac{1}{2} \sum_k \sum_j \gamma_{kj}^* \log p_k \log p_j$$

$$(3) \quad \log b(p) = \log a(p) + \beta_0 \prod_k p_k^{\beta_k}$$

so that the *AIDS* cost function is written

$$(4) \quad \log c(u, p) = a_0 + \sum_k \alpha_k \log p_k + \frac{1}{2} \sum_k \sum_j \gamma_{kj}^* \log p_k \log p_j + u \beta_0 \prod_k p_k^{\beta_k}$$

where α_i , β_i , and γ_{ij}^* are parameters. It can easily be checked that $c(u, p)$ is linearly ho-

mogeneous in p (as it must be to be a valid representation of preferences) provided that $\sum_i \alpha_i = 1$, $\sum_j \gamma_{kj}^* = \sum_k \gamma_{kj}^* = \sum_j \beta_j = 0$. It is also straightforward to check that (4) has enough parameters for it to be a flexible functional form provided it is borne in mind that, since utility is ordinal, we can always choose a normalization such that, at a point, $\partial^2 \log c / \partial u^2 = 0$. The choice of the functions $a(p)$ and $b(p)$ in (2) and (3) is governed partly by the need for a flexible functional form. However, the main justification is that this particular choice leads to a system of demand functions with the desirable properties which we demonstrate below.

The demand functions can be derived directly from equation (4). It is a fundamental property of the cost function (see Ronald Shephard, 1953, 1970, or Diewert's 1974 survey paper) that its price derivatives are the quantities demanded: $\partial c(u, p) / \partial p_i = q_i$. Multiplying both sides by $p_i / c(u, p)$ we find

$$(5) \quad \frac{\partial \log c(u, p)}{\partial \log p_i} = \frac{p_i q_i}{c(u, p)} = w_i$$

where w_i is the budget share of good i . Hence, logarithmic differentiation of (4) gives the budget shares as a function of prices and utility:

$$(6) \quad w_i = \alpha_i + \sum_j \gamma_{ij} \log p_j + \beta_i u \beta_0 \prod_k p_k^{\beta_k}$$

where

$$(7) \quad \gamma_{ij} = \frac{1}{2} (\gamma_{ij}^* + \gamma_{ji}^*)$$

For a utility-maximizing consumer, total expenditure x is equal to $c(u, p)$ and this equality can be inverted to give u as a function of p and x , the indirect utility function. If we do this for (4) and substitute the result into (6) we have the budget shares as a function of p and x ; these are the *AIDS* demand functions in budget share form:

$$(8) \quad w_i = \alpha_i + \sum_j \gamma_{ij} \log p_j + \beta_i \log \{x / P\}$$

where P is a price index defined by

$$(9) \quad \log P = \alpha_0 + \sum_k \alpha_k \log p_k + \frac{1}{2} \sum_j \sum_k \gamma_{kj} \log p_k \log p_j$$

The restrictions on the parameters of (4) plus equation (7) imply restrictions on the parameters of the AIDS equation (8). We take these in three sets

$$(10) \quad \sum_{i=1}^n \alpha_i = 1 \quad \sum_{i=1}^n \gamma_{ij} = 0 \quad \sum_{i=1}^n \beta_i = 0$$

$$(11) \quad \sum_j \gamma_{ij} = 0$$

$$(12) \quad \gamma_{ij} = \gamma_{ji}$$

Provided (10), (11), and (12) hold, equation (8) represents a system of demand functions which add up to total expenditure ($\sum w_i = 1$), are homogeneous of degree zero in prices and total expenditure taken together, and which satisfy Slutsky symmetry. Given these, the AIDS is simply interpreted: in the absence of changes in relative prices and "real" expenditure (x/P) the budget shares are constant and this is the natural starting point for predictions using the model. Changes in relative prices work through the terms γ_{ij} ; each γ_{ij} represents 10^2 times the effect on the i th budget share of a 1 percent increase in the j th price with (x/P) held constant. Changes in real expenditure operate through the β_i coefficients; these add to zero and are positive for luxuries and negative for necessities. Further interpretation is best done in terms of the claims made in the introduction.

A. Aggregation Over Households

The aggregation theory developed in Muellbauer (1975, 1976, of which the main relevant points are summarized in the Appendix) implies that exact aggregation is possible if, for an individual household h , behavior is described by the generalization

of (8):

$$(8') \quad w_{ih} = \alpha_i + \sum_j \gamma_{ij} \log p_j + \beta_i \log \{x_h/k_h P\}$$

The parameters k_h can be interpreted as a sophisticated measure of household size which, in principle, could take account of age composition, other household characteristics, and economies of household size; and which is used to deflate the budget x_h to bring it to a "needs corrected" *per capita* level. This allows a limited amount of taste variation across households. The share of aggregate expenditure on good i in the aggregate budget of all households, denoted \bar{w}_i is given by

$$\sum_h p_i q_{ih} / \sum_h x_h \equiv \sum_h x_h w_{ih} / \sum_h x_h$$

which, when applied to (8') gives

$$(8'') \quad \bar{w}_i = \alpha_i + \sum_j \gamma_{ij} \log p_j - \beta_i \log P + \beta_i \left\{ \sum_h x_h \log (x_h/k_h) / \sum_h x_h \right\}$$

Define the aggregate index k by

$$(13) \quad \log(\bar{x}/k) \equiv \sum_h x_h \log(x_h/k_h) / \sum_h x_h$$

where \bar{x} is the average level of total expenditure x_h . Hence (8'') becomes

$$(8''') \quad \bar{w}_i = \alpha_i + \sum_j \gamma_{ij} \log p_j + \beta_i \log(\bar{x}/kP)$$

This is identical in form to (8') and this confirms that under these assumptions aggregate budget shares correspond to the decisions of a rational representative household whose preferences are given by the AIDS cost function (4) and whose budget is given by \bar{x}/k , the "representative budget level."

The index k has an interesting interpretation. If each household had the same tastes ($k_h = 1$, all h), k would be an index of the

equality of the distribution of household budgets. In fact, this index is identical to Theil's (1972) entropy measure of equality Z deflated by the number of households H , where $\log Z = -\sum(x_h/X)\log(x_h/X)$ and X is the aggregate budget; Z reaches its maximum level of H when there is perfect equality so that $x_h = \bar{x}$, all h . Therefore as inequality increases, $k = Z/H$ decreases and the representative budget level increases. When k_h differs across households, for example, because of differences in household composition, the index k reflects not only the distribution of budgets but the demographic structure. Ideally, one might attempt to model the variation of k_h with household characteristics in a cross-section study and, given time-series data on the joint distribution of household budgets and characteristics, construct a series for k for use in fitting (8''). Data limitations have prevented us from carrying out this proposal in the empirical application below. To the extent that k is constant or uncorrelated with \bar{x} or p , no omitted variable bias arises from our procedure of omitting k and redefining $\alpha_i^* = \alpha_i - \beta_i \log k^*$ where k^* is the constant or sample mean value of k .

When the distribution of household budgets and household characteristics is invariant except for equiproportional changes in household budgets, k is constant. In this case there is considerable extra scope for taste variations in the individual demand functions without altering the validity of the representative consumer hypothesis embodied in (8''). Indeed, it turns out that not only α_{ih} , all i , but also γ_{jh} , all i,j , can differ over households. The α_i and γ_{ij} parameters in (8'') are then weighted averages of the micro parameters.

B. Generality of the Model

The flexible functional form property of the AIDS cost function implies that the demand functions derived from it are first-order approximations to any set of demand functions derived from utility-maximizing behavior. The AIDS is thus as general as

other flexible forms such as the translog or the Rotterdam models. However, if maximizing behavior is not assumed but it is simply held that demands are continuous functions of the budget and of prices, then the AIDS demand functions (8) (without the restrictions (11) and (12)) can still provide a first-order approximation. In general, without maximizing assumptions, we can think of the budget shares w_i as being unknown functions of $\log p$ and $\log x$. From (8) and (9), the AIDS has derivatives $\partial w_i / \partial \log x = \beta_i$ and $\partial w_i / \partial \log p_j = \gamma_{ij} - \beta_i \alpha_j - \beta_i \sum \gamma_{jk} \log p_k$ so that, at any point, β and γ can be chosen so that the derivatives of the AIDS will be identical to those of any true model. Given that the α parameters act as intercepts, the AIDS can thus provide a local first-order approximation to any true demand system, whether derived from the theory of choice or not. This property is important since it means that tests of homogeneity of symmetry are set within a maintained hypothesis which makes sense and would be widely accepted in its own right.

Generality is not without its problems, however. There is a large number of parameters in (18) and on most data sets these are unlikely to be all well determined. It is thus important that there should exist some straightforward procedure for eliminating unnecessary parameters without untoward consequences for the properties of the model. In the AIDS, this can be done by placing whatever restrictions on γ_{ij} parameters are thought to be empirically or theoretically plausible. As we shall see below, in many cases it will be possible to impose these restrictions on a single equation basis. One obvious restriction is that for some pairs (i,j) , γ_{ij} should be zero; for such pairs the budget share of each is independent of the price of the other if (x/P) is held constant. It can be shown that γ_{ij} has approximately the same sign as the compensated cross-price elasticity and this is also useful in suggesting prior restrictions. We should not however expect all the γ_{ij} s to be zero; the resulting model, $w_i = \alpha_i + \beta_i \log(x/P)$ is extremely restrictive and has been tested and rejected by Deaton (1978).

C. Restrictions

If we start from equations (8) and (9) as our maintained hypothesis, we can examine the effects of the restrictions (10)–(12) which are required to make the model consistent with the theory of demand. The conditions (10) are the *adding-up* restrictions; as can easily be checked from (8), these ensure that $\sum w_i \equiv 1$. *Homogeneity* of the demand functions requires restriction (11) which can be tested equation by equation. Slutsky *symmetry* is satisfied by (8) if and only if the symmetry restriction (12) holds. As is true of other flexible functional forms, *negativity* cannot be ensured by any restrictions on the parameters alone. It can however be checked for any given estimates by calculating the eigenvalues of the Slutsky matrix s_{ij} , say. In practice, it is easier to use not s_{ij} but $k_{ij} = p_i p_j s_{ij} / x$, the eigenvalues of which have the same signs as those of s_{ij} and which are given by

$$(14) \quad k_{ij} = \gamma_{ij} + \beta_i \beta_j \log \frac{x}{P} - w_i \delta_{ij} + w_i w_j$$

where δ_{ij} is the Kronecker *delta*. Note that apart from this negativity condition, all the restrictions are expressible as linear constraints involving only the parameters and so can be imposed globally by standard techniques.

D. Estimation

In general, estimation can be carried out by substituting (9) in (8) to give

$$(15) \quad w_i = (\alpha_i - \beta_i \alpha_0) + \sum_j \gamma_{ij} \log p_j + \beta_i \left\{ \log x - \sum_k \alpha_k \log p_k - \frac{1}{2} \sum_k \sum_j \gamma_{kj} \log p_k \log p_j \right\}$$

and estimating this non-linear system of equations by maximum likelihood or other methods with and without the restrictions (11) and (12). (Note that since the data add up by construction, (10) is not testable.) Equation (15) is not particularly difficult to estimate since the first-order conditions for

likelihood maximization are linear in α and γ given β and vice versa so that “concentration” allows iteration on a subset of the parameters (see for example, Deaton, 1975, pp. 46–49). Although all the parameters in (15) are identified given sufficient variation in the independent variables, in many examples the practical identification of α_0 is likely to be problematical. This parameter is only identified from the α_i s in (15) by the presence of these latter inside the term in braces, originally in the formula for $\log P$, equation (9). However, in situations where individual prices are closely collinear, $\log P$ is unlikely to be very sensitive to its weights so that changes in the intercept term in (15) due to variations in α_0 can be offset in the α s with minimal effect on $\log P$. This can be overcome in practice by assigning a value to α_0 a priori. Since the parameter can be interpreted as the outlay required for a minimal standard of living when prices are unity (usually in the base year; see the Appendix), choosing a plausible value is not difficult.

However, in many situations, it is possible to exploit the collinearity of the prices to yield a much simpler estimation technique. Note from (8) that if P were known, the model would be linear in the parameters α , β , and γ , and estimation (at least without cross-equation restrictions such as symmetry) can be done equation by equation by *OLS* which, in this case and given normally distributed errors, is equivalent to maximum likelihood estimation for the system as a whole. The adding-up constraints (10) will be automatically satisfied by these estimates. In situations where prices are closely collinear, it may well be adequate to approximate P as proportional to some known index P^* , say. One obvious candidate in view of (8) and (9) is Stone's (1953) index $\log P^* = \sum w_k \log p_k$. If $P \approx \phi P^*$ say, then (8) can be estimated as

$$(16)$$

$$w_i = (\alpha_i - \beta_i \log \phi) + \sum_j \gamma_{ij} \log p_j + \beta_i \log \left(\frac{x}{P^*} \right)$$

Note that in this framework the α_i parameters are identified only up to a scalar multiple of β_i ; if we write $\alpha_i^* = \alpha_i - \beta_i \log \phi$, it is

easily seen that $\sum \alpha_k^* = 0$ is still required for adding up, since $\sum \beta_k = 0$.

In the empirical results below we shall estimate both (15) and (16), and show that the latter is an excellent approximation to the former. However, it must be emphasized that (16) exists only as an approximation to (15) and will only be accurate in specific circumstances, albeit widely occurring ones in time-series estimation. Note finally that if single equation estimation is used to investigate likely restrictions amongst the γ parameters, as is suggested in Part B above, the *constrained OLS* estimates will no longer automatically be maximum likelihood, efficient, or satisfy adding up. Hence, once the restrictions have been selected, (15) should be used to reestimate the whole system.

E. Relationship with Budget Studies and with the Rotterdam Model

The Engel curves corresponding to (8) take the form $p_i q_i = \xi_i x + \beta_i x \log x$ for appropriate functions of prices ξ_i . These are clearly not linear except in the proportional case when $\beta_i = 0$. The model thus allows a possible reconciliation between time-series models, which have to date required linearity of Engel curves for aggregation with cross-section results, which typically find evidence of nonlinearity. Indeed, the *PIG-LOG* Engel curve $w_i = \xi_i + \beta_i \log x$ was used as early as 1943 by Holbrook Working and has recently been recommended by Claus Leser (1963, 1976) as providing an excellent fit to cross-section data in a wide range of circumstances.

In the time-series context, the *AIDS* has a close relationship to the Rotterdam model of Theil (1965, 1976) and Barten. The first difference form of (8) is

$$(17) \quad \Delta w_i = \beta_i \Delta \log\left(\frac{x}{P}\right) + \sum_j \gamma_{ij} \Delta \log p_j$$

which no longer involves the α parameters except through $\Delta \log P$. This dependence can

be seen by writing (17) in full, i.e.,

$$(18) \quad \Delta w_i = \beta_i \left\{ \Delta \log x - \sum_k \alpha_k \Delta \log p_k \right. \\ \left. - \frac{1}{2} \sum_k \sum_j \gamma_{kj} \Delta (\log p_j \log p_k) \right\} + \sum_j \gamma_{ij} \Delta \log p_j$$

Again, all the parameters (except α_0) are theoretically identified, but in practice the substitutability between γ_{ij} s and $\beta_i \alpha_j$ s in fitting (18) if prices are nearly collinear means that, in such cases, the only practical way of estimating (17) is to replace $\Delta \log P$ by some index, for example, $\Delta(\sum w_k \log p_k)$ as before or by its approximation $\sum w_k \Delta \log p_k$. In the latter case, the right-hand side of (17) becomes identical to the right-hand side of the Rotterdam model which is

$$(19) \quad w_i \Delta \log q_i = b_i \{ \Delta \log x - \sum w_k \Delta \log p_k \}$$

$$+ \sum_j c_{ij} \Delta \log p_j$$

The dependent variable is different in the *AIDS*; instead of $w_i \Delta \log q_i$ we have Δw_i or $w_i \Delta \log w_i$. Thus, by replacing the dependent variable $w_i \Delta \log q_i$ in the Rotterdam model by $w_i \Delta \log w_i$, an addition of $w_i \Delta \log(p_i/x)$, we generate the first-difference form of the *AIDS*. The similarity between the two models is quite striking in this form; both are effectively linear and both can be used to test homogeneity and symmetry with only linear restrictions on constant parameters. Note however that the parameters have quite different interpretations in the two models so that, for example, the negativity condition applies directly to the matrix of price effects in the Rotterdam model which is not the case for the *AIDS*. The crucial difference between the two models is that (17), unlike (19), is derived from explicit demand functions, (8), and an explicit characterization of preferences, (4). For the prediction of demand this difference may not be vitally important, but in many other contexts, for example in calculating cost-of-living indices, household equivalence scales, or optimal tax rates, the ability to link

estimated parameter values to preferences themselves becomes of great significance.

II. An Application to Postwar British Data

In this section we estimate the model using annual British data from 1954 to 1974 inclusive on eight nondurable groups of consumers' expenditure, namely, food, clothing, housing services, fuel, drink and tobacco, transport and communication services, other goods, and other services. As discussed in Section I, Part A above, if we assume that the index k in (8'') is either constant or that its deviations are independently distributed from those of the average budget \bar{x} and of prices, no biases result from its omission. In particular, we allow the intercepts in (8'') to absorb the $-\beta_i \log k$ terms. We then proceed by first following the strategy outlined in Section I, Part D, setting $\log P^* = \sum w_k \log p_k$ for each year and estimating equation (16) for each good separately by OLS. The system is then reestimated, equation by equation, and again using P^* , in order to test the homogeneity condition. Equation (11) is imposed by substitution so that instead of (16), we estimate

$$(20) \quad \bar{w}_i = \alpha_i^* + \sum_{j=1}^{n-1} \gamma_{ij} \log\left(\frac{p_j}{p_n}\right) + \beta_i \log\left(\frac{\bar{x}}{P^*}\right)$$

At this stage, F -ratios are calculated equation by equation to test the validity of the restriction.

The next stage is to impose symmetry of γ , at this point replacing P^* by the "correct" price index (9) with α_0 set to some appropriate value. Since symmetry, unlike homogeneity or the unrestricted model, involves cross-equation restrictions, the variance-covariance matrix of the residuals for the first time plays a part in the estimation. Since this is unknown a priori, normal practice would be to replace it by its maximum likelihood estimate. However, with only twenty-one observations, this is not practicable for equation (15) since, with so many parameters in each equation, the likelihood can be made arbitrarily large by making any

one equation fit perfectly.¹ This difficulty can only be resolved by assuming a particular structure for the variance-covariance matrix of the residuals. Following Deaton (1975, p. 39), we assume $V = \sigma^2(I - ii')$, where V is the variance-covariance matrix of the residuals, σ^2 is a (positive) parameter to be estimated, I is an $n \times n$ identity matrix and i is a vector each of the elements of which is $(n)^{-1/2}$. In this case, maximum-likelihood estimation reduces to least squares so that instead of minimizing the determinant of the matrix of residual cross products, we minimize its trace. The likelihood values quoted below are calculated on this assumption. Once again, maximum use is made of substitution in the estimation so that, under symmetry, (15) is estimated using only the fourteen independent α_i^* 's and β_i 's and the twenty-eight parameters forming the upper right-hand triangle of γ with its final row and column deleted. We now check that P^* and P are sufficiently close to allow comparison of likelihoods both by direct evaluation of both indices and by reestimation of the unrestricted and homogeneous models using P as evaluated from the symmetric estimates. It is also possible to check concavity at this stage by using the symmetric parameter estimates to calculate the eigenvalues of the matrix in (14). Finally, the whole process is repeated with the model written in first differences, that is, equation (17) with the addition of intercepts. Collinearity prevented any successful attempt to link P to the parameter estimates in these regressions; instead, the value of P calculated from the symmetric estimation in levels was used throughout.

Note that we choose to test symmetry whether or not homogeneity is rejected. This procedure has been criticized by Grayham Mizon who suggests that optimal inference requires that further testing be abandoned as soon as a rejection is encountered. Mizon's criticism would be correct if we were certain of our maintained hypothesis, but to some extent this is a matter of choice. Many economists would choose not to test

¹We are grateful to Teun Kloek for pointing this out to us.

TABLE 1—THE UNCONSTRAINED PARAMETER ESTIMATES AND TESTS OF HOMOGENEITY
(*t*-Values in Parentheses)

Commodity <i>i</i>	α_i^*	β_i	γ_{i1}	γ_{i2}	γ_{i3}	γ_{i4}	γ_{i5}	γ_{i6}	γ_{i7}	γ_{i8}	$\sum_j \gamma_{ij}$	(10^{-2})	R^2	D.W.
Food	1.221	-0.160	0.186	-0.077	-0.013	-0.020	-0.058	0.032	0.015	-0.098	-0.033	.113	0.999	2.33
	(7.4)	(-6.1)	(9.8)	(-4.3)	(-0.8)	(-1.1)	(-6.2)	(1.3)	(0.7)	(-4.2)	(-4.4)	.180	0.998	1.74
Clothing	-0.482	0.091	0.033	0.016	-0.024	-0.026	-0.029	0.014	0.033	-0.049	-0.032	.106	0.984	2.29
	(-3.1)	(3.7)	(1.8)	(1.0)	(-1.6)	(-1.5)	(-3.3)	(0.6)	(1.6)	(-2.2)	(-4.5)	.171	0.955	1.55
Housing	0.793	-0.104	-0.082	-0.9	0.088	0.9	0.033	-0.055	-0.030	0.098	0.051	.086	0.999	1.89
	(6.3)	(-5.1)	(-5.6)	(-0.7)	(7.2)	(0.7)	(4.7)	(-2.9)	(-1.8)	(5.5)	(9.1)	.241	0.992	1.29
Fuel	-0.159	0.033	-0.042	0.010	-0.011	0.037	-0.004	0.022	0.007	-0.031	-0.010	.140	0.883	2.25
	(-0.8)	(1.0)	(-1.8)	(0.4)	(-0.5)	(1.6)	(-0.3)	(0.7)	(0.3)	(-1.1)	(-1.1)	.141	0.870	2.03
Drink and Tobacco	-0.043	0.028	-0.043	0.034	-0.027	-0.020	0.056	0.005	-0.018	0.014	0.001	.099	0.969	2.96
	(-0.3)	(1.2)	(-2.6)	(2.2)	(-1.9)	(-1.2)	(6.9)	(0.2)	(-0.9)	(0.7)	(0.0)	.095	0.969	2.93
Transport and Communication	-0.061	0.029	-0.022	-0.012	-0.002	0.011	0.060	-0.023	-0.024	0.053	0.040	.047	1.000	2.24
	(-0.9)	(2.6)	(-2.7)	(-1.6)	(-0.3)	(1.4)	(15.2)	(-2.2)	(-2.6)	(5.3)	(13.1)	.184	0.992	1.36
Other Goods	-0.038	0.022	0.001	-0.003	-0.001	-0.006	-0.030	0.007	0.032	-0.006	-0.005	.108	0.885	1.92
	(-0.2)	(0.9)	(0.0)	(-0.2)	(-0.0)	(-0.3)	(-3.4)	(0.3)	(1.5)	(-0.2)	(0.7)	.106	0.880	1.91
Other Services	-0.231	0.060	-0.032	0.041	-0.011	0.014	-0.028	-0.003	-0.015	0.019	-0.014	.107	0.843	2.27
	(-1.5)	(2.4)	(-1.8)	(2.4)	(-0.7)	(0.8)	(-3.1)	(-0.1)	(-0.7)	(0.9)	(-2.0)	.119	0.788	1.98

homogeneity, treating absence of money illusion as a maintained hypothesis; the test of symmetry would then be the interesting one. Even if the maintained hypothesis turns out to be false, tests based on it are not necessarily without interest. Few if any tests in econometrics are carried out within the framework of maintained hypotheses which are even widely accepted, let alone of unchallengeable validity.

Table 1 reports the first-stage estimates of (16) using P^* and without any constraints on the parameters save (10) which are automatically and costlessly satisfied. The estimates of β classify food and housing as necessities while the other goods are luxuries. A large number of γ coefficients are significantly different from zero; twenty-two out of sixty-four have *t*-values absolutely larger than 2. Even so, none of the variables considered have any detectable effect on the value share for fuel and very few have influence in the other goods or other services equations. Similarly, the prices of fuel, of transport and communication, and of other services have little or no effect anywhere (except, of course, through P^* and the value share itself), while the prices of food, drink and tobacco, and of other services appear with considerable regularity. The total expenditure and own-price elasticities are shown in the first two col-

umns of Table 2 and, although food has an (insignificant) positive price elasticity, these numbers appear both credible and in line with other studies. Note the general price inelasticity of demand; only transport and communication appear to be price elastic.

Table 1 also shows, in the column headed $\sum_j \gamma_{ij}$, the row sums of the unconstrained γ_{ij} matrix; this number shows 10^2 times the absolute effect on each value share of a 1 percent increase in all prices and total expenditure. Under homogeneity, this should be zero and the bracketed numbers given are *t*-tests of the significance of the deviation from zero. These numbers are, of course, identical to the square roots of the *F*-ratios obtained by comparing the residual sums of squares of equations (16) and (20). Hence, a proportional increase in prices and expenditure will decrease expenditure on food and on clothing, and increase expenditure on housing and transport and communication. These are also the commodities for which the elasticities suffer the largest changes between columns 1 and 2 and columns 3 and 4 of Table 2. Other deviations from homogeneity appear not to be significant. The final columns of Table 1 give equations standard errors, the R^2 and Durbin-Watson (*D.W.*) statistics for free and restricted estimation. Note that for the four commodity groups where homogeneity

TABLE 2—TOTAL EXPENDITURE AND OWN-PRICE ELASTICITIES

	Levels Model				First-Differences Model			
	Unconstrained e_i	Homogeneous e_{ii}	Unconstrained e_i	Homogeneous e_{ii}	Unconstrained e_i	Homogeneous e_{ii}	Unconstrained e_i	Homogeneous e_{ii}
Food	0.21	0.07	0.04	-0.01	0.04	0.22	0.17	-0.00
Clothing	2.00	-0.92	1.51	-0.48	2.83	-0.94	2.92	-0.94
Housing	0.30	-0.31	0.79	-0.16	0.04	-0.31	-0.02	-0.30
Fuel	1.67	-0.28	1.37	0.10	1.00	0.00	0.86	-0.08
Drink and Tobacco	1.22	-0.60	1.22	-0.62	1.37	-0.67	1.36	-0.68
Transport and Communication	1.23	-1.21	1.73	-0.92	1.14	-1.23	1.05	-1.17
Other goods	1.21	-0.72	1.15	-0.77	2.03	-0.52	1.92	-0.47
Other services	1.40	-0.93	1.28	-0.78	1.03	-0.78	1.06	-0.74

is rejected, the *D.W.* statistic shows a sharp fall in each case.

The failure of homogeneity is not a new result (see, for example, Barten; Ray Byron; Deaton, 1974a), and can be ascribed to a number of possible causes. However, as far as we are aware, the introduction of serial correlation through the imposition of homogeneity is a result which has not been previously remarked, although it may have been implicit in earlier work. There are a number of plausible explanations for this phenomenon. For example, expenditure on several items may be relatively inflexible in the short run; housing is the obvious case here. The explanation of such items may require other variables such as stocks, lagged dependent variables, or time trends which can perhaps be proxied by the absolute price level. The omission of such variables will thus lead to a rejection of homogeneity associated with an introduction of serial correlation in the residuals of the restricted equations. In principle one could easily include such conditioning variables in the AIDS cost function, for example by allowing the α s to vary linearly with them, and this is likely to be an important topic for future research. A second explanation is the omission of price expectations—the argument advanced by Deaton (1977a) would suggest that factors such as the frequency of purchase for different goods will be relevant in assessing the response of expenditures to changes in price, especially when there is rapid relative or absolute price change. A third possibility, suggested by the

discussion of aggregation in Section I, Part A above, is that it may be incorrect to assume that k , the index reflecting the distribution of household budgets and demographic structure, is independent of the average budget and the price vector. Finally, the assumption of weak intertemporal separability of nondurable goods in the intertemporal utility function, which is required to justify the conventional static utility-maximizing model, may be inappropriate. It is not difficult to construct other models which produce the result and without extensive further empirical work it is extremely difficult to discriminate between them.

In moving to the symmetric estimates (not reported here), in which P^* is replaced by P , we must first check the closeness of the approximation. Table 3 reproduces the two series, $P^* = \exp\{\sum w_k \log p_k\}$ and P scaled to be unity in the base year, i.e., $\exp\{\sum \alpha_k \log p_k + \frac{1}{2} \sum \sum \gamma_{kj} \log p_k \log p_j\}$ evaluated at the symmetric parameter estimates. Both series are based on unity in 1970. Clearly the differences are small; the absolute magnitude of the difference is never greater than .008.

The reestimation of the unconstrained and homogeneous models using P rather than P^* confirmed the empirical unimportance of the difference. Both sets of likelihoods are given in Table 4. It must be reemphasized that these findings are conditional on the kind of relative price movements that took place in our sample. However, even if relative price changes had been greater, the results suggest that the proce-

TABLE 3—COMPARISON OF PRICE INDICES

	P^*	P		P^*	P		P^*	P
1954	0.566	0.571	1961	0.684	0.686	1968	0.894	0.888
1955	0.587	0.595	1962	0.712	0.715	1969	0.946	0.944
1956	0.611	0.617	1963	0.729	0.730	1970	1.000	1.000
1957	0.631	0.636	1964	0.754	0.758	1971	1.084	1.084
1958	0.648	0.653	1965	0.793	0.797	1972	1.161	1.161
1959	0.655	0.661	1966	0.827	0.830	1973	1.271	1.279
1960	0.663	0.666	1967	0.851	0.852	1974	1.465	1.461

TABLE 4—COMPARATIVE VALUES OF 2 LOG LIKELIHOOD

	Levels		First Differences	
	Using P^*	Using P	Using P^*	Using P
Unrestricted	1722.5	1723.8	1560.0	1560.3
Homogeneous	1579.6	1585.1(7)	1546.6	1547.9(7)
Symmetric	—	1491.0(21)	—	1508.8(21)

Note: Number of restrictions in parentheses. Numbers can only be compared within columns, not between levels and first differences.

dure of starting with P^* , calculating OLS regressions, computing a new P , and repeating will be a computationally efficient way of obtaining good estimates of the full nonlinear system.

Symmetry, unlike homogeneity, cannot be tested on an equation-by-equation basis and we must rely on a large-sample likelihood-ratio test for the system as a whole. For comparison, twice the logarithm of the likelihood is 1722.5 for the unconstrained system, 1579.6 for the homogeneous model (a fall which reflects the individual restrictions) and this falls further to 1491.0 under symmetry. Since symmetry embodies twenty-one constraints over and above the seven of homogeneity, the restriction is rejected on an asymptotically valid χ^2 -test whether or not the maintained hypothesis is taken to contain homogeneity. Once again, this is consistent with earlier results although rejection of symmetry given homogeneity is not always clear-cut in the studies cited above. The interpretation of the rejection is not clear without some convincing explanation of the lack of homogeneity. Without this, it is impossible to know whether or not we should expect symmetry to hold. For exam-

ple, it is possible to introduce habits into demand functions so that, if they are allowed for, symmetry can be expected to hold, while if ignored, symmetry will be destroyed.

The full set of symmetric parameter estimates are not included here for reasons of space. The most interesting property of these, apart from symmetry, is in their implications for negativity. To assess this, the K matrix of equation (14) was evaluated for each year in the sample and its eigenvalues calculated. One of these is identically zero and, for concavity, the others should be negative. Contrary to this, we found one positive eigenvalue for the early part of the period, increasing to two by the end. The most obvious symptom of nonconcavity in the symmetric estimates was an estimated positive compensated own-price elasticity for fuel throughout the sample period. This may seem to be of limited importance given that the symmetric homogeneous model has already been rejected; if the cost function doesn't exist, why worry about its concavity? However, for several reasons it would be extremely useful to have parameter estimates for a reasonably general concave

homogeneous cost function. For example, we frequently wish to calculate price and quantity index numbers or to use optimal taxation formulae to derive numerical values for tax rates. All such calculations require numerical estimates of cost functions and, if they are to make any sense at all, these cost functions must be both homogeneous and concave. Consequently, in cases where empirical estimates of demand equations have been used in applied welfare analysis, the linear expenditure system has invariably been used; see, for example, Anthony Atkinson and Joseph Stiglitz, Muellbauer (1974), or Deaton (1977b). With the linear expenditure system, the model is so restrictive that concavity of the cost function is virtually guaranteed provided inferior goods do not appear. But this restrictiveness is also known to be empirically false (see, for example, Deaton, 1974b or 1978) so that it would be of considerable value to have estimates of a concave cost function which allowed considerably more substitution than does the linear expenditure system. Consequently, it will be of considerable interest in future work to attempt to restrict the parameters further so that the estimated cost function is concave.

Finally, we turn to the estimation of the model in first-difference form. Here we use equation (17) plus intercepts, i.e.,

$$(21) \quad \Delta w_i = \eta_i + \beta_i \Delta \log \left(\frac{\bar{x}}{P} \right) + \sum_k \gamma_{ik} \Delta \log p_k$$

where the constants η_i are introduced primarily for econometric reasons but, if significant, would imply time trends in the original model which expresses the variables in levels. The P is taken as in Table 3. In these regressions homogeneity is only rejected for food and for transport and communication; clothing and housing, which rejected homogeneity in the earlier regressions, now yield insignificant F -ratios. Closer inspection reveals that for both these cases, the constant term η_i , which is insignificant without constraints, becomes significant when homogeneity is imposed. Similarly, for transport and communication η_i

becomes significant when homogeneity is imposed, but in this case the F -ratio remains significant. This would support our earlier conjectures as to the possible role of time trends, stocks, or other omitted variables in explaining nonhomogeneity. Likewise, in Table 2 the expenditure elasticities from the first-difference model tend to be higher than the levels estimates when stock effects are likely to be important (clothing, other goods) and lower when one would expect short-run total expenditure effects to be limited (food, housing, transport). Otherwise, the first-difference parameter estimates, homogeneous or unconstrained, are rather close to the values originally obtained. As with levels, tests of concavity with the first-difference model revealed several violations. The likelihoods for the two models are summarized in Table 4; the fact that homogeneity cannot be rejected overall at the 5 percent level reflects the importance of the time trends in the housing, clothing, and transport and communication equations. Note too that in this case, from the last column of the table, symmetry is only just rejected given homogeneity. Hence, if we make some allowance for the asymptotic nature of the test, these final results would suggest that the introduction of (arbitrary) time trends removes much of the conflict between the data and the hypothesis of a representative consumer maximizing a conventional static utility function.

III. Summary and Conclusions

In this paper we have introduced a new system of demand equations, the *AIDS*, in which the budget shares of the various commodities are linearly related to the logarithm of real total expenditure and the logarithms of relative prices. The model is shown to possess most of the properties usually thought desirable in conventional demand analysis, and to do so in a way not matched by any single competing system. Fitted to postwar British data, the *AIDS* is capable of explaining a high proportion of the variance of the commodity budget shares but, unless allowance is made for omitted variables by the arbitrary use of

time trends, does so in a way which is inconsistent with the hypothesis of consumers making decisions according to the model's demand functions governed by the conventional static budget constraint. These results suggest that influences other than current prices and current total expenditure must be systematically modelled if even the broad pattern of demand is to be explained in a theoretically coherent and empirically robust way. Whether these developments generalize the static framework by including stock effects, errors in price perceptions, or by going beyond the assumption of weak intertemporal separability on which the static model rests, we believe that the AIDS, with its simplicity of structure, generality, and conformity with the theory, offers a platform on which such developments can proceed.

APPENDIX: AIDS IN THE CONTEXT OF AGGREGATION THEORY

In Muellbauer (1975, 1976), a definition of the existence conditions for a representative consumer is given which allows more general behavior than the parallel linear Engel curves which are required if average demands are to be a function of the average budget. We know that in general, the average budget share

$$\bar{w}_i = \sum_h p_i q_{ih} / \sum_h x_h \equiv \sum_h x_h w_{ih} / \sum_h x_h$$

is a function of prices and the complete distribution vector (x_1, x_2, \dots, x_H) . A representative consumer exists in Muellbauer's sense if each \bar{w}_i can be written as a function of prices and the same single scalar x_0 , itself a function of prices and the distribution vector. This scalar, which can be thought of as marking a position in the distribution of xs, is the representative budget level. Muellbauer shows that for an x_0 to exist such that

(A1)

$$\sum_h x_h w_{ih}(x_h, p) / \sum_h x_h = w_i \{ x_0(x_1, \dots, x_H, p), p \}$$

the individual budget share equations must

have the "generalized linear" (GL) form:

(A2)

$$w_{ih}(x_h, p) = v_h(x_h, p) A_i(p) + B_i(p) + C_{ih}(p)$$

where v_h , A_i , B_i , and C_{ih} are functions satisfying $\sum_i A_i = \sum_i C_{ih} = \sum_h C_{ih} = 0$ and $\sum B_i = 1$. Clearly, (A1) goes beyond the usual formulation of $x_0 = \bar{x}$, and, as we shall see below, allows us to incorporate into the demand functions features of the expenditure distribution other than the mean.

Of particular interest is the case where x_0 is independent of prices, depending only on the individual xs. This occurs if, and only if, the v_h function in (A2) restricts to

$$(A3) \quad v_h(x_h, p) = \{1 - (x_h/k_h)^{-\alpha}\}^{\alpha^{-1}}$$

where α is a constant and k_h , although not a function of x_h , and p is free to vary from household to household. In this case, the budget shares are said to have the "price-independent generalized linear" form (PIGL). Note the special case of (A3) as $\alpha \rightarrow 0$, i.e.,

$$(A4) \quad v_h(x_h, p) = \log(x_h/k_h)$$

For obvious reasons, this is referred to as the PIGLOG case. By substituting (A3) in (A2), (A1) can be used to give an explicit form for x_0 , viz.,

$$(A5) \quad x_0 = \left\{ \sum \left(\frac{x_h}{k_h} \right)^{-\alpha} / \sum x_h \right\}^{-1/\alpha}$$

If we assume that individual behavior is preference consistent, the cost function corresponding to PIGL takes the form

$$(A6) \quad \{c(u_h, p)/k_h\}^\alpha$$

$$= (1 - u_h) \{a(p)\}^\alpha + u_h \{b(p)\}^\alpha$$

which as α tends to zero takes the PIGLOG form

$$(A7) \quad \log \{c(u_h, p)/k_h\}$$

$$= (1 - u_h) \log \{a(p)\} + u_h \log \{b(p)\}$$

where $a(p)$ and $b(p)$ are linear homogeneous concave functions, α is the constant

parameter of (A3), and (with some exceptions discussed below) $0 < u < 1$. The quantity k_h can be used to allow for family composition effects within *PIGL*; for the standard or "reference" household k_h is unity.

Since the *AIDS* is a member of the *PIGLOG* family, and hence of *PIGL*, we can achieve maximum generality by discussing some of the important properties of this class. If, omitting the household subscript, we write q_i for the quantity demanded of good i , then, by the derivative property of the cost function, $q_i = \partial c / \partial p_i$ so that $w_i = p_i q_i / x = \partial \log c / \partial \log p_i$. Hence from (A6), taking $k_h = 1$, and differentiating

$$(A8) \quad \alpha c^\alpha \frac{\partial \log c}{\partial \log p_i} = \alpha a^\alpha a_i (1 - u) + u a b^\alpha b_i$$

where $a_i = \partial \log a / \partial \log p_i$ and $b_i = \partial \log b / \partial \log p_i$. Hence, substituting x for c ,

$$(A9) \quad w_i = (1 - u) \left(\frac{a}{x} \right)^\alpha a_i + u \left(\frac{b}{x} \right)^\alpha b_i$$

where, from (A6), $u = (x^\alpha - a^\alpha) / (b^\alpha - a^\alpha)$ or from (A7), $u = (\log x - \log a) / (\log b - \log a)$. Similarly, when $\alpha = 0$,

$$(A10) \quad w_i = (1 - u) a_i + u b_i$$

Equations (A9) and (A10) have attractive interpretations. Cost $c(u, p)$ is increasing in utility as long as $b(p)$ is greater than $a(p)$ —note that this does not depend on the sign of α —so that as u increases from 0 to 1, $c(u, p)$ increases from $a(p)$ to $b(p)$ with w_i moving from a_i to b_i . Hence a total expenditure of $a(p)$ can be thought of as "poverty" expenditure with associated expenditure pattern a_i , while $b(p)$ is "affluence" expenditure with budget shares b_i . On this interpretation $x = a(p)$ and $x = b(p)$ are the equations of the tangents to the poverty and affluence indifference curves, respectively, $u = 0$ and $u = 1$. From (A6) therefore, we see that the tangent to the indifference curve actually attained is the mean of order α of poverty and affluence tangents, the weights depending on the welfare level or

outlay of the household. This averaging is even more obvious in the value share equations (A9) and (A10). Since $(1 - u)(a/x)^\alpha$ and $u(b/x)^\alpha$ sum to unity, as do $(1 - u)$ and u , these equations give the actual budget shares as weighted averages of a_i and b_i .

Since the value shares of luxuries increase with total outlay and hence with u , we can characterize luxuries and necessities simply by whether b_i is greater than or less than a_i . Inferior goods are not excluded under *PIGL* and it is straightforward to construct examples from both (A9) and (A10).

Note finally that there are restrictions on the possible set of x and p over which the cost function and the associated demands are valid. One set of restrictions is implied by the necessity that, for all i , $0 < w_i \leq 1$. The upper bound is relevant when $b_i > a_i$ and implies that $u = (x^\alpha - a^\alpha) / (b^\alpha - a^\alpha) \leq \min_i \{ (1 - a_i(a/x)^\alpha) / ((b/x)^\alpha b_i - (a/x)^\alpha a_i) \}$. The lower bound, relevant when $b_i < a_i$ requires similarly that $u = (x^\alpha - a^\alpha) / (b^\alpha - a^\alpha) \geq \max_i \{ (a/x)^\alpha a_i / ((a/x)^\alpha a_i - (b/x)^\alpha b_i) \}$. The second set of restrictions are those required to ensure that the cost function is concave. From (A6), we can see that a sufficient condition for concavity of $c(u, p)$ is that $a(p)$ and $b(p)$ be concave and that $0 < u < 1$. However, this is by no means necessary. If $b(p)$ is "more concave" than $a(p)$, then $c(u, p)$ is concave for $u > 0$ and for a range of $u > 1$. It can be shown that the *PIGL* cost function is concave for all $x > 0$, $p > 0$ if and only if $a_i = b_i$ for all i , and $a(p)$ and $b(p)$ are concave. In this not very interesting case, preferences are homothetic.

The practical application of the *PIGL* class requires selection of specific functional forms for the functions $a(p)$ and $b(p)$; those leading to the *AIDS* have been discussed in the text. However, the *PIGL* class is related to two other well-known models. Note first that if $\alpha = 1$, and $k_h = 1$, (A6) becomes the Gorman polar form. The *PIGL* class thus includes all models with linear Engel curves, for example, linear expenditure system, the quadratic utility function, as special cases. Perhaps less obviously, a weakly restricted form of the indirect translog is also *PIGLOG*. From Jorgenson and Lau the translog

indirect utility function is

$$(A11) \quad u = a_0 + \sum_i \alpha_i \log\left(\frac{p_i}{x}\right) + \frac{1}{2} \sum_i \sum_j \beta_{ij} \log\left(\frac{p_i}{x}\right) \log\left(\frac{p_j}{x}\right)$$

where we can choose $\sum \alpha_i = -1$ and $\beta_{ij} = \beta_{ji}$ as arbitrary normalizations. Write $\sum_k \beta_{ki} = \beta_{Mi}$, then if we impose the additional restriction that $\sum_i \beta_{Mi} = 0$, (A11) solves explicitly for $\log c(u, p)$ to

$$(A12) \quad \log c(u, p) = \frac{u - \frac{1}{2} \sum \beta_{Mi} \log p_i \log p_j - \sum \alpha_i \log p_i - a_0}{1 + \sum_i \beta_{Mi} \log p_i}$$

This is of the general form $\log c(u, p) = \log a(p) + u / \log h(p)$ for appropriate choice of $a(p)$ and $h(p)$. Using $\log h(p) = 1 / \log\{b(p)/a(p)\}$ to define $b(p)$ and substituting, we see that (A12) is identical to (A7). Hence, in this case ($\sum_i \beta_{Mi} = 0$) and in this case only, the indirect translog allows consistent aggregation. No such result holds for any interesting subcase of the direct translog.

REFERENCES

- A. B. Atkinson and J. E. Stiglitz**, "The Structure of Indirect Taxation and Economic Efficiency," *J. Publ. Econ.*, Apr. 1972, 1, 97-119.
- A. P. Barten**, "Maximum Likelihood Estimation of a Complete System of Demand Equations," *Euro. Econ. Rev.*, Fall 1969, 1, 7-73.
- E. R. Berndt, M. N. Darrough, and W. E. Diewert**, "Flexible Functional Forms and Expenditure Distributions: An Application to Canadian Consumer Demand Functions," *Int. Econ. Rev.*, Oct. 1977, 18, 651-75.
- R. P. Byron**, "A Simple Method for Estimating Demand Systems under Separable Utility Assumptions," *Rev. Econ. Stud.*, Apr. 1970, 37, 261-74.
- L. R. Christensen, D. W. Jorgenson, and L. J. Lau**, "Transcendental Logarithmic Utility Functions," *Amer. Econ. Rev.*, June 1975, 65, 367-83.
- A. S. Deaton**, (1974a) "The Analysis of Consumer Demand in the United Kingdom, 1900-1970," *Econometrica*, Mar. 1974, 42, 351-67.
- _____, (1974b) "A Reconsideration of the Empirical Implications of Additive Preferences," *Econ. J.*, June 1974, 84, 338-48.
- _____, *Models and Projections of Demand in Post-War Britain*, London 1975.
- _____, (1977a) "Involuntary Saving through Unanticipated Inflation," *Amer. Econ. Rev.*, Dec. 1977, 67, 899-910.
- _____, (1977b) "Equity, Efficiency and the Structure of Indirect Taxation," *J. Publ. Econ.*, Apr. 1977, 8, 299-312.
- _____, "Specification and Testing in Applied Demand Analysis," *Econ. J.*, Sept. 1978, 88, 524-36.
- W. E. Diewert**, "An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function," *J. Polit. Econ.*, May/June 1971, 79, 481-507.
- _____, "Applications of Duality Theory," in Michael D. Intriligator and David A. Kendrick, eds, *Frontiers of Quantitative Economics*, Vol. 2, Amsterdam 1974, ch. 3.
- D. W. Jorgenson and L. J. Lau**, "The Structure of Consumer Preferences," *Annals Econ. Soc. Measure.*, Winter 1975, 4, 49-101.
- C. E. V. Leser**, "Forms of Engel Functions," *Econometrica*, Oct. 1963, 31, 694-703.
- _____, "Income, Household Size and Price Changes 1953-1973," *Oxford Bull. Econ. Statist.*, Feb. 1976, 38, 1-10.
- G. E. Mizon**, "Inferential Procedures in Non-Linear Models: An Application in a U.K. Industrial Cross-Section Study of Factor Substitution and Returns to Scale," *Econometrica*, July 1977, 45, 1221-42.
- J. Muellbauer**, "Recent U.K. Experience of Prices and Inequality: An Application of True Cost of Living and Real Income Indices," *Econ. J.*, Mar. 1974, 84, 32-55.
- _____, "Aggregation, Income Distribution and Consumer Demand," *Rev. Econ.*

- Stud.*, Oct. 1975, 62, 525-43.
- _____, "Community Preferences and the Representative Consumer," *Econometrica*, Sept. 1976, 44, 979-99.
- R. W. Shephard**, *Cost and Production Functions*, Princeton 1953.
- _____, *Theory of Cost and Production Functions*, Princeton 1970.
- J. R. N. Stone**, *The Measurement of Consumers' Expenditure and Behaviour in the United Kingdom, 1920-1938*, Vol. 1, Cambridge 1953.
- _____, "Linear Expenditure Systems and Demand Analysis: An Application to the Pattern of British Demand," *Econ. J.*, Sept. 1954, 64, 511-27.
- Henri Theil**, "The Information Approach to Demand Analysis," *Econometrica*, Jan. 1965, 33, 67-87.
- _____, *Statistical Decomposition Analysis with Applications in the Social and Administrative Sciences*, Amsterdam 1972.
- _____, *Theory and Measurement of Consumer Demand*, Vols. 1 and 2, Amsterdam 1976.
- H. Working**, "Statistical Laws of Family Expenditure," *J. Amer. Statist. Assn.*, Mar. 1943, 38, 43-56.

On the Impossibility of Informationally Efficient Markets

By SANFORD J. GROSSMAN AND JOSEPH E. STIGLITZ*

If competitive equilibrium is defined as a situation in which prices are such that all arbitrage profits are eliminated, is it possible that a competitive economy always be in equilibrium? Clearly not, for then those who arbitrage make no (private) return from their (privately) costly activity. Hence the assumptions that all markets, including that for information, are always in equilibrium and always perfectly arbitrated are inconsistent when arbitrage is costly.

We propose here a model in which there is an equilibrium degree of disequilibrium: prices reflect the information of informed individuals (arbitrageurs) but only partially, so that those who expend resources to obtain information do receive compensation. How informative the price system is depends on the number of individuals who are informed; but the number of individuals who are informed is itself an endogenous variable in the model.

The model is the simplest one in which prices perform a well-articulated role in conveying information from the informed to the uninformed. When informed individuals observe information that the return to a security is going to be high, they bid its price up, and conversely when they observe information that the return is going to be low. Thus the price system makes publicly available the information obtained by informed individuals to the uninformed. In general, however, it does this imperfectly; this is perhaps lucky, for were it to do it perfectly, an equilibrium would not exist.

In the introduction, we shall discuss the general methodology and present some con-

jectures concerning certain properties of the equilibrium. The remaining analytic sections of the paper are devoted to analyzing in detail an important example of our general model, in which our conjectures concerning the nature of the equilibrium can be shown to be correct. We conclude with a discussion of the implications of our approach and results, with particular emphasis on the relationship of our results to the literature on "efficient capital markets."

I. The Model

Our model can be viewed as an extension of the noisy rational expectations model introduced by Robert Lucas and applied to the study of information flows between traders by Jerry Green (1973); Grossman (1975, 1976, 1978); and Richard Kihlstrom and Leonard Mirman. There are two assets: a safe asset yielding a return R , and a risky asset, the return to which, u , varies randomly from period to period. The variable u consists of two parts,

$$(1) \quad u = \theta + \varepsilon$$

where θ is observable at a cost c , and ε is unobservable.¹ Both θ and ε are random variables. There are two types of individuals, those who observe θ (informed traders), and those who observe only price (uninformed traders). In our simple model, all individuals are, *ex ante*, identical; whether they are informed or uninformed just depends on whether they have spent c to obtain information. Informed traders' demands will depend on θ and the price of the risky asset P . Uninformed traders' demands

*University of Pennsylvania and Princeton University, respectively. Research support under National Science Foundation grants SOC76-18771 and SOC77-15980 is gratefully acknowledged. This is a revised version of a paper presented at the Econometric Society meetings, Winter 1975, at Dallas, Texas.

¹An alternative interpretation is that θ is a "measurement" of u with error. The mathematics of this alternative interpretation differ slightly, but the results are identical.

will depend only on P , but we shall assume that they have rational expectations; they learn the relationship between the distribution of return and the price, and use this in deriving their demand for the risky assets. If x denotes the supply of the risky asset, an equilibrium when a given percentage, λ , of traders are informed, is thus a price function $P_\lambda(\theta, x)$ such that, when demands are formulated in the way described, demand equals supply. We assume that uninformed traders do not observe x . Uninformed traders are prevented from learning θ via observations of $P_\lambda(\theta, x)$ because they cannot distinguish variations in price due to changes in the informed trader's information from variations in price due to changes in aggregate supply. Clearly, $P_\lambda(\theta, x)$ reveals some of the informed trader's information to the uninformed traders.

We can calculate the expected utility of the informed and the expected utility of the uninformed. If the former is greater than the latter (taking account of the cost of information), some individuals switch from being uninformed to being informed (and conversely). An overall equilibrium requires the two to have the same expected utility. As more individuals become informed, the expected utility of the informed falls relative to the uninformed for two reasons:

(a) The price system becomes more informative because variations in θ have a greater effect on aggregate demand and thus on price when more traders observe θ . Thus, more of the information of the informed is available to the uninformed. Moreover, the informed gain more from trade with the uninformed than do the uninformed. The informed, on average, buy securities when they are "underpriced" and sell them when they are "overpriced" (relative to what they would have been if information were equalized).² As the price system becomes more informative, the difference in their information—and hence the magnitude by

which the informed can gain relative to the uninformed—is reduced.

(b) Even if the above effect did not occur, the increase in the ratio of informed to uninformed means that the relative gains of the informed, on a per capita basis, in trading with the uninformed will be smaller.

We summarize the above characterization of the equilibrium of the economy in the following two conjectures:

Conjecture 1: The more individuals who are informed, the more informative is the price system.

Conjecture 2: The more individuals who are informed, the lower the ratio of expected utility of the informed to the uninformed.

(Conjecture 1 obviously requires a definition of "more informative"; this is given in the next section and in fn. 7.)

The equilibrium number of informed and uninformed individuals in the economy will depend on a number of critical parameters: the cost of information, how informative the price system is (how much noise there is to interfere with the information conveyed by the price system), and how informative the information obtained by an informed individual is.

Conjecture 3: The higher the cost of information, the smaller will be the equilibrium percentage of individuals who are informed.

Conjecture 4: If the quality of the informed trader's information increases, the more their demands will vary with their information and thus the more prices will vary with θ . Hence, the price system becomes more informative. The equilibrium proportion of informed to uninformed may be either increased or decreased, because even though the value of being informed has increased due to the increased quality of θ , the value of being uninformed has also increased because the price system becomes more informative.

Conjecture 5: The greater the magnitude of noise, the less informative will the price system be, and hence the lower the expected utility of uninformed individuals. Hence, in equilibrium the greater the magnitude of noise, the larger the proportion of informed individuals.

²The framework described herein does not explicitly model the effect of variations in supply, i.e., x on commodity storage. The effect of futures markets and storage capabilities on the informativeness of the price system was studied by Grossman (1975, 1977).

Conjecture 6: In the limit, when there is no noise, prices convey all information, and there is no incentive to purchase information. Hence, the only possible equilibrium is one with no information. But if everyone is uninformed, it clearly pays some individual to become informed.³ Thus, there does not exist a competitive equilibrium.⁴

Trade among individuals occurs either because tastes (risk aversions) differ, endowments differ, or beliefs differ. This paper focuses on the last of these three. An interesting feature of the equilibrium is that beliefs may be precisely identical in either one of two situations: when all individuals are informed or when all individuals are uninformed. This gives rise to:

Conjecture 7: That, other things being equal, markets will be thinner under those conditions in which the percentage of individuals who are informed (λ) is either near zero or near unity. For example, markets will be thin when there is very little noise in the system (so λ is near zero), or when costs of information are very low (so λ is near unity).

In the last few paragraphs, we have provided a number of conjectures describing the nature of the equilibrium when prices convey information. Unfortunately, we have not been able to obtain a general proof of any of these propositions. What we have been able to do is to analyze in detail an interesting example, entailing constant absolute risk-aversion utility functions and normally distributed random variables. In this example, the equilibrium price distribution can actually be calculated, and all of

³That is, with no one informed, an individual can only get information by paying c dollars, since no information is revealed by the price system. By paying c dollars an individual will be able to predict better than the market when it is optimal to hold the risky asset as opposed to the risk-free asset. Thus his expected utility will be higher than an uninformed person gross of information costs. Thus for c sufficiently low all uninformed people will desire to be informed.

⁴See Grossman (1975, 1977) for a formal example of this phenomenon in futures markets. See Stiglitz (1971, 1974) for a general discussion of information and the possibility of nonexistence of equilibrium in capital markets.

the conjectures provided above can be verified. The next sections are devoted to solving for the equilibrium in this particular example.⁵

II. Constant Absolute Risk-Aversion Model

A. The Securities

The i th trader is assumed to be endowed with stocks of two types of securities: \bar{M}_i , the riskless asset, and \bar{X}_i , a risky asset. Let P be the current price of risky assets and set the price of risk free assets equal to unity. The i th trader's budget constraint is

$$(2) \quad PX_i + M_i = W_{0i} \equiv \bar{M}_i + P\bar{X}_i$$

Each unit of the risk free asset pays R "dollars" at the end of the period, while each unit of the risky asset pays u dollars. If at the end of the period, the i th trader holds a portfolio (M_i, X_i) , his wealth will be

$$(3) \quad W_{1i} = RM_i + uX_i$$

B. Individual's Utility Maximization

Each individual has a utility function $V_i(W_{1i})$. For simplicity, we assume all individuals have the same utility function and so drop the subscripts i . Moreover, we assume the utility function is exponential, i.e.,

$$V(W_{1i}) = -e^{-aW_{1i}}, \quad a > 0$$

where a is the coefficient of absolute risk aversion. Each trader desires to maximize expected utility, using whatever information is available to him, and to decide on what information to acquire on the basis of the consequences to his expected utility.

Assume that in equation (1) θ and ϵ have a multivariate normal distribution, with

$$(4) \quad E\epsilon = 0$$

$$(5) \quad E\theta\epsilon = 0$$

$$(6) \quad \text{Var}(u^*|\theta) = \text{Var}\epsilon^* \equiv \sigma_\epsilon^2 > 0$$

⁵The informational equilibria discussed here may not, in general, exist. See Green (1977). Of course, for the utility function we choose equilibrium does exist.

since θ and ϵ are uncorrelated. Throughout this paper we will put a * above a symbol to emphasize that it is a random variable. Since W_{1i} is a linear function of ϵ , for a given portfolio allocation, and a linear function of a normally distributed random variable is normally distributed, it follows that W_{1i} is normal conditional on θ . Then, using (2) and (3) the expected utility of the *informed* trader with information θ can be written

$$(7) \quad E(V(W_{1i}^*)|\theta) = -\exp\left(-a\left\{E[W_{1i}^*|\theta] - \frac{a}{2} \text{Var}[W_{1i}^*|\theta]\right\}\right)$$

$$= -\exp\left(-a\left[RW_{0i} + X_I\{E(u^*|\theta) - RP\} - \frac{a}{2} X_I^2 \text{Var}(u^*|\theta)\right]\right)$$

$$= -\exp\left(-a\left[RW_{0i} + X_I(\theta - RP) - \frac{a}{2} X_I^2 \sigma_\epsilon^2\right]\right)$$

where X_I is an informed individual's demand for the risky security. Maximizing (7) with respect to X_I yields a demand function for risky assets:

$$(8) \quad X_I(P, \theta) = \frac{\theta - RP}{a\sigma_\epsilon^2}$$

The right-hand side of (8) shows the familiar result that with constant absolute risk aversion, a trader's demand does not depend on wealth; hence the subscript i is not on the left-hand side of (8).

We now derive the demand function for the uninformed. Let us assume the only source of "noise" is the per capita supply of the risky security x .

Let $P^*(\cdot)$ be some particular price function of (θ, x) such that u^* and P^* are jointly normally distributed. (We will prove that this exists below.)

Then, we can write for the uninformed individual

$$(7') \quad E(V(W_{1i}^*)|P^*) = -\exp\left[-a\left\{E[W_{1i}^*|P^*]\right.\right.$$

$$\left.\left.- \frac{a}{2} \text{Var}[W_{1i}^*|P^*]\right\}\right]$$

$$= -\exp\left[-a\left\{RW_{0i} + X_U(E[u^*|P^*] - RP)\right.\right.$$

$$\left.\left.- \frac{a}{2} X_U^2 \text{Var}[u^*|P^*]\right\}\right]$$

The demands of the uninformed will thus be a function of the price function P^* and the actual price P .

$$(8') \quad X_U(P; P^*) = \frac{E[u^*|P^*(\theta, x) = P] - RP}{a \text{Var}[u^*|P^*(\theta, x) = P]}$$

C. Equilibrium Price Distribution

If λ is some particular fraction of traders who decide to become informed, then define an equilibrium price system as a function of (θ, x) , $P_\lambda(\theta, x)$, such that for all (θ, x) per capita demands for the risky assets equal supplies:

$$(9) \quad \lambda X_I(P_\lambda(\theta, x), \theta) + (1 - \lambda) X_U(P_\lambda(\theta, x); P_\lambda^*) = x$$

The function $P_\lambda(\theta, x)$ is a statistical equilibrium in the following sense. If over time uninformed traders observe many realizations of (u^*, P_λ^*) , then they learn the joint distribution of (u^*, P_λ^*) . After all learning about the joint distribution of (u^*, P_λ^*) ceases, all traders will make allocations and form expectations such that this joint distribution persists over time. This follows from (8), (8'), and (9), where the market-clearing price that comes about is the one which takes into account the fact that uninformed traders have learned that it contains information.

We shall now prove that there exists an equilibrium price distribution such that P^* and u^* are jointly normal. Moreover, we shall be able to characterize the price distribution. We define

$$(10a) \quad w_\lambda(\theta, x) = \theta - \frac{a\sigma_e^2}{\lambda}(x - Ex^*)$$

for $\lambda > 0$, and define $w_0(\theta, x)$ as the number:

$$(10b) \quad w_0(\theta, x) = x \quad \text{for all } (\theta, x)$$

where w_λ is just the random variable θ , plus noise.⁶ The magnitude of the noise is inversely proportional to the proportion of informed traders, but is proportional to the variance of ϵ . We shall prove that the equilibrium price is just a linear function of w_λ . Thus, if $\lambda > 0$, the price system conveys information about θ , but it does so imperfectly.

D. Existence of Equilibrium and a Characterization Theorem

THEOREM 1: *If $(\theta^*, \epsilon^*, x^*)$ has a nondegenerate joint normal distribution such that θ^* , ϵ^* , and x^* are mutually independent, then there exists a solution to (9) which has the form $P_\lambda(\theta, x) = \alpha_1 + \alpha_2 w_\lambda(\theta, x)$, where α_1 and α_2 are real numbers which may depend on λ , such that $\alpha_2 > 0$. (If $\lambda = 0$, the price contains no information about θ .) The exact form of $P_\lambda(\theta, x)$ is given in equation (A10) in Appendix B. The proof of this theorem is also in Appendix B.*

The importance of Theorem 1 rests in the simple characterization of the information in the equilibrium price system: P_λ^* is informationally equivalent to w_λ^* . From (10) w_λ^* is a "mean-preserving spread" of θ ; i.e., $E[w_\lambda^*|\theta] = \theta$ and

$$(11) \quad \text{Var}[w_\lambda^*|\theta] = \frac{a^2\sigma_e^4}{\lambda^2} \text{Var} x^*$$

⁶If $y' = y + Z$, and $E[Z|y] = 0$, then y' is just y plus noise.

For each replication of the economy, θ is the information that uninformed traders would like to know. But the noise x^* prevents w_λ^* from revealing θ . How well-informed uninformed traders can become from observing P_λ^* (equivalently w_λ^*) is measured by $\text{Var}[w_\lambda^*|\theta]$. When $\text{Var}[w_\lambda^*|\theta]$ is zero, w_λ^* and θ are perfectly correlated. Hence when uninformed firms observe w_λ^* , this is equivalent to observing θ . On the other hand, when $\text{Var}[w_\lambda^*|\theta]$ is very large, there are "many" realizations of w_λ^* that are associated with a given θ . In this case the observation of a particular w_λ^* tells very little about the actual θ which generated it.⁷

From equation (11) it is clear that large noise (high $\text{Var} x^*$) leads to an imprecise price system. The other factor which determines the precision of the price system ($a^2\sigma_e^4/\lambda^2$) is more subtle. When a is small (the individual is not very risk averse) or σ_e^2 is small (the information is very precise), an informed trader will have a demand for risky assets which is very responsive to changes in θ . Further, the larger λ is, the more responsive is the total demand of informed traders. Thus small ($a^2\sigma_e^4/\lambda^2$) means that the aggregate demand of informed traders is very responsive to θ . For a fixed amount of noise (i.e., fixed $\text{Var} x^*$) the larger are the movements in aggregate demand which are due to movements in θ , the more will price movements be due to movements in θ . That is, x^* becomes less important relative to θ in determining price movements. Therefore, for small ($a^2\sigma_e^4/\lambda^2$) uninformed traders are able to confidently know that price is, for example, unusually high due to θ being high. In this way information from informed traders is transferred to uninformed traders.

⁷Formally, w_λ^* is an experiment in the sense of Blackwell which gives information about θ . It is easy to show that, *ceteris paribus*, the smaller $\text{Var}(w_\lambda^*|\theta)$ the more "informative" (or sufficient) in the sense of Blackwell, is the experiment; see Grossman, Kihlstrom, and Mirman (p. 539).

E. Equilibrium in the Information Market

What we have characterized so far is the equilibrium price distribution for given λ . We now define an *overall* equilibrium to be a pair (λ, P_λ^*) such that the expected utility of the informed is equal to that of the uninformed if $0 < \lambda < 1$; $\lambda = 0$ if the expected utility of the informed is less than that of the uninformed at P_0^* ; $\lambda = 1$ if the expected utility of the informed is greater than the uninformed at P_1^* . Let

$$(12a) \quad W_{li}^\lambda \equiv R(W_{0i} - c) + [u - RP_\lambda(\theta, x)]X_i(P_\lambda(\theta, x), \theta)$$

$$(12b) \quad W_{U1}^\lambda \equiv RW_{0i} + [u - RP_\lambda(\theta, x)]X_U(P_\lambda(\theta, x); P_\lambda^*)$$

where c is the cost of observing a realization of θ^* . Equation (12a) gives the end of period wealth of a trader if he decides to become informed, while (12b) gives his wealth if he decides to be uninformed. Note that end of period wealth is random due to the randomness of W_{0i} , u , θ , and x .

In evaluating the expected utility of W_{li}^λ , we do not assume that a trader knows which realization of θ^* he gets to observe if he pays c dollars. A trader pays c dollars and then gets to observe some realization of θ^* . The overall expected utility of W_{li}^λ averages over all possible θ^* , ϵ^* , x^* , and W_{0i} . The variable W_{0i} is random for two reasons. First from (2) it depends on $P_\lambda(\theta, x)$, which is random as (θ, x) is random. Secondly, in what follows we will assume that X_i is random.

We will show below that $EV(W_{li}^\lambda)/EV(W_{Ui}^\lambda)$ is independent of i , but is a function of λ , a , c , and σ_e^2 . More precisely, in Appendix B we prove

THEOREM 2: *Under the assumptions of Theorem 1, and if \bar{X}_i is independent of (u^*, θ^*, x^*) then*

$$(13) \quad \frac{EV(W_{li}^\lambda)}{EV(W_{Ui}^\lambda)} = e^{ac} \sqrt{\frac{Var(u^*|\theta)}{Var(u^*|w_\lambda)}}$$

F. Existence of Overall Equilibrium

Theorem 2 is useful, both in proving the uniqueness of overall equilibrium and in analyzing comparative statics. Overall equilibrium, it will be recalled, requires that for $0 < \lambda < 1$, $EV(W_{li}^\lambda)/EV(W_{Ui}^\lambda) = 1$. But from (13)

$$(14) \quad \frac{EV(W_{li}^\lambda)}{EV(W_{Ui}^\lambda)} = e^{ac} \sqrt{\frac{Var(u^*|\theta)}{Var(u^*|w_\lambda)}} \equiv \gamma(\lambda)$$

Hence overall equilibrium simply requires, for $0 < \lambda < 1$,

$$(15) \quad \gamma(\lambda) = 1$$

More precisely, we now prove

THEOREM 3: *If $0 < \lambda < 1$, $\gamma(\lambda) = 1$, and P_λ^* is given by (A10) in Appendix B, then (λ, P_λ^*) is an overall equilibrium. If $\gamma(1) < 1$, then $(1, P_1^*)$ is an overall equilibrium. If $\gamma(0) > 1$, then $(0, P_0^*)$ is an overall equilibrium. For all price equilibria P_λ which are monotone functions of w_λ , there exists a unique overall equilibrium (λ, P_λ^*) .*

PROOF:

The first three sentences follow immediately from the definition of overall equilibrium given above equation (12), and Theorems 1 and 2. Uniqueness follows from the monotonicity of $\gamma(\cdot)$ which follows from (A11) and (14). The last two sentences in the statement of the theorem follow immediately.

In the process of proving Theorem 3, we have noted

COROLLARY 1: $\gamma(\lambda)$ is a strictly monotone increasing function of λ .

This looks paradoxical; we expect the ratio of informed to uninformed expected utility to be a decreasing function of λ . But, we have defined utility as negative. Therefore

as λ rises, the expected utility of informed traders does go down relative to uninformed traders.

Note that the function $\gamma(0) = e^{ac} (\text{Var}(u^*|\theta)) / \text{Var}(u^*)^{1/2}$. Figure 1 illustrates the determination of the equilibrium λ . The figure assumes that $\gamma(0) < 1 < \gamma(1)$.

G. Characterization of Equilibrium

We wish to provide some further characterization of the equilibrium. Let us define

$$(16a) \quad m = \left(\frac{a\sigma_e^2}{\lambda} \right)^2 \frac{\sigma_x^2}{\sigma_\theta^2}$$

$$(16b) \quad n = \frac{\sigma_\theta^2}{\sigma_e^2}$$

Note that m is inversely related to the informativeness of the price system since the squared correlation coefficient between P_λ^* and θ^* , ρ_θ^2 is given by

$$(17) \quad \rho_\theta^2 = \frac{1}{1+m}$$

Similarly, n is directly related to the quality of the informed trader's information because $n/(1+n)$ is the squared correlation coefficient between θ^* and u^* .

Equations (14) and (15) show that the cost of information c , determines the equilibrium ratio of information quality between informed and uninformed traders ($\text{Var}(u^*|\theta)) / \text{Var}(u^*|w_\lambda)$). From (1), (A11) of Appendix A, and (16), this can be written as

(18)

$$\frac{\text{Var}(u^*|\theta)}{\text{Var}(u^*|w_\lambda)} = \frac{1+m}{1+m+nm} = \left(1 + \frac{nm}{1+m} \right)^{-1}$$

Substituting (18) into (14) and using (15) we obtain, for $0 < \lambda < 1$, in equilibrium

$$(19a) \quad m = \frac{e^{2ac} - 1}{1 + n - e^{2ac}}$$

or

$$(19b) \quad 1 - \rho_\theta^2 = \frac{e^{2ac} - 1}{n}$$

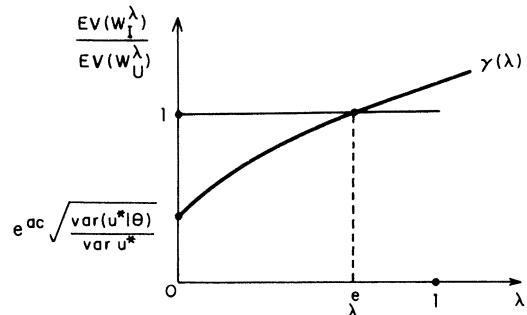


FIGURE 1

Note that (19) holds for $\gamma(0) < 1 < \gamma(1)$, since these conditions insure that the equilibrium λ is between zero and one. Equation (19b) shows that the equilibrium informativeness of the price system is determined completely by the cost of information c , the quality of the informed trader's information n , and the degree of risk aversion a .

H. Comparative Statics

From equation (19b), we immediately obtain some basic comparative statics results:

1) An increase in the quality of information (n) increases the informativeness of the price system.

2) A decrease in the cost of information increases the informativeness of the price system.

3) A decrease in risk aversion leads informed individuals to take larger positions, and this increases the informativeness of the price system.

Further, all other changes in parameters, such that n , a , and c remain constant, do not change the equilibrium degree of informativeness of the price system; other changes lead only to particular changes in λ of a magnitude to exactly offset them. For example:

4) An increase in noise (σ_x^2) increases the proportion of informed traders. At any given λ , an increase in noise reduces the informativeness of the price system; but it increases the returns to information and leads more individuals to become informed; the remarkable result obtained above establishes that the two effects exactly offset each

other so that the equilibrium informativeness of the price system is unchanged. This can be illustrated diagrammatically if we note from (16a) that for a given λ , an increase in σ_x^2 raises m which from (18) lowers $(Var(u^*|\theta))/Var(u^*|w_\lambda)$. Thus from (14) a rise in σ_x^2 leads to a vertical downward shift of the $\gamma(\lambda)$ curve in Figure 1, and thus a higher value of λ^e .

5) Similarly an increase in σ_e^2 for a constant n (equivalent to an increase in the variance of u since n is constant) leads to an increased proportion of individuals becoming informed—and indeed again just enough to offset the increased variance, so that the degree of *informativeness* of the price system remains unchanged. This can also be seen from Figure 1 if (16) is used to note that an increase in σ_e^2 with n held constant by raising σ_θ^2 leads to an increase in m for a given λ . From (18) and (14) this leads to a vertical downward shift of the $\gamma(\lambda)$ curve and thus a higher value of λ^e .

6) It is more difficult to determine what happens if, say σ_θ^2 increases, keeping σ_u^2 constant (implying a fall in σ_e^2), that is, *the information obtained is more informative*. This leads to an increase in n , which from (19b) implies that the equilibrium informativeness of the price system rises. From (16) it is clear that m and nm both fall when σ_θ^2 rises (keeping $\sigma_u^2 = \sigma_\theta^2 + \sigma_e^2$ constant). This implies that the $\gamma(\lambda)$ curve may shift up or down depending on the precise values of c , a , and n .⁸ This ambiguity arises because an

⁸From (14) and (18) it is clear that λ rises if and only if $Var(u^*|\theta) + Var(u^*|w_\lambda)$ falls due to the rise in σ_θ^2 for a given λ . This occurs if and only if $nm/(1+m)$ rises. Using (16) to differentiate $nm/(1+m)$ with respect to σ_e^2 subject to the constraint that $d\sigma_u^2=0$ (i.e., $d\sigma_\theta^2=-d\sigma_e^2$), we find that the sign of

$$\begin{aligned} \frac{d}{d\sigma_\theta^2} \left(\frac{nm}{1+m} \right) &= sgn \left[m \left(\frac{n+1}{n} \right) - 1 \right] \\ &= sgn \left[\left(\frac{\gamma}{n-\gamma} \right) \left(\frac{n+1}{n} \right) - 1 \right] \end{aligned}$$

where $\gamma \equiv e^{2ac} - 1$ and the last equality follows from equation (19a). Thus for n very large the derivative is negative so that λ falls due to an increase in the precision of the informed trader's information. Similarly if n is sufficiently small, the derivative is positive and thus λ rises.

improvement in the precision of informed traders' information, with the cost of the information fixed, increases the benefit of being informed. However, some of the improved information is transmitted, via a more informative price system, to the uninformed; this increases the benefits of being uninformed. If n is small, both the price system m is not very informative and the marginal value of information to informed traders is high. Thus the *relative* benefits of being informed rises when n rises; implying that the equilibrium λ rises. Conversely when n is large the price system is very informative and the marginal value of information is low to informed traders so the relative benefits of being uninformed rises.

7) From (14) it is clear that an increase in the cost of information c shifts the $\gamma(\lambda)$ curve up and thus decreases the percentage of informed traders.

The above results are summarized in the following theorem.

THEOREM 4: *For equilibrium λ such that $0 < \lambda < 1$:*

A. *The equilibrium informativeness of the price system, ρ_θ^2 , rises if n rises, c falls, or a falls.*

B. *The equilibrium informativeness of the price system is unchanged if σ_x^2 changes, or if σ_u^2 changes with n fixed.*

C. *The equilibrium percentage of informed traders will rise if σ_x^2 rises, σ_u^2 rises for a fixed n , or c falls.*

D. *If \bar{n} satisfies $(e^{2ac} - 1)/(\bar{n} - (e^{2ac} - 1)) = \bar{n}/(\bar{n} + 1)$, then $n > \bar{n}$ ($<$) implies that λ falls (rises) due to an increase in n .*

PROOF:

Parts A–C are proved in the above remarks. Part D is proved in footnote 8.

I. Price Cannot Fully Reflect Costly Information

We now consider certain limiting cases, for $\gamma(0) < 1 < \gamma(1)$, and show that equilibrium does not exist if $c > 0$ and price is fully informative.

1) As the cost of information goes to zero, the price system becomes more infor-

mative, but at a positive value of c , say \hat{c} , all traders are informed. From (14) and (15) \hat{c} satisfies

$$e^{\hat{c}} \sqrt{\frac{Var(u^*|\theta)}{Var(u^*|w_1)}} = 1$$

2) From (19a) as the precision of the informed trader's information n goes to infinity, i.e., $\sigma_\epsilon^2 \rightarrow 0$ and $\sigma_\theta^2 \rightarrow \sigma_u^2$, σ_u^2 held fixed, the price system becomes perfectly informative. Moreover the percentage of informed traders goes to zero! This can be seen from (18) and (15). That is, as $\sigma_\epsilon^2 \rightarrow 0$, $nm/(1+m)$ must stay constant for equilibrium to be maintained. But from (19b) and (17), m falls as σ_ϵ^2 goes to zero. Therefore nm must fall, but nm must not go to zero or else $nm/(1+m)$ would not be constant. From (16) $nm = (a/\lambda)^2 \sigma_\epsilon^2 \sigma_x^2$, and thus λ must go to zero to prevent nm from going to zero as $\sigma_\epsilon^2 \rightarrow 0$.

3) From (16a) and (19a) it is clear that as noise σ_x^2 goes to zero, the percentage of informed traders goes to zero. Further, since (19a) implies that m does not change as σ_x^2 changes, the informativeness of the price system is unchanged as $\sigma_x^2 \rightarrow 0$.

Assume that c is small enough so that it is worthwhile for a trader to become informed when no other trader is informed. Then if $\sigma_x^2 = 0$ or $\sigma_\epsilon^2 = 0$, there exists no competitive equilibrium. To see this, note that equilibrium requires either that the ratio of expected utility of the informed to the uninformed be equal to unity, or that if the ratio is larger than unity, no one be informed. We shall show that when no one is informed, it is less than unity so that $\lambda = 0$ cannot be an equilibrium; but when $\lambda > 0$, it is greater than unity. That is, if $\sigma_x^2 = 0$ or $\sigma_\epsilon^2 = 0$, the ratio of expected utilities is not a continuous function of λ at $\lambda = 0$.

This follows immediately from observing that at $\lambda = 0$, $Var(u^*|w_0) = Var u^*$, and thus by (14)

$$(20) \quad \frac{EV(W_{Ii}^0)}{EV(W_{Ui}^0)} = e^{ac} \sqrt{\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_\theta^2}} \\ = e^{ac} \sqrt{\frac{1}{1+n}}$$

while if $\lambda > 0$, by (18)

$$\frac{EV(W_{Ii}^\lambda)}{EV(W_{Ui}^\lambda)} = e^{ac} \sqrt{\frac{1}{1+n \frac{m}{m+1}}}$$

But if $\sigma_x^2 = 0$ or $\sigma_\epsilon^2 = 0$, then $m = 0$, $nm = 0$ for $\lambda > 0$, and hence

$$(21) \quad \lim_{\lambda \rightarrow 0} \frac{EV(W_{Ii}^\lambda)}{EV(W_{Ui}^\lambda)} = e^{ac}$$

It immediately follows that

THEOREM 5: (a) *If there is no noise ($\sigma_x^2 = 0$), an overall equilibrium does not exist if (and only if) $e^{ac} < \sqrt{1+n}$.* (b) *If information is perfect ($\sigma_\epsilon^2 = 0, n = \infty$), there never exists an equilibrium.*

PROOF:

(a) If $e^{ac} < \sqrt{1+n}$, then by (20) and (21), $\gamma(\lambda)$ is discontinuous at $\lambda = 0$; $\lambda = 0$ is not an equilibrium since by (20) $\gamma(0) < 1$; $\lambda > 0$ is not an equilibrium since by (21) $\gamma(\lambda) > 1$.

(b) If $\sigma_\epsilon^2 = 0$ and $\sigma_\theta^2 = \sigma_u^2$ so that information is perfect, then for $\lambda > 0$, $nm = 0$ by (16) and hence $\gamma(\lambda) > 1$ by (21). From (20) $\gamma(0) = 0 < 1$.

If there is no noise and some traders become informed, then *all* their information is transmitted to the uninformed by the price system. Hence each informed trader acting as a price taker thinks the informativeness of the price system will be unchanged if he becomes uninformed, so $\lambda > 0$ is not an equilibrium. On the other hand, if no traders are informed, then each uninformed trader learns nothing from the price system, and thus he has a desire to become informed (if $e^{ac} < (1+n)^{1/2}$). Similarly if the informed traders get perfect information, then their demands are very sensitive to their information, so that the market-clearing price becomes very sensitive to their information and thus reveals θ to the uninformed. Hence all traders desire to be uninformed. But if all traders are uninformed, each trader can eliminate the risk of his portfolio by the purchase of information, so each trader desires to be informed.

In the next section we show that the non-existence of competitive equilibrium can be thought of as the breakdown of competitive markets due to lack of trade. That is, we will show that as σ_x^2 gets very small, trade goes to zero and markets serve no function. Thus competitive markets close for lack of trade "before" equilibrium ceases to exist at $\sigma_x^2=0$.

III. On the Thinness of Speculative Markets

In general, trade takes place because traders differ in endowments, preferences, or beliefs. Grossman (1975, 1977, 1978) has argued that differences in preferences are not a major factor in explaining the magnitude of trade in speculative markets. For this reason the model in Section II gave all traders the same risk preferences (note that none of the results in Section II are affected by letting traders have different coefficients of absolute risk aversion). In this section we assume that trade requires differences in endowments or beliefs and dispense with differences in risk preference as an explanatory variable.⁹

There is clearly some fixed cost in operating a competitive market. If traders have to bear this cost, then trade in the market must be beneficial. Suppose traders have the same endowments and beliefs. Competitive equilibrium will leave them with allocations which are identical with their initial endowments. Hence, if it is costly to enter such a competitive market, no trader would ever enter. We will show below that in an important class of situations, there is continuity in the amount of net trade. That is, when initial endowments are the same and peo-

ples' beliefs differ *slightly*, then the competitive equilibrium allocation that an individual gets will be only *slightly* different from his initial endowment. Hence, there will only be a slight benefit to entering the competitive market. This could, for sufficiently high operating costs, be outweighed by the cost of entering the market.

The amount of trade occurring at any date is a random variable; a function of θ and x . It is easy to show that it is a normally distributed random variable. Since one of the primary determinants of the size of markets is differences in beliefs, one might have conjectured that markets will be thin, in some sense, if almost all traders are either informed or uninformed. This is not, however, obvious, since the amount of trade by any single trader may be a function of λ as well, and a few active traders can do the job of many small traders. In our model, there is a sense, however, in which our conjecture is correct.

We first calculate the magnitude of trades as a function of the exogenous parameters, θ and x . Let $h \equiv \sigma_e^2$, $\bar{x} = Ex^*$, and $\bar{\theta} \equiv E\theta^*$. (The actual trades will depend on the distribution of random endowments across all of the traders, but these we shall net out.) Per capita net trade is¹⁰

$$(22) \quad X_I - x = (1-\lambda) \left[\left(nm + \frac{ah}{\lambda} \right) (x - \bar{x}) \right. \\ \left. + [(m+1)n-1](\theta - \bar{\theta}) + \bar{x}nm \right] \\ + [1 + m + \lambda nm]$$

¹⁰Calculation of distribution of net trades

$$\frac{\lambda}{ah}(\theta - RP_\lambda) \\ + \frac{(1-\lambda) \left[(\bar{\theta} - RP_\lambda)(1+m)n + \theta - \bar{\theta} - \frac{ah}{\lambda}(x - \bar{x}) \right]}{ah(1+m+nm)n} = x \\ \text{or } \frac{(\theta - RP_\lambda)}{ah} \left(\lambda + \frac{(1-\lambda)(1+m)}{1+m+nm} \right) \\ = \left(\frac{\theta - RP_\lambda}{ah} \right) \left(\frac{1+m+\lambda nm}{1+m+nm} \right) \\ = x + \frac{(1-\lambda) \left[[(m+1)n-1](\theta - \bar{\theta}) + \frac{ah}{\lambda}(x - \bar{x}) \right]}{ah(1+m+\lambda nm)n}$$

⁹In the model described in Section II it was assumed that an individual's endowment \bar{X}_i is independent of the market's per capita endowment x^* . This was done primarily so there would not be useful information in an individual's endowment about the total market endowment. Such information would be useful in equilibrium because an individual observes $P_\lambda(\theta, x)$. If due to observing \bar{X}_i , he knows something about x , then by observing $P_\lambda(\theta, x)$, \bar{X}_i is valuable in making inferences about θ . To take this into account is possible, but would add undue complication to a model already overburdened with computations.

Thus, the mean of total informed trade is

$$(23) \quad E\lambda(X_I - x) = \frac{(1-\lambda)\lambda m \bar{x}}{1+m+\lambda nm}$$

and its variance is

$$(24) \quad \sigma_\theta^2(1-\lambda)^2\lambda^2 \left[[(m+1)n-1]^2 + \left(nm + \frac{a\sigma_\epsilon^2}{\lambda} \right)^2 \frac{\sigma_x^2}{\sigma_\theta^2} \right] \div (1+m+\lambda nm)^2 n^2$$

In the last section we considered limiting values of the exogenous variables with the property that $\lambda \rightarrow 0$. The following theorem will show that the mean and variance of trade go to zero as $\lambda \rightarrow 0$. That is, the distribution of $\lambda(X_I - x)$ becomes degenerate at zero as $\lambda \rightarrow 0$. This is not trivial because as $\lambda \rightarrow 0$ due to $n \rightarrow \infty$ (very precise information), the informed trader's demand $X_I(P, \theta)$ goes to infinity at most prices because the risky asset becomes riskless with perfect information.

THEOREM 6: (a) *For sufficiently large or small c , the mean and variance of trade is zero.* (b) *As the precision of informed traders' information n goes to infinity, the mean and variance of trade go to zero.*

PROOF:

(a) From remark 1) in Section II, Part I, $\lambda = 1$ if $c \leq \hat{c}$, which from (23) and (24) implies trade is degenerate at zero. From (14), for c sufficiently large, say c^0 , $\gamma(0) = 1$, so

$$\text{or } X_I = \frac{1+m+nm}{1+m+\lambda nm}$$

$$\times \left[x + \frac{(1-\lambda)((m+1)-1)(\theta - \bar{\theta}) + \frac{ah}{\lambda}(x - \bar{x})}{ah(1+m+nm)n} \right]$$

$$X_I - x =$$

$$\frac{(1-\lambda) \left[\left(nm + \frac{ah}{\lambda} \right) (x - \bar{x}) + ((m+1)-1)(\theta - \bar{\theta}) + \bar{x}nm \right]}{(1+m+\lambda nm)n}$$

the equilibrium $\lambda = 0$. As c goes to c^0 from below $\lambda \rightarrow 0$, and from (14), (15), and (18) $\lim_{c \uparrow c^0} (1+nm/(1+m))^{-1/2} = e^{-ac^0}$. Hence $\lim_{c \uparrow c^0} (nm/1+m)$ is a finite positive number. Thus from (22) mean trade goes to zero as $c \uparrow c^0$. If the numerator and the denominator of (24) are divided by $(1+m)^2$, then again using the fact that $m/1+m$ has a finite limit gives the result that as $c \uparrow c^0$, $\lambda \rightarrow 0$, and variance of trade goes to zero.

(b) By (14), (15), and (18), $nm/(1+m)$ is constant as $n \rightarrow \infty$. Further, from remark 2) of Section II, Part I, $\lambda \rightarrow 0$ as $n \rightarrow \infty$. Hence from (23) and (24), the mean and variance of trade go to zero.

(c) From remark 3) in Section II, Part I, m is constant and λ goes to zero as $\sigma_x^2 \rightarrow 0$. Therefore mean trade goes to zero. In (24), note that $(nm + a\sigma_\epsilon^2/\lambda)^2 \sigma_x^2 / \sigma_\theta^2 = (nmo_x / \sigma_\theta + (m)^{1/2})^2$ by (16a). Hence the variance of trade goes to zero as $\sigma_x^2 \rightarrow 0$.

Note further that $\lambda(X_I - x) + (1-\lambda)(X_U - x) = 0$ implies that no trade will take place as $\lambda \rightarrow 1$. Thus, the result that competitive equilibrium is incompatible with informationally efficient markets should be interpreted as meaning that speculative markets where prices reveal a lot of information will be very thin because it will be composed of individuals with very similar beliefs.

IV. On the Possibility of Perfect Markets

In Section II we showed that the price system reveals the signal w_λ^* to traders, where

$$w_\lambda \equiv \theta - \frac{a\sigma_\epsilon^2}{\lambda} (x - Ex^*)$$

Thus, for given information of informed traders θ , the price system reveals a noisy version of θ . The noise is $(a\sigma_\epsilon^2/\lambda)(x - Ex^*)$. Uninformed traders learn θ to within a random variable with mean zero and variance $(a\sigma_\epsilon^2/\lambda)^2 Var x^*$, where σ_ϵ^2 is the precision of informed traders' information, $Var x^*$ is the amount of endowment uncertainty, λ the fraction of informed traders, and a is the degree of absolute risk aversion. Thus, in general the price system does not reveal all

the information about "the true value" of the risky asset. (θ is the true value of the risky asset in that it reflects the best available information about the asset's worth.)

The only way informed traders can earn a return on their activity of information gathering, is if they can use their information to take positions in the market which are "better" than the positions of uninformed traders. "Efficient Markets" theorists have claimed that "at any time prices fully reflect all available information" (see Eugene Fama, p. 383). If this were so then informed traders could not earn a return on their information.

We showed that when the efficient markets hypothesis is true and information is costly, competitive markets break down. This is because when $\sigma_e^2 = 0$ or $Var x^* = 0$, w_λ , and thus price, does reflect all the information. When this happens, each informed trader, because he is in a competitive market, feels that he could stop paying for information and do as well as a trader who pays nothing for information. But all informed traders feel this way. Hence having any positive fraction informed is not an equilibrium. Having no one informed is also not an equilibrium, because then each trader, taking the price as given, feels that there are profits to be made from becoming informed.

Efficient Markets theorists seem to be aware that costless information is a *sufficient* condition for prices to fully reflect all available information (see Fama, p. 387); they are not aware that it is a *necessary* condition. But this is a *reducto ad absurdum*, since price systems and competitive markets are important only when information is costly (see Fredrick Hayek, p. 452).

We are attempting to redefine the Efficient Markets notion, not destroy it. We have shown that when information is very inexpensive, or when informed traders get very precise information, then equilibrium exists and the market price will reveal most of the informed traders' information. However, it was argued in Section III that such markets are likely to be thin because traders have almost homogeneous beliefs.

There is a further conflict. As Grossman (1975, 1977) showed, whenever there are differences in beliefs that are not completely arbitrated, there is an incentive to create a market. (Grossman, 1977, analyzed a model of a storable commodity whose spot price did not reveal all information because of the presence of noise. Thus traders were left with differences in beliefs about the future price of the commodity. This led to the opening of a futures market. But then uninformed traders had two prices revealing information to them, implying the elimination of noise.) But, because differences in beliefs are themselves endogenous, arising out of expenditure on information and the informativeness of the price system, the creation of markets eliminates the differences of beliefs which gave rise to them, and thus causes those markets to disappear. If the creation of markets were costless, as is conventionally assumed in equilibrium analyses, equilibrium would never exist. For instance, in our model, were we to introduce an additional security, say a security which paid

$$z = \begin{cases} 1 & \text{if } u > E\theta^* \\ 0 & \text{if } u \leq E\theta^* \end{cases}$$

then the demand y for this security by the informed would depend on its price, say q on p and on θ , while the uninformed demand depends only on p and q :

$$\lambda y_I(q, p, \theta) + (1 - \lambda)y_u(q, p) = 0$$

is the condition that demand equals (supply is zero for a pure security). Under weak assumptions, q and p would convey all the information concerning θ . Thus, the market would be "noiseless" and no equilibrium could exist.

Thus, we could argue as soon as the assumptions of the conventional perfect capital markets model are modified to allow even a slight amount of information imperfection and a slight cost of information, the traditional theory becomes untenable. There cannot be as many securities as states of nature. For if there were, competitive equilibrium would not exist.

It is only because of costly transactions and the fact that this leads to there being a limited number of markets, that competitive equilibrium can be established.

We have argued that because information is costly, prices cannot perfectly reflect the information which is available, since if it did, those who spent resources to obtain it would receive no compensation. There is a fundamental conflict between the efficiency with which markets spread information and the incentives to acquire information. However, we have said nothing regarding the social benefits of information, nor whether it is socially optimal to have "informationally efficient markets." We hope to examine the welfare properties of the equilibrium allocations herein in future work.

APPENDIX A

Here we collect some facts on conditional expectations used in the text. If X^* and Y^* are jointly normally distributed then

$$(A1) \quad E[X^*|Y^* = Y] = EX^* + \frac{Cov(X^*, Y^*)}{Var(Y^*)} \{ Y - EY^* \}$$

$$(A2) \quad Var[X^*|Y^* = Y] = Var(X^*) - \frac{[Cov(X^*, Y^*)]^2}{Var(Y^*)}$$

(See Paul Hoel, p. 200.) From (A1) note that $E[X^*|Y^*]$ is a function of Y . If the expectation of both sides of (A1) is taken, we see that

$$(A3) \quad E\{E[X^*|Y^* = Y]\} = EX^*$$

Note that $Var[X^*|Y^* = Y]$ is not a function of Y , as $Var(X^*)$, $Cov(X^*, Y^*)$, and $Var(Y^*)$ are just parameters of the joint distribution of X^* and Y^* .

Two other relevant properties of conditional expectation are

$$(A4) \quad E\{E[Y^*|F(X^*)]|X^*\} = E[Y^*|F(X^*)]$$

$$(A5) \quad E\{E[Y^*|X]|F(X^*)\} = E[Y^*|F(X^*)]$$

where $F(\cdot)$ is a given function on the range of X^* (see Robert Ash, p. 260).

APPENDIX B

PROOF of Theorem 1:

(a) Suppose $\lambda = 0$; then (9) becomes

$$(A6) \quad X_U(P_0(\theta, x), P_0^*) = x$$

Define

$$(A7) \quad P_0(\theta, x) \equiv \frac{E\theta^* - ax\sigma_u^2}{R}$$

where σ_u^2 is the variance of u . Note that $P_0(\theta^*, x^*)$ is uncorrelated with u^* , as x^* is uncorrelated with u^* . Hence

$$(A8) \quad E[u^*|P_0^* = P_0(\theta, x)] = Eu^* = E\theta^*$$

$$\text{and } Var[u^*|P_0^* = P_0(\theta, x)] = Var[u^*]$$

Substitution of (A8) in (8) yields

$$(A9) \quad X_U(P_0^*, P_0(\theta, x)) = \frac{E\theta^* - RP_0(\theta, x)}{a Var u}$$

Substitution of (A7) in the right-hand side of (A9) yields $X_U(P_0^*(\theta, x), P_0^*) = x$ which was to be shown.

(b) Suppose $0 < \lambda \leq 1$. Let

$$(A10) \quad P_\lambda(\theta, x) = \frac{\frac{\lambda w_\lambda}{a\sigma_\epsilon^2} + \frac{(1-\lambda)E[u^*|w_\lambda]}{a Var[u^*|w_\lambda]} - Ex^*}{R \left[\frac{\lambda}{a\sigma_\epsilon^2} + \frac{(1-\lambda)}{a Var[u^*|w_\lambda]} \right]}$$

Note that from equations (1), (10), (A1) and (A2):

$$(A11a) \quad E(u^*|w_\lambda) = E\theta^* + \frac{\sigma_\theta^2}{Var w_\lambda} \cdot (w_\lambda - E\theta^*)$$

$$(A11b) \quad Var(u^*|w_\lambda) = \sigma_\theta^2 + \sigma_\epsilon^2 - \frac{\sigma_\theta^2}{Var w_\lambda}$$

$$(A11c) \quad Var w_\lambda = \sigma_\theta^2 + \left(\frac{a\sigma_\epsilon^2}{\lambda} \right)^2 Var x^*$$

Since $P_\lambda(\theta, x)$ is a linear function of w_λ , it is immediate that $E(u^*|w_\lambda) \equiv E(u^*|P_\lambda)$, $Var(u^*|w_\lambda) = Var(u^*|P_\lambda)$, etc. To see that P_λ^* is an equilibrium, we must show that the following equation holds as an identity in (θ, x) , for $P_\lambda(\cdot)$ defined by (A10):

$$(A12) \quad \lambda \cdot \frac{\theta - RP_\lambda}{a\sigma_e^2} + (1-\lambda) \frac{E[u^*|w_\lambda] - RP_\lambda}{a Var[u^*|w_\lambda]} = x$$

It is immediate from (10) that (A12) holds as an identity in θ and x .

PROOF of Theorem 2:

(a) *Calculation of the expected utility of the informed.* Using the fact that W_{li}^λ is normally distributed conditional on (\bar{X}_i, θ, x)

$$(A13) \quad E[V(W_{li}^\lambda)|\bar{X}_i, \theta, x] = \exp \left[-a \left\{ E[W_{li}^\lambda|\bar{X}_i, \theta, x] - \frac{a}{2} Var[W_{li}^\lambda|\bar{X}_i, \theta, x] \right\} \right]$$

Using (8), (12), and the fact that (θ, x) determines a particular P ,

$$(A14a) \quad E[W_{li}^\lambda|\bar{X}_i, \theta, x] = R(W_{0i} - c) + \frac{(E[u^*|\theta] - RP_\lambda)^2}{a\sigma_e^2}$$

$$(A14b) \quad Var[W_{li}^\lambda|\bar{X}_i, \theta, x] = \frac{(E[u^*|\theta] - RP_\lambda)^2}{a^2\sigma_e^2}$$

Substitution of (A14) into (A13) yields

$$(A15) \quad E[V(W_{li}^\lambda)|\bar{X}_i, \theta, x] = -\exp \left[-aR(W_{0i} - c) - \frac{1}{2\sigma_e^2} (E[u^*|\theta] - RP_\lambda)^2 \right]$$

Note that, as $P_\lambda^*(\cdot) = P_\lambda(\theta, x)$,

$$(A16) \quad E \left(E \left[V(W_{li}^\lambda) | \bar{X}_i, \theta, x \right] | P_\lambda, \bar{X}_i \right) = E \left[V(W_{li}^\lambda) | P_\lambda, \bar{X}_i \right]$$

(see (A5)). Note that since W_{0i} is nonstochastic conditional on (P_λ, \bar{X}_i) , equation (A15) implies

$$(A17) \quad E \left[V(W_{li}^\lambda) | P_\lambda, \bar{X}_i \right] = -\exp[-aR(W_{0i}^\lambda - c)].$$

Note that by Theorem 1, conditioning on w_λ^* is equivalent to conditioning on P_λ^* . Define

$$(A18) \quad h_\lambda \equiv Var(E[u^*|\theta]|w_\lambda) = Var(\theta|w_\lambda), h_0 \equiv \sigma_e^2 \equiv h$$

$$(A19) \quad Z \equiv \frac{E[u^*|\theta] - RP_\lambda}{\sqrt{h_\lambda}}$$

Using (3) and (A18), equation (A17) can be written as

$$(A20) \quad E[V(W_{li}^\lambda)|P_\lambda, \bar{X}_i] = e^{ac} V(RW_{0i}) E \left[\exp \left[-\frac{h_\lambda}{2\sigma_e^2} Z^2 \right] | w_\lambda \right]$$

since \bar{X}_i and w_λ are independent. Conditional on w_λ , P_λ is nonstochastic and $E[u^*|\theta]$ is normal. Hence conditional on w_λ , $(Z^*)^2$ has a noncentral chi-square distribution (see C. Rao, p. 181). Then for $t > 0$ the moment generating function for $(Z^*)^2$ can be written

$$(A21) \quad E[e^{-tZ^2}|w_\lambda] = \frac{1}{\sqrt{1+2t}} \exp \left[\frac{-(E[Z|w_\lambda])^2 t}{1+2t} \right]$$

Note that $E[u^*|\theta] = E[u^*|\theta, x]$. Hence

$$(A22) \quad E[E[u^*|\theta]|w_\lambda] = E[u^*|w_\lambda] \\ = E\theta^* + \frac{\sigma_\theta^2}{Var w_\lambda} (w_\lambda - E\theta^*)$$

since w_λ is just a function of (θ, x) . Therefore

$$(A23) \quad E[Z^*|w_\lambda] = \frac{E[u^*|w_\lambda] - RP_\lambda}{\sqrt{h_\lambda}}$$

Since $u = \theta + \epsilon$

$$(A24) \quad Var(u^*|w_\lambda) = \sigma_\epsilon^2 + Var(\theta^*|w_\lambda) = \sigma_\epsilon^2 + h_\lambda$$

The nondegeneracy assumptions on (x^*, ϵ^*, u^*) imply $h_\lambda > 0$. Set $t = (h_\lambda/2\sigma_\epsilon^2)$; and evaluate (A21) using (A23) and (A24):

$$(A25) \quad E\left[\exp\left(-\frac{h_\lambda}{2\sigma_\epsilon^2} Z^2\right)|w_\lambda\right] = \sqrt{\frac{Var(u^*|\theta)}{Var(u^*|w_\lambda)}} \\ \cdot \exp\left(\frac{-(E(u^*|w_\lambda) - RP_\lambda)^2}{2Var(u^*|w_\lambda)}\right)$$

This permits the evaluation of (A20).

(b) *Calculation of expected utility of the uninformed.* Equations (8), (5), and the normality of W_{Ui}^λ conditional on w_λ can be used to show, by calculations parallel to (A13)–(A25), that

$$(A26) \quad E[V(W_{Ui}^\lambda)|w_\lambda, \bar{X}_i] \\ = V(RW_{0i}) \exp\left(\frac{-(E(u^*|w_\lambda) - RP_\lambda)^2}{2Var(u^*|w_\lambda)}\right)$$

Hence

$$(A27) \quad E[V(W_{Ui}^\lambda)|w_\lambda, \bar{X}_i] - E[V(W_{Ui}^\lambda)|w_\lambda, \bar{X}_i] \\ = \left[e^{ac} \sqrt{\frac{Var(u^*|\theta)}{Var(u^*|w_\lambda)}} - 1 \right] \\ \times E[V(W_{Ui}^\lambda)|w_\lambda, \bar{X}_i]$$

Taking expectations of both sides of (A27) yields:

$$(A28) \quad E[V(W_{Ui}^\lambda)] - E[V(W_{Ui}^\lambda)] \\ = \left[e^{ac} \sqrt{\frac{Var(u^*|\theta)}{Var(u^*|w_\lambda)}} - 1 \right] EV(W_{Ui}^\lambda)$$

Equation (13) follows immediately from (A28).

REFERENCES

- Robert B. Ash**, *Real Analysis and Probability*, New York 1972.
- E. Fama**, "Efficient Capital Markets: A Review of Theory and Empirical Work," *J. Finance*, May 1970, 25, 383–417.
- J. R. Green**, "Information, Efficiency and Equilibrium," disc. paper no. 284, Harvard Inst. Econ. Res., Mar. 1973.
- _____, "The Non-Existence of Informational Equilibria," *Rev. Econ. Stud.*, Oct. 1977, 44, 451–64.
- S. Grossman**, "Essays on Rational Expectations," unpublished doctoral dissertation, Univ. Chicago 1975.
- _____, "On the Efficiency of Competitive Stock Markets Where Traders Have Diverse Information," *J. Finance*, May 1976, 31, 573–85.
- _____, "The Existence of Futures Markets, Noisy Rational Expectations and Informational Externalities," *Rev. Econ. Stud.*, Oct. 1977, 64, 431–49.
- _____, "Further Results on the Informational Efficiency of Competitive Stock Markets," *J. Econ. Theory*, June 1978, 18, 81–101.
- _____, **R. Kihlstrom**, and **L. Mirman**, "A Bayesian Approach to the Production of Information and Learning by Doing," *Rev. Econ. Stud.*, Oct. 1977, 64, 533–47.
- F. H. Hayek**, "The Use of Knowledge in Society," *Amer. Econ. Rev.*, Sept. 1945, 35, 519–30.
- Paul G. Hoel**, *Introduction to Mathematical Statistics*, New York 1962.
- R. Kihlstrom** and **L. Mirman**, "Information and Market Equilibrium," *Bell. J. Econ.*, Spring 1975, 6, 357–76.

- R. E. Lucas, Jr., "Expectations and the Neutrality of Money," *J. Econ. Theory*, Apr. 1972, 4, 103-24.
- C. Rao, *Linear Statistical Inference and Its Applications*, New York 1965.

- J. E. Stiglitz, "Perfect and Imperfect Capital Markets," paper presented to the Econometric Society, New Orleans 1971.
- _____, "Information and Capital Markets," mimeo., Oxford Univ. 1974.

Scale Economies, Product Differentiation, and the Pattern of Trade

By PAUL KRUGMAN*

For some time now there has been considerable skepticism about the ability of comparative cost theory to explain the actual pattern of international trade. Neither the extensive trade among the industrial countries, nor the prevalence in this trade of two-way exchanges of differentiated products, make much sense in terms of standard theory. As a result, many people have concluded that a new framework for analyzing trade is needed.¹ The main elements of such a framework—economies of scale, the possibility of product differentiation, and imperfect competition—have been discussed by such authors as Bela Balassa, Herbert Grubel (1967, 1970), and Irving Kravis, and have been “in the air” for many years. In this paper I present a simple formal analysis which incorporates these elements, and show how it can be used to shed some light on some issues which cannot be handled in more conventional models. These include, in particular, the causes of trade between economies with similar factor endowments, and the role of a large domestic market in encouraging exports.

The basic model of this paper is one in which there are economies of scale in production and firms can costlessly differentiate their products. In this model, which is derived from recent work by Avinash Dixit and Joseph Stiglitz, equilibrium takes the form of Chamberlinian monopolistic competition: each firm has some monopoly power, but entry drives monopoly profits to zero. When two imperfectly competitive economies of this kind are allowed to trade, increasing returns produce trade and gains

from trade even if the economies have identical tastes, technology, and factor endowments. This basic model of trade is presented in Section I. It is closely related to a model I have developed elsewhere; in this paper a somewhat more restrictive formulation of demand is used to make the analysis in later sections easier.

The rest of the paper is concerned with two extensions of the basic model. In Section II, I examine the effect of transportation costs, and show that countries with larger domestic markets will, other things equal, have higher wage rates. Section III then deals with “home market” effects on trade patterns. It provides a formal justification for the commonly made argument that countries will tend to export those goods for which they have relatively large domestic markets.

This paper makes no pretense of generality. The models presented rely on extremely restrictive assumptions about cost and utility. Nonetheless, it is to be hoped that the paper provides some useful insights into those aspects of international trade which simply cannot be treated in our usual models.

I. The Basic Model

A. Assumptions of the Model

There are assumed to be a large number of potential goods, all of which enter symmetrically into demand. Specifically, we assume that all individuals in the economy have the same utility function,

$$(1) \quad U = \sum_i c_i^\theta \quad 0 < \theta < 1$$

where c_i is consumption of the i th good. The number of goods actually produced, n ,

*Yale University and Massachusetts Institute of Technology.

¹A paper which points out the difficulties in explaining the actual pattern of world trade in a comparative cost framework is the study of Gary Hufbauer and John Chilas.

will be assumed to be large, although smaller than the potential range of products.²

There will be assumed to be only one factor of production, labor. All goods will be produced with the same cost function:

$$(2) \quad l_i = \alpha + \beta x_i \quad \alpha, \beta > 0$$

$$i = 1, \dots, n$$

where l_i is labor used in producing the i th good and x_i is output of that good. In other words, I assume a fixed cost and constant marginal cost. Average cost declines at all levels of output, although at a diminishing rate.

Output of each good must equal the sum of individual consumptions. If we can identify individuals with workers, output must equal consumption of a representative individual times the labor force:

$$(3) \quad x_i = L c_i \quad i = 1, \dots, n$$

We also assume full employment, so that the total labor force must just be exhausted by labor used in production:

$$(4) \quad L = \sum_{i=1}^n (\alpha + \beta x_i)$$

Finally, we assume that firms maximize profits, but that there is free entry and exit of firms, so that in equilibrium profits will always be zero.

B. Equilibrium in a Closed Economy

We can now proceed to analyze equilibrium in a closed economy described by the assumptions just laid out. The analysis proceeds in three stages. First I analyze consumer behavior to derive demand functions. Then profit-maximizing behavior by firms is derived, treating the number of firms as given. Finally, the assumption of free entry is used to determine the equilibrium number of firms.

²To be fully rigorous, we would have to use the concept of a continuum of potential products.

The reason that a Chamberlinian approach is useful here is that, in spite of imperfect competition, the equilibrium of the model is determinate in all essential respects because the special nature of demand rules out strategic interdependence among firms. Because firms can costlessly differentiate their products, and all products enter symmetrically into demand, two firms will never want to produce the same product; each good will be produced by only one firm. At the same time, if the number of goods produced is large, the effect of the price of any one good on the demand for any other will be negligible. The result is that each firm can ignore the effect of its actions on other firms' behavior, eliminating the indeterminacies of oligopoly.

Consider, then, an individual maximizing (1) subject to a budget constraint. The first-order conditions from that maximum problem have the form

$$(5) \quad \theta c_i^{\theta-1} = \lambda p_i \quad i = 1, \dots, n$$

where p_i is the price of the i th good and λ is the shadow price on the budget constraint, that is, the marginal utility of income. Since all individuals are alike, (5) can be rearranged to show the demand curve for the i th good, which we have already argued is the demand curve facing the single firm producing that good:

$$(6) \quad p_i = \theta \lambda^{-1} (x_i / L)^{\theta-1} \quad i = 1, \dots, n$$

Provided that there are a large number of goods being produced, the pricing decision of any one firm will have a negligible effect on the marginal utility of income. In that case, (6) implies that each firm faces a demand curve with an elasticity of $1/(1-\theta)$, and the profit-maximizing price is therefore

$$(7) \quad p_i = \theta^{-1} \beta w \quad i = 1, \dots, n$$

where w is the wage rate, and prices and wages can be defined in terms of any (common!) unit. Note that since θ , β , and w are the same for all firms, prices are the same

for all goods and we can adopt the shorthand $p=p_i$ for all i .

The price p is independent of output given the special assumptions about cost and utility (which is the reason for making these particular assumptions). To determine profitability, however, we need to look at output. Profits of the firm producing good i are

$$(8) \quad \pi_i = px_i - \{\alpha + \beta x_i\}w \quad i = 1, \dots, n$$

If profits are positive, new firms will enter, causing the marginal utility of income to rise and profits to fall until profits are driven to zero. In equilibrium, then $\pi=0$, implying for the output of a representative firm:

$$(9) \quad x_i = \alpha/(p/w - \beta) = \alpha\theta/\beta(1-\theta)$$

$$i = 1, \dots, n$$

Thus output per firm is determined by the zero-profit condition. Again, since α , β , and θ are the same for all firms we can use the shorthand $x=x_i$ for all i .

Finally, we can determine the number of goods produced by using the condition of full employment. From (4) and (9), we have

$$(10) \quad n = \frac{L}{\alpha + \beta x} = \frac{L(1-\theta)}{\alpha}$$

C. Effects of Trade

Now suppose that two countries of the kind just analyzed open trade with one another at zero transportation cost. To make the point most clearly, suppose that the countries have the same tastes and technologies; since we are in a one-factor world there cannot be any differences in factor endowments. What will happen?

In this model there are none of the conventional reasons for trade; but there will nevertheless be both trade and gains from trade. Trade will occur because, in the presence of increasing returns, each good (i.e., each differentiated product) will be produced in only one country—for the same reasons that each good is produced by only one firm. Gains from trade will occur because the world economy will produce a

greater diversity of goods than would either country alone, offering each individual a wider range of choice.

We can easily characterize the world economy's equilibrium. The symmetry of the situation ensures that the two countries will have the same wage rate, and that the price of any good produced in either country will be the same. The number of goods produced in each country can be determined from the full-employment condition

$$(11) \quad n = L(1-\theta)/\alpha; \quad n^* = L^*(1-\theta)/\alpha$$

where L^* is the labor force of the second country and n^* the number of goods produced there.

Individuals will still maximize the utility function (1), but they will now distribute their expenditure over both the n goods produced in the home country and the n^* goods produced in the foreign country. Because of the extended range of choice, welfare will increase even though the "real wage" w/p (i.e., the wage rate in terms of a representative good) remains unchanged. Also, the symmetry of the problem allows us to determine trade flows. It is apparent that individuals in the home country will spend a fraction $n^*/(n+n^*)$ of their income on foreign goods, while foreigners spend $n/(n+n^*)$ of their income on home country products. Thus the value of home country imports measured in wage units is $Ln^*/(n+n^*) = LL^*/(L+L^*)$. This equals the value of foreign country imports, confirming that with equal wage rates in the two countries we will have balance-of-payments equilibrium.

Notice, however, that while the *volume* of trade is determinate, the *direction* of trade—which country produces which goods—is not. This indeterminacy seems to be a general characteristic of models in which trade is a consequence of economies of scale. One of the convenient features of the models considered in this paper is that nothing important hinges on who produces what within a group of differentiated products. There is an indeterminacy, but it doesn't matter. This result might not hold up in less special models.

Finally, I should note a peculiar feature of the effects of trade in this model. Both before and after trade, equation (9) holds; that is, there is no effect of trade on the scale of production, and the gains from trade come solely through increased product diversity. This is an unsatisfactory result. In another paper I have developed a slightly different model in which trade leads to an increase in scale of production as well as an increase in diversity.³ That model is, however, more difficult to work with, so that it seems worth sacrificing some realism to gain tractability here.

II. Transport Costs

In this section I extend the model to allow for some transportation costs. This is not in itself an especially interesting extension although the main result—that the larger country will, other things equal, have the higher wage rate—is somewhat surprising. The main purpose of the extension is, however, to lay the groundwork for the analysis of home market effects in the next section. (These effects can obviously occur only if there are transportation costs.) I begin by describing the behavior of individual agents, then analyze the equilibrium.

A. Individual Behavior

Consider a world consisting of two countries of the type analyzed in Section I, able to trade but only at a cost. Transportation costs will be assumed to be of the "iceberg" type, that is, only a fraction g of any good shipped arrives, with $1-g$ lost in transit. This is a major simplifying assumption, as will be seen below.

³To get an increase in scale, we must assume that the demand facing each individual firm becomes more elastic as the number of firms increases, whereas in this model the elasticity of demand remains unchanged. Increasing elasticity of demand when the variety of products grows seems plausible, since the more finely differentiated are the products, the better substitutes they are likely to be for one another. Thus an increase in scale as well as diversity is probably the "normal" case. The constant elasticity case, however, is much easier to work with, which is my reason for using it in this paper.

An individual in the home country will have a choice over n products produced at home and n^* products produced abroad. The price of a domestic product will be the same as that received by the producer p . Foreign products, however, will cost more than the producer's price; if foreign firms charge p^* , home country consumers will have to pay the c.i.f. price $\hat{p}^* = p^*/g$. Similarly, foreign buyers of domestic products will pay $\hat{p} = p/g$.

Since the prices to consumers of goods of different countries will in general not be the same, consumption of each imported good will differ from consumption of each domestic good. Home country residents, for example, in maximizing utility will consume $(p/\hat{p}^*)^{1/(1-\theta)}$ units of a representative imported good for each unit of a representative domestic good they consume.

To determine world equilibrium, however, it is not enough to look at consumption; we must also take into account the quantities of goods used up in transit. If a domestic resident consumes one unit of a foreign good, his combined direct and indirect demand is for $1/g$ units. For determining total demand, then, we need to know the ratio of total demand by domestic residents for each foreign product to demand for each domestic product. Letting σ denote this ratio, and σ^* the corresponding ratio for the other country, we can show that

$$(12) \quad \sigma = (p/p^*)^{1/(1-\theta)} g^{\theta/(1-\theta)}$$

$$\sigma^* = (p/p^*)^{-1/(1-\theta)} g^{\theta/(1-\theta)}$$

The overall demand pattern of each individual can then be derived from the requirement that his spending just equal his wage; that is, in the home country we must have $(np + \sigma n^* p^*)d = w$, where d is the consumption of a representative domestic good; and similarly in the foreign country.

This behavior of individuals can now be used to analyze the behavior of firms. The important point to notice is that the elasticity of export demand facing any given firm is $1/(1-\theta)$, which is the same as the elasticity of domestic demand. Thus transportation

costs have no effect on firms' pricing policy; and the analysis of Section I can be carried out as before, showing that transportation costs also have no effect on the number of firms or output per firm in either country.

Writing out these conditions again, we have

$$(13) \quad p = w\beta/\theta; \quad p^* = w^*\beta/\theta$$

$$n = L(1-\theta)/\alpha; \quad n^* = L^*(1-\theta)/\alpha$$

The only way in which introducing transportation costs modifies the results of Section I is in allowing the possibility that wages may not be equal in the two countries; the number and size of firms are not affected. This strong result depends on the assumed form of the transport costs, which shows at the same time how useful and how special the assumed form is.

B. Determination of Equilibrium

The model we have been working with has a very strong structure—so strong that transport costs have no effect on either the numbers of goods produced in the countries, n and n^* , or on the prices relative to wages, p/w and p^*/w^* . The only variable which can be affected is the relative wage rate $w/w^* = \omega$, which no longer need be equal to one.

We can determine ω by looking at any one of three equivalent market-clearing conditions: (i) equality of demand and supply for home country labor; (ii) equality of demand and supply for foreign country labor; (iii) balance-of-payments equilibrium. It will be easiest to work in terms of the balance of payments. If we combine (12) with the other equations of the model, it can be shown that the home country's balance of payments, measured in wage units of the other country, is

$$(14) \quad B = \frac{\sigma^* n \omega}{\sigma^* n + n^*} L^* - \frac{\sigma n^*}{n + \sigma n^*} \omega L \\ = \omega LL^* \left[\frac{\sigma^*}{\sigma^* L + L^*} - \frac{\sigma}{L + \sigma L^*} \right]$$

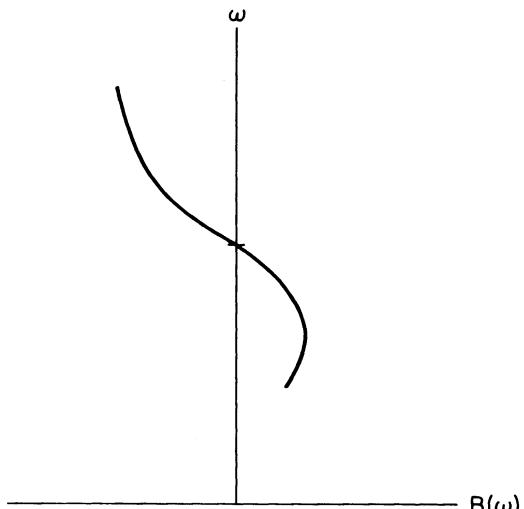


FIGURE 1

Since σ and σ^* are both functions of $p/p^* = \omega$, the condition $B=0$ can be used to determine the relative wage. The function $B(\omega)$ is illustrated in Figure 1. The relative wage $\bar{\omega}$ is that relative wage at which the expression in brackets in (4) is zero, and at which trade is therefore balanced. Since σ is an increasing function of ω and σ^* a decreasing function of ω , $B(\omega)$ will be negative (positive) if and only if ω is greater (less) than $\bar{\omega}$, which shows that $\bar{\omega}$ is the unique equilibrium relative wage.

We can use this result to establish a simple proposition: *that the larger country, other things equal, will have the higher wage*. To see this, suppose that we were to compute $B(\omega)$ for $\omega=1$. In that case we have $\sigma=\sigma^*<1$. The expression for the balance of payments reduces to

$$(14') \quad B = LL^* \left[\frac{1}{\sigma L + L^*} - \frac{1}{L + \sigma L^*} \right]$$

But (14') will be positive if $L > L^*$, negative if $L < L^*$. This means that the equilibrium relative wage ω must be greater than one if $L > L^*$, less than one if $L < L^*$.

This is an interesting result. In a world characterized by economies of scale, one

would expect workers to be better off in larger economies, because of the larger size of the local market. In this model, however, there is a secondary benefit in the form of better terms of trade with workers in the rest of the world. This does, on reflection, make intuitive sense. If production costs were the same in both countries, it would always be more profitable to produce near the larger market, thus minimizing transportation costs. To keep labor employed in both countries, this advantage must be offset by a wage differential.

III. "Home Market" Effects on the Pattern of Trade

In a world characterized both by increasing returns and by transportation costs, there will obviously be an incentive to concentrate production of a good near its largest market, even if there is some demand for the good elsewhere. The reason is simply that by concentrating production in one place, one can realize the scale economies, while by locating near the larger market, one minimizes transportation costs. This point—which is more often emphasized in location theory than in trade theory—is the basis for the common argument that countries will tend to export those kinds of products for which they have relatively large domestic demand. Notice that this argument is wholly dependent on increasing returns; in a world of diminishing returns strong domestic demand for a good will tend to make it an import rather than an export. But the point does not come through clearly in models where increasing returns take the form of external economies (see W. M. Corden). One of the main contributions of the approach developed in this paper is that by using this approach the home market can be given a simple formal justification.

I will begin by extending the basic closed economy model to one in which there are two industries (with many differentiated products within each industry). It will then be shown for a simple case that when two countries of this kind trade, each will be a net exporter in the industry for whose prod-

ucts it has the relatively larger demand. Finally, some extensions and generalizations will be discussed.

A. A Two-Industry Economy

As in Section I, we begin by analyzing a closed economy. Assume that there are two classes of products, *alpha* and *beta*, with many potential products within each class. A tilde will distinguish *beta* products from *alpha* products; for example, consumption of products in the first class will be represented as c_1, \dots, c_n while consumption of products in second are $\tilde{c}_1, \dots, \tilde{c}_n$.

Demand for the two classes of products will be assumed to arise from the presence of two groups in the population.⁴ There will be one group with L members, which derives utility only from consumption of *alpha* products; and another group with \tilde{L} members, deriving utility only from *beta* products. The utility functions of representative members of the two classes may be written

$$(15) \quad U = \sum_i c_i^\theta; \quad \tilde{U} = \sum_j \tilde{c}_j^\theta \quad 0 < \theta < 1$$

For simplicity assume that not only the form of the utility function but the parameter θ is the same for both groups.

On the cost side, the two kinds of products will be assumed to have identical cost functions:

$$(16) \quad l_i = \alpha + \beta x_i \quad i = 1, \dots, n$$

$$\tilde{l}_j = \alpha + \beta \tilde{x}_j \quad j = 1, \dots, \tilde{n}$$

where, l_i, \tilde{l}_j are labor used in production on typical goods in each class, and x_i, \tilde{x}_j are total outputs of the goods.

The demand conditions now depend on the population shares. By analogy with (3),

⁴An alternative would be to have all people alike, with a taste for both kinds of goods. The results are similar. In fact, if each industry receives a fixed share of expenditure, they will be identical.

we have

$$(17) \quad x_i = Lc_i \quad i = 1, \dots, n$$

$$\tilde{x}_j = \tilde{L}\tilde{c}_j \quad j = 1, \dots, \tilde{n}$$

The full-employment condition, however, applies to the economy as a whole:

$$(18) \quad \sum_{i=1}^n l_i + \sum_{j=1}^{\tilde{n}} \tilde{l}_j = L + \tilde{L}$$

Finally, we continue to assume free entry, driving profits to zero. Now it is immediately apparent that the economy described by equations (15)–(18) is very similar to the economy described in equations (1)–(4). The price and output of a representative good—of either class—and the total number of products $n+\tilde{n}$ are determined just as if all goods belonged to a single industry. The only modification we must make to the results of Section I is that we must divide the total production into two industries. A simple way of doing this is to note that the sales of each industry must equal the income of the appropriate group in the population:

$$(19) \quad npx = wL; \quad \tilde{n}\tilde{p}\tilde{x} = \tilde{w}\tilde{L}$$

But wages of the two groups must be equal, as must the prices and outputs of any products of either industry. So this reduces to the result $n/\tilde{n} = L/\tilde{L}$: the shares of the industries in the value of output equal the shares of the two demographic groups in the population.

This extended model clearly differs only trivially from the model developed in Section I when the economy is taken to be closed. When two such economies are allowed to trade, however, the extension allows some interesting results.

B. Demand and the Trade Pattern: A Simple Case

We can begin by considering a particular case of trade between a pair of two-industry countries in which the role of the domestic

market appears particularly clearly. Suppose that there are two countries of the type just described, and that they can trade with transport costs of the type analyzed in Section II.

In the home country, some fraction f of the population will be consumers of *alpha* products. The crucial simplification I will make is to assume that the other country is a *mirror image* of the home country. The labor forces will be assumed to be equal, so that

$$(20) \quad L + \tilde{L} = L^* + \tilde{L}^* = \bar{L}$$

But in the foreign country the population shares will be reversed, so that we have

$$(21) \quad L = f\bar{L}; \quad L^* = (1-f)\bar{L}$$

If f is greater than one-half, then the home country has the larger domestic market for the *alpha* industry's products; and conversely. In this case there is a very simple home market proposition: *that the home country will be a net exporter of the first industry's products if $f > 0.5$* . This proposition turns out to be true.

The first step in showing this is to notice that this is a wholly symmetrical world, so that wage rates will be equal, as will the output and prices of all goods. (The case was constructed for that purpose.) It follows that the ratio of demand for each imported product to the demand for each domestic product is the same in both countries.

$$(22) \quad \sigma = \sigma^* = g^{\theta/(1-\theta)} < 1$$

Next we want to determine the pattern of production. The expenditure on goods in an industry is the sum of domestic residents' and foreigners' expenditures on the goods, so we can write the expressions

$$(23) \quad npx = \frac{n}{n + \sigma n^*} wL + \frac{\sigma n}{\sigma n + n^*} wL^*$$

$$n^*px = \frac{\sigma n^*}{n + \sigma n^*} wL + \frac{n^*}{\sigma n + n^*} wL^*$$

where the price p of each product and the

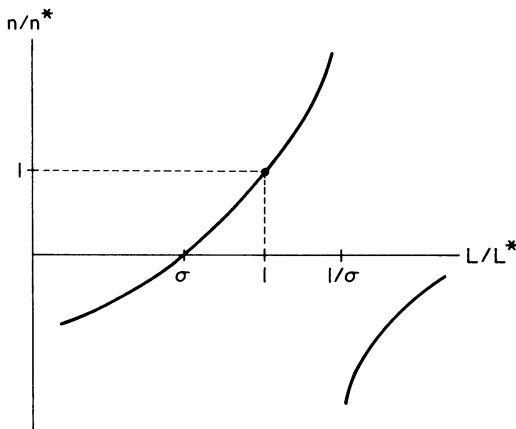


FIGURE 2

output x are the same in the two countries. We can use (23) to determine the relative number of products produced in each country, n/n^* .

To see this, suppose *provisionally* that some products in the *alpha* industry are produced in both countries; i.e., $n > 0$, $n^* > 0$. We can then divide the equations (23) through by n and n^* , respectively, and rearrange to get

$$(24) \quad L/L^* = (n + \sigma n^*) / (\sigma n + n^*)$$

which can be rearranged to give

$$(25) \quad n/n^* = \frac{L/L^* - \sigma}{1 - \sigma L/L^*}.$$

Figure 2 shows the relationship (25). If $L/L^* = 1$, so does n/n^* ; that is, if the demand patterns of the two countries are the same, their production patterns will also be the same, as we would expect. And as the relative size of either country's home market rises for *alpha* goods, so does its domestic production, as long as L/L^* lies in the range $\sigma < L/L^* < 1/\sigma$.

Outside that range, (25) appears to give absurd results. Recall, however, that the derivation of (24) was made on the provisional assumption that n and n^* were both non-zero. Clearly, if L/L^* lies outside the range

from σ to $1/\sigma$, this assumption is not valid. What the figure suggests is that if L/L^* is less than σ , $n=0$; the home country specialized entirely in *beta* products, producing no *alpha* products (while the foreign country produces only *alpha* products). Conversely, if L/L^* is greater than $1/\sigma$, $n^*=0$, and we have the opposite pattern of specialization.

We can easily demonstrate that this solution is in fact an equilibrium. Suppose that the home country produced no *alpha* products, and that a firm attempted to start production of a single product. This firm's profit-maximizing f.o.b. price would be the same as that of the foreign firm's. But its sales would be less, in the ratio

$$\frac{\sigma^{-1}L + \sigma L^*}{L + L^*} < 1$$

Thus such a firm could not compete.

This gives us our first result on the effect of the home market. It says that if the two countries have sufficiently dissimilar tastes each will specialize in the industry for which it has the larger home market. Obviously, also, each will be a net exporter of the class of goods in which it specializes. Thus the idea that the pattern of exports is determined by the home market is quite nicely confirmed.

We also get some illuminating results on the conditions under which specialization will be incomplete. Incomplete specialization and two-way trade within the two classes of products will occur if the relative size of the domestic markets for *alpha* goods lies in the range from σ to $1/\sigma$, where $\sigma = g^{\theta/(1-\theta)}$. But g measures transportation costs, while $\theta/(1-\theta)$ is, in equilibrium, the ratio of variable to fixed costs;⁵ that is, it is an index of the importance of scale economies. So we have shown that the possibility of incomplete specialization is greater, the greater are transport costs and the less important are economies of scale.

A final result we can take from this special case concerns the pattern of trade when

⁵One can see this by rearranging equation (9) to get $\beta x/\alpha = \theta/(1-\theta)$.

specialization is incomplete. In this case each country will both import and export products in *both* classes (though not the same products). But it remains true that, if one country has the larger home market for *alpha* producers, it will be a *net exporter* in the *alpha* class and a *net importer* in the other. To see this, note that we can write the home country's trade balance in *alpha* products as

$$(26) \quad B_\alpha = \frac{\sigma n}{\sigma n + n^*} wL^* - \frac{\sigma n^*}{n + \sigma n^*} wL$$

$$= wL^* \left[\frac{\sigma n}{\sigma n + n^*} - \frac{\sigma n^*}{n + \sigma n^*} \frac{L}{L^*} \right]$$

$$= \frac{\sigma wL^*}{\sigma n + n^*} [n - n^*]$$

where we used (24) to eliminate the relative labor supplies. This says that the sign of the trade balance depends on whether the number of *alpha* products produced in the home country is more or less than the number produced abroad. But we have already seen that n/n^* is an increasing function of L/L^* in the relevant range. So the country with the larger home market for the *alpha*-type products will be a *net exporter* of those goods, even if specialization is not complete.

C. Generalizations and Extensions

The analysis we have just gone through shows that there is some justification for the idea that countries export what they have home markets for. The results were arrived at, however, only for a special case designed to make matters as simple as possible. Our next question must be the extent to which these results generalize.

One way in which generalization might be pursued is by abandoning the "mirror image" assumption: we can let the countries have arbitrary populations and demand patterns, while retaining all the other assumptions of the model. It can be shown that in that case, although the derivations become more complicated, the basic home market result is unchanged. Each country will be a *net exporter* in the industry for whose goods it has a relatively larger demand. The dif-

ference is that wages will in general not be equal; in particular, smaller countries with absolutely smaller markets for both kinds of goods will have to compensate for this disadvantage with lower wages.

Another, perhaps more interesting, generalization would be to abandon the assumed symmetry between the industries. Again, we would like to be able to make sense of some arguments made by practical men. For example, is it true that large countries will have an advantage in the production and export of goods whose production is characterized by sizeable economies of scale? This is an explanation which is sometimes given for the United States' position as an exporter of aircraft.

A general analysis of the effects of asymmetry between industries would run to too great a length. We can learn something, however, by considering another special case. Suppose that the *alpha* production is the same as in our last analysis, but that the production of *beta* goods is characterized by *constant* returns to scale and perfect competition. For simplicity, also assume that *beta* goods can be transported costlessly.

It is immediately apparent that in this case the possibility of trade in *beta* products will ensure that wage rates are equal. But this in turn means that we can apply the analysis of Part B, above, to the *alpha* industry. Whichever country has the larger market for the products of that industry will be a *net exporter* of *alpha* products and a *net importer* of *beta* products. In particular: if two countries have the same composition of demand, the larger country will be a *net exporter* of the products whose production involves economies of scale.

The analysis in this section has obviously been suggestive rather than conclusive. It relies heavily on very special assumptions and on the analysis of special cases. Nonetheless, the analysis does seem to confirm the idea that, in the presence of increasing returns, countries will tend to export the goods for which they have large domestic markets. And the implications for the pattern of trade are similar to those suggested by Steffan Linder, Grubel (1970), and others.

REFERENCES

- Bela Balassa**, *Trade Liberalization Among Industrial Countries*, New York 1967.
- W. M. Corden**, "A Note on Economies of Scale, the Size of the Domestic Market and the Pattern of Trade," in I. A. McDougall and R. H. Snape, eds., *Studies in International Economics*, Amsterdam 1970.
- A. Dixit and J. Stiglitz**, "Monopolistic Competition and Optimum Product Diversity," *Amer. Econ. Rev.*, June 1977, 67, 297–308.
- H. Grubel**, "Intra-Industry Specialization and the Pattern of Trade," *Can. J. Econ.*, Aug. 1967, 33, 374–388.
- _____, "The Theory of Intra-Industry Trade," in I. A. McDougall and R. H. Snape, eds., *Studies in International Economics*, Amsterdam 1970.
- G. Hufbauer and J. Chilas**, "Specialization by Industrial Countries: Extent and Consequences," in Herbert Giersch, ed., *The International Division of Labor*, Tübingen 1974.
- I. Kravis**, "The Current Case for Import Limitations," in *United States Economic Policy in an Interdependent World, Commission on International Trade and Investment Policy*, Washington 1971.
- P. Krugman**, "Increasing Returns, Monopolistic Competition, and International Trade," *J. Int. Econ.*, Nov. 1979, 9, 469–80.
- Steffan Linder**, *An Essay on Trade and Transformation*, New York 1961.

Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?

By ROBERT J. SHILLER*

A simple model that is commonly used to interpret movements in corporate common stock price indexes asserts that real stock prices equal the present value of rationally expected or optimally forecasted future real dividends discounted by a constant real discount rate. This valuation model (or variations on it in which the real discount rate is not constant but fairly stable) is often used by economists and market analysts alike as a plausible model to describe the behavior of aggregate market indexes and is viewed as providing a reasonable story to tell when people ask what accounts for a sudden movement in stock price indexes. Such movements are then attributed to "new information" about future dividends. I will refer to this model as the "efficient markets model" although it should be recognized that this name has also been applied to other models.

It has often been claimed in popular discussions that stock price indexes seem too "volatile," that is, that the movements in stock price indexes could not realistically be attributed to any objective new information, since movements in the price indexes seem to be "too big" relative to actual subsequent events. Recently, the notion that financial asset prices are too volatile to accord with efficient markets has received some econometric support in papers by Stephen LeRoy

and Richard Porter on the stock market, and by myself on the bond market.

To illustrate graphically why it seems that stock prices are too volatile, I have plotted in Figure 1 a stock price index p_t with its *ex post* rational counterpart p_t^* (data set 1).¹ The stock price index p_t is the real Standard and Poor's Composite Stock Price Index (detrended by dividing by a factor proportional to the long-run exponential growth path) and p_t^* is the present discounted value of the actual subsequent real dividends (also as a proportion of the same long-run growth factor).² The analogous series for a modified Dow Jones Industrial Average appear in Figure 2 (data set 2). One is struck by the smoothness and stability of the *ex post* rational price series p_t^* when compared with the actual price series. This behavior of p^* is due to the fact that the present value relation relates p^* to a long-weighted moving average of dividends (with weights corresponding to discount factors) and moving averages tend to smooth the series averaged. Moreover, while real dividends did vary over this sample period, they did not vary long enough or far enough to cause major movements in p^* . For example, while one normally thinks of the Great Depression as a time when business was bad, real dividends were substantially below their long-run exponential growth path (i.e., 10–25 percent below the

*Associate professor, University of Pennsylvania, and research associate, National Bureau of Economic Research. I am grateful to Christine Amsler for research assistance, and to her as well as Benjamin Friedman, Irwin Friend, Sanford Grossman, Stephen LeRoy, Stephen Ross, and Jeremy Siegel for helpful comments. This research was supported by the National Bureau of Economic Research as part of the Research Project on the Changing Roles of Debt and Equity in Financing U.S. Capital Formation sponsored by the American Council of Life Insurance and by the National Science Foundation under grant SOC-7907561. The views expressed here are solely my own and do not necessarily represent the views of the supporting agencies.

¹The stock price index may look unfamiliar because it is deflated by a price index, expressed as a proportion of the long-run growth path and only January figures are shown. One might note, for example, that the stock market decline of 1929–32 looks smaller than the recent decline. In real terms, it was. The January figures also miss both the 1929 peak and 1932 trough.

²The price and dividend series as a proportion of the long-run growth path are defined below at the beginning of Section I. Assumptions about public knowledge or lack of knowledge of the long-run growth path are important, as shall be discussed below. The series p^* is computed subject to an assumption about dividends after 1978. See text and Figure 3 below.

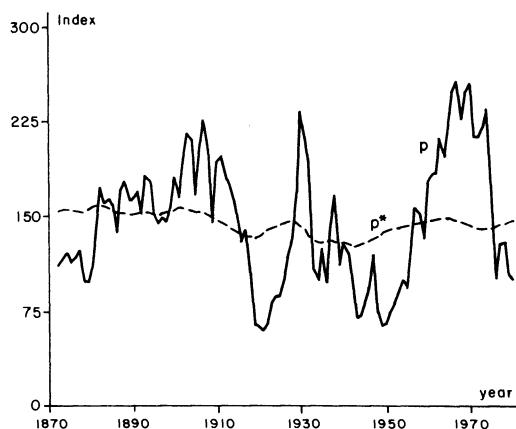


FIGURE 1

Note: Real Standard and Poor's Composite Stock Price Index (solid line p) and *ex post* rational price (dotted line p^*), 1871–1979, both detrended by dividing a long-run exponential growth factor. The variable p^* is the present value of actual subsequent real detrended dividends, subject to an assumption about the present value in 1979 of dividends thereafter. Data are from Data Set 1, Appendix.

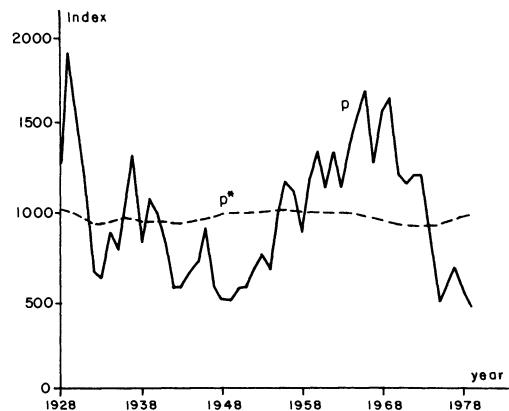


FIGURE 2

Note: Real modified Dow Jones Industrial Average (solid line p) and *ex post* rational price (dotted line p^*), 1928–1979, both detrended by dividing by a long-run exponential growth factor. The variable p^* is the present value of actual subsequent real detrended dividends, subject to an assumption about the present value in 1979 of dividends thereafter. Data are from Data Set 2, Appendix.

growth path for the Standard and Poor's series, 16–38 percent below the growth path for the Dow Series) only for a few depression years: 1933, 1934, 1935, and 1938. The moving average which determines p^* will smooth out such short-run fluctuations. Clearly the stock market decline beginning in 1929 and ending in 1932 could not be rationalized in terms of subsequent dividends! Nor could it be rationalized in terms of subsequent earnings, since earnings are relevant in this model only as indicators of later dividends. Of course, the efficient markets model does not say $p=p^*$. Might one still suppose that this kind of stock market crash was a rational mistake, a forecast error that rational people might make? This paper will explore here the notion that the very volatility of p (i.e., the tendency of big movements in p to occur again and again) implies that the answer is no.

To give an idea of the kind of volatility comparisons that will be made here, let us consider at this point the simplest inequality which puts limits on one measure of volatility: the standard deviation of p . The efficient markets model can be described as asserting

that $p_t = E_t(p_t^*)$, i.e., p_t is the mathematical expectation conditional on all information available at time t of p_t^* . In other words, p_t is the optimal forecast of p_t^* . One can define the forecast error as $u_t = p_t^* - p_t$. A fundamental principle of optimal forecasts is that the forecast error u_t must be uncorrelated with the forecast; that is, the covariance between p_t and u_t must be zero. If a forecast error showed a consistent correlation with the forecast itself, then that would in itself imply that the forecast could be improved. Mathematically, it can be shown from the theory of conditional expectations that u_t must be uncorrelated with p_t .

If one uses the principle from elementary statistics that the variance of the sum of two uncorrelated variables is the sum of their variances, one then has $\text{var}(p^*) = \text{var}(u) + \text{var}(p)$. Since variances cannot be negative, this means $\text{var}(p) \leq \text{var}(p^*)$ or, converting to more easily interpreted standard deviations,

$$(1) \quad \sigma(p) \leq \sigma(p^*)$$

This inequality (employed before in the

papers by LeRoy and Porter and myself) is violated dramatically by the data in Figures 1 and 2 as is immediately obvious in looking at the figures.³

This paper will develop the efficient markets model in Section I to clarify some theoretical questions that may arise in connection with the inequality (1) and some similar inequalities will be derived that put limits on the standard deviation of the innovation in price and the standard deviation of the change in price. The model is restated in innovation form which allows better understanding of the limits on stock price volatility imposed by the model. In particular, this will enable us to see (Section II) that the standard deviation of Δp is highest when information about dividends is revealed smoothly and that if information is revealed in big lumps occasionally the price series may have higher kurtosis (fatter tails) but will have *lower* variance. The notion expressed by some that earnings rather than dividend data should be used is discussed in Section III, and a way of assessing the importance of time variation in real discount rates is shown in Section IV. The inequalities are compared with the data in Section V.

This paper takes as its starting point the approach I used earlier (1979) which showed evidence suggesting that long-term bond yields are too volatile to accord with simple expectations models of the term structure of interest rates.⁴ In that paper, it was shown

³Some people will object to this derivation of (1) and say that one might as well have said that $E_t(p_t) = p_t^*$, i.e., that forecasts are correct "on average," which would lead to a reversal of the inequality (1). This objection stems, however, from a misinterpretation of conditional expectations. The subscript t on the expectations operator E means "taking as given (i.e., nonrandom) all variables known at time t ." Clearly, p_t is known at time t and p_t^* is not. In practical terms, if a forecaster gives as his forecast anything other than $E_t(p_t^*)$, then high forecast is not optimal in the sense of expected squared forecast error. If he gives a forecast which equals $E_t(p_t^*)$ only on average, then he is adding random noise to the optimal forecast. The amount of noise apparent in Figures 1 or 2 is extraordinary. Imagine what we would think of our local weather forecaster if, say, actual local temperatures followed the dotted line and his forecasts followed the solid line!

⁴This analysis was extended to yields on preferred stocks by Christine Amsler.

how restrictions implied by efficient markets on the cross-covariance function of short-term and long-term interest rates imply inequality restrictions on the spectra of the long-term interest rate series which characterize the smoothness that the long rate should display. In this paper, analogous implications are derived for the volatility of stock prices, although here a simpler and more intuitively appealing discussion of the model in terms of its innovation representation is used. This paper also has benefited from the earlier discussion by LeRoy and Porter which independently derived some restrictions on security price volatility implied by the efficient markets model and concluded that common stock prices are too volatile to accord with the model. They applied a methodology in some ways similar to that used here to study a stock price index and individual stocks in a sample period starting after World War II.

It is somewhat inaccurate to say that this paper attempts to contradict the extensive literature of efficient markets (as, for example, Paul Cootner's volume on the random character of stock prices, or Eugene Fama's survey).⁵ Most of this literature really examines different properties of security prices. Very little of the efficient markets literature bears directly on the characteristic feature of the model considered here: that expected *real* returns for the aggregate stock market are constant through time (or approximately so). Much of the literature on efficient markets concerns the investigation of nominal "profit opportunities" (variously defined) and whether transaction costs prohibit their exploitation. Of course, if real stock prices are "too volatile" as it is defined here, then there may well be a sort of real profit opportunity. Time variation in expected real interest rates does not itself imply that any

⁵It should not be inferred that the literature on efficient markets uniformly supports the notion of efficiency put forth there, for example, that no assets are dominated or that no trading rule dominates a buy and hold strategy, (for recent papers see S. Basu; Franco Modigliani and Richard Cohn; William Brainard, John Shoven and Lawrence Weiss; and the papers in the symposium on market efficiency edited by Michael Jensen).

trading rule dominates a buy and hold strategy, but really large variations in expected returns might seem to suggest that such a trading rule exists. This paper does not investigate this, or whether transaction costs prohibit its exploitation. This paper is concerned, however, instead with a more interesting (from an economic standpoint) question: what accounts for movements in real stock prices and can they be explained by new information about subsequent real dividends? If the model fails due to excessive volatility, then we will have seen a new characterization of how the simple model fails. The characterization is not equivalent to other characterizations of its failure, such as that one-period holding returns are forecastable, or that stocks have not been good inflation hedges recently.

The volatility comparisons that will be made here have the advantage that they are insensitive to misalignment of price and dividend series, as may happen with earlier data when collection procedures were not ideal. The tests are also not affected by the practice, in the construction of stock price and dividend indexes, of dropping certain stocks from the sample occasionally and replacing them with other stocks, so long as the volatility of the series is not misstated. These comparisons are thus well suited to existing long-term data in stock price averages. The robustness that the volatility comparisons have, coupled with their simplicity, may account for their popularity in casual discourse.

I. The Simple Efficient Markets Model

According to the simple efficient markets model, the real price P_t of a share at the beginning of the time period t is given by

$$(2) \quad P_t = \sum_{k=0}^{\infty} \gamma^{k+1} E_t D_{t+k} \quad 0 < \gamma < 1$$

where D_t is the real dividend paid at (let us say, the end of) time t , E_t denotes mathematical expectation conditional on information available at time t , and γ is the constant real discount factor. I define the constant

real interest rate r so that $\gamma = 1/(1+r)$. Information at time t includes P_t and D_t and their lagged values, and will generally include other variables as well.

The one-period holding return $H_t \equiv (\Delta P_{t+1} + D_t)/P_t$ is the return from buying the stock at time t and selling it at time $t+1$. The first term in the numerator is the capital gain, the second term is the dividend received at the end of time t . They are divided by P_t to provide a rate of return. The model (2) has the property that $E_t(H_t) = r$.

The model (2) can be restated in terms of series as a proportion of the long-run growth factor: $p_t = P_t/\lambda^{-T}$, $d_t = D_t/\lambda^{t+1-T}$ where the growth factor is $\lambda^{-T} = (1+g)^{t-T}$, g is the rate of growth, and T is the base year. Dividing (2) by λ^{-T} and substituting one finds⁶

$$(3) \quad p_t = \sum_{k=0}^{\infty} (\lambda\gamma)^{k+1} E_t d_{t+k} \\ = \sum_{k=0}^{\infty} \bar{\gamma}^{k+1} E_t d_{t+k}$$

The growth rate g must be less than the discount rate r if (2) is to give a finite price, and hence $\bar{\gamma} \equiv \lambda\gamma < 1$, and defining \bar{r} by $\bar{\gamma} \equiv 1/(1+\bar{r})$, the discount rate appropriate for the p_t and d_t series is $\bar{r} > 0$. This discount rate \bar{r} is, it turns out, just the mean dividend divided by the mean price, i.e., $\bar{r} = E(d)/E(p)$.⁷

⁶No assumptions are introduced in going from (2) to (3), since (3) is just an algebraic transformation of (2). I shall, however, introduce the assumption that d_t is jointly stationary with information, which means that the (unconditional) covariance between d_t and z_{t-k} , where z_t is any information variable (which might be d_t itself or p_t), depends only on k , not t . It follows that we can write expressions like $\text{var}(p)$ without a time subscript. In contrast, a realization of the random variable the *conditional expectation* $E_t(d_{t+k})$ is a function of time since it depends on information at time t . Some stationarity assumption is necessary if we are to proceed with any statistical analysis.

⁷Taking unconditional expectations of both sides of (3) we find

$$E(p) = \frac{\bar{\gamma}}{1-\bar{\gamma}} E(d)$$

using $\bar{\gamma} = 1/(1+\bar{r})$ and solving we find $\bar{r} = E(d)/E(p)$.

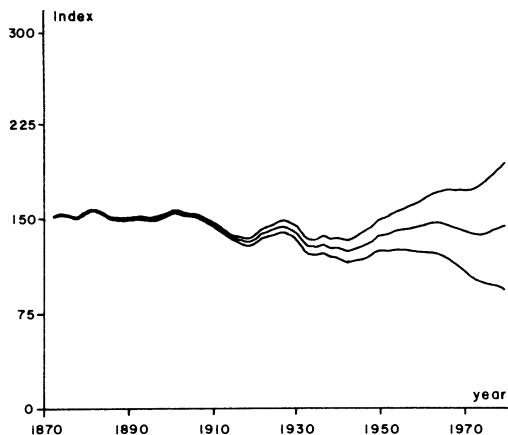


FIGURE 3

Note: Alternative measures of the *ex post* rational price p^* , obtained by alternative assumptions about the present value in 1979 of dividends thereafter. The middle curve is the p^* series plotted in Figure 1. The series are computed recursively from terminal conditions using dividend series d of Data Set 1.

We may also write the model as noted above in terms of the *ex post* rational price series p_t^* (analogous to the *ex post* rational interest rate series that Jeremy Siegel and I used to study the Fisher effect, or that I used to study the expectations theory of the term structure). That is, p_t^* is the present value of actual subsequent dividends:

$$(4) \quad p_t = E_t(p_t^*)$$

$$\text{where } p_t^* = \sum_{k=0}^{\infty} \bar{\gamma}^{k+1} d_{t+k}$$

Since the summation extends to infinity, we never observe p_t^* without some error. However, with a long enough dividend series we may observe an approximate p_t^* . If we choose an arbitrary value for the terminal value of p_t^* (in Figures 1 and 2, p^* for 1979 was set at the average detrended real price over the sample) then we may determine p_t^* recursively by $p_t^* = \bar{\gamma}(p_{t+1}^* + d_t)$ working backward from the terminal date. As we move back from the terminal date, the importance of the terminal value chosen declines. In data set (1) as shown in Figure 1, $\bar{\gamma}$ is .954 and $\bar{\gamma}^{108} = .0063$ so that at the beginning of the sample the terminal value chosen has a negligible weight in the determination of p_t^* . If we had chosen a different terminal condi-

TABLE 1—DEFINITIONS OF PRINCIPAL SYMBOLS

γ	= real discount factor for series before detrending;	
	$\gamma = 1/(1+r)$	
$\bar{\gamma}$	= real discount factor for detrended series; $\bar{\gamma} = \lambda\gamma$	
D_t	= real dividend accruing to stock index (before detrending)	
d_t	= real detrended dividend; $d_t = D_t/\lambda^{t+1-T}$	
Δ	= first difference operator $\Delta x_t \equiv x_t - x_{t-1}$	
δ_t	= innovation operator; $\delta_t x_{t+k} \equiv E_t x_{t+k} - E_{t-1} x_{t+k}; \delta_t x_t \equiv \delta_t x_t$	
E	= unconditional mathematical expectations operator.	
	$E(x)$ is the true (population) mean of x .	
E_t	= mathematical expectations operator conditional on information at time t ; $E_t x_t \equiv E(x_t	I_t)$ where I_t is the vector of information variables known at time t .
λ	= trend factor for price and dividend series; $\lambda \equiv 1+g$ where g is the long-run growth rate of price and dividends.	
P_t	= real stock price index (before detrending)	
p_t	= real detrended stock price index; $p_t = P_t/\lambda^{t-T}$	
p_t^*	= *ex post* rational stock price index (expression 4)	
r	= one-period real discount rate for series before detrending	
\bar{r}	= real discount rate for detrended series; $\bar{r} = (1-\bar{\gamma})/\bar{\gamma}$	
\bar{r}_2	= two-period real discount rate for detrended series; $\bar{r}_2 = (1+\bar{r})^2 - 1$	
t	= time (year)	
T	= base year for detrending and for wholesale price index; $p_T = P_T$ = nominal stock price index at time T	

tion, the result would be to add or subtract an exponential trend from the p^* shown in Figure 1. This is shown graphically in Figure 3, in which p^* is shown computed from alternative terminal values. Since the only thing we need know to compute p^* about dividends after 1978 is p^* for 1979, it does not matter whether dividends are "smooth" or not after 1978. Thus, Figure 3 represents our uncertainty about p^* .

There is yet another way to write the model, which will be useful in the analysis which follows. For this purpose, it is convenient to adopt notation for the innovation in a variable. Let us define the innovation operator $\delta_t \equiv E_t - E_{t-1}$ where E_t is the conditional expectations operator. Then for any variable X_t the term $\delta_t X_{t+k}$ equals $E_t X_{t+k} - E_{t-1} X_{t+k}$ which is the change in the conditional expectation of X_{t+k} that is made in response to new information arriving between $t-1$ and t . The time subscript t may be dropped so that δX_k denotes $\delta_t X_{t+k}$ and

δX denotes δX_0 or $\delta_t X_t$. Since conditional expectations operators satisfy $E_j E_k = E_{\min(j,k)}$ it follows that $E_{t-m} \delta_t X_{t+k} = E_{t-m} (E_t X_{t+k} - E_{t-1} X_{t+k}) = E_{t-m} X_{t+k} - E_{t-m} X_{t+k} = 0$, $m \geq 0$. This means that $\delta_t X_{t+k}$ must be uncorrelated for all k with all information known at time $t-1$ and must, since lagged innovations are information at time t , be uncorrelated with $\delta_{t'} X_{t+j}$, $t' < t$, all j , i.e., innovations in variables are serially uncorrelated.

The model implies that the innovation in price $\delta_t p_t$ is observable. Since (3) can be written $p_t = \bar{\gamma}(d_t + E_t p_{t+1})$, we know, solving, that $E_t p_{t+1} = p_t / \bar{\gamma} - d_t$. Hence $\delta_t p_t \equiv E_t p_t - E_{t-1} p_t = p_t + d_{t-1} - p_{t-1} / \bar{\gamma} = \Delta p_t + d_{t-1} - \bar{r} p_{t-1}$. The variable which we call $\delta_t p_t$ (or just δp) is the variable which Clive Granger and Paul Samuelson emphasized should, in contrast to $\Delta p_t \equiv p_t - p_{t-1}$, by efficient markets, be unforecastable. In practice, with our data, $\delta_t p_t$ so measured will approximately equal Δp_t .

The model also implies that the innovation in price is related to the innovations in dividends by

$$(5) \quad \delta_t p_t = \sum_{k=0}^{\infty} \bar{\gamma}^{k+1} \delta_t d_{t+k}$$

This expression is identical to (3) except that δ_t replaces E_t . Unfortunately, while $\delta_t p_t$ is observable in this model, the $\delta_t d_{t+k}$ terms are not directly observable, that is, we do not know when the public gets information about a particular dividend. Thus, in deriving inequalities below, one is obliged to assume the "worst possible" pattern of information accrual.

Expressions (2)–(5) constitute four different representations of the same efficient markets model. Expressions (4) and (5) are particularly useful for deriving our inequalities on measures of volatility. We have already used (4) to derive the limit (1) on the standard deviation of p given the standard deviation of p^* , and we will use (5) to derive a limit on the standard deviation of δp given the standard deviation of d .

One issue that relates to the derivation of (1) can now be clarified. The inequality (1) was derived using the assumption that the

forecast error $u_t = p_t^* - p_t$ is uncorrelated with p_t . However, the forecast error u_t is not serially uncorrelated. It is uncorrelated with all information known at time t , but the lagged forecast error u_{t-1} is not known at time t since p_{t-1}^* is not discovered at time t . In fact, $u_t = \sum_{k=1}^{\infty} \bar{\gamma}^k \delta_{t+k} p_{t+k}$, as can be seen by substituting the expressions for p_t and p_t^* from (3) and (4) into $u_t = p_t^* - p_t$, and rearranging. Since the series $\delta_t p_t$ is serially uncorrelated, u_t has first-order autoregressive serial correlation.⁸ For this reason, it is inappropriate to test the model by regressing $p_t^* - p_t$ on variables known at time t and using the ordinary t -statistics of the coefficients of these variables. However, a generalized least squares transformation of the variables would yield an appropriate regression test. We might thus regress the transformed variable $u_t - \bar{\gamma} u_{t+1}$ on variables known at time t . Since $u_t - \bar{\gamma} u_{t+1} = \bar{\gamma} \delta_{t+1} p_{t+1}$, this amounts to testing whether the innovation in price can be forecasted. I will perform and discuss such regression tests in Section V below.

To find a limit on the standard deviation of δp for a given standard deviation of d_t , first note that d_t equals its unconditional expectation plus the sum of its innovations:

$$(6) \quad d_t = E(d) + \sum_{k=0}^{\infty} \delta_{t-k} d_t$$

If we regard $E(d)$ as $E_{-\infty}(d_t)$, then this expression is just a tautology. It tells us, though, that d_t , $t=0, 1, 2, \dots$ are just different linear combinations of the same innovations in dividends that enter into the linear combination in (5) which determine $\delta_t p_t$, $t=0, 1, 2, \dots$. We can thus ask how large $\text{var}(\delta p)$ might be for given $\text{var}(d)$. Since innovations are serially uncorrelated, we know from (6) that the variance of the sum is

⁸It follows that $\text{var}(u) = \text{var}(\delta p) / (1 - \bar{\gamma}^2)$ as LeRoy and Porter noted. They base their volatility tests on our inequality (1) (which they call theorem 2) and an equality restriction $\sigma^2(p) + \sigma^2(\delta p) / (1 - \bar{\gamma}^2) = \sigma^2(p^*)$ (their theorem 3). They found that, with postwar Standard and Poor earnings data, both relations were violated by sample statistics.

the sum of the variances:

$$(7) \quad \text{var}(d) = \sum_{k=0}^{\infty} \text{var}(\delta d_k) = \sum_{k=0}^{\infty} \sigma_k^2$$

Our assumption of stationarity for d_t implies that $\text{var}(\delta_{t-k} d_t) \equiv \text{var}(\delta d_k) \equiv \sigma_k^2$ is independent of t .

In expression (5) we have no information that the variance of the sum is the sum of the variances since all the innovations are time t innovations, which may be correlated. In fact, for given $\sigma_0^2, \sigma_1^2, \dots$, the maximum variance of the sum in (5) occurs when the elements in the sum are perfectly positively correlated. This means then that so long as $\text{var}(\delta d) \neq 0$, $\delta_t d_{t+k} = a_k \delta_t d_t$, where $a_k = \sigma_k / \sigma_0$. Substituting this into (6) implies

$$(8) \quad \hat{d}_t = \sum_{k=0}^{\infty} a_k \epsilon_{t-k}$$

where a hat denotes a variable minus its mean: $\hat{d}_t \equiv d_t - E(d)$ and $\epsilon_t \equiv \delta_t d_t$. Thus, if $\text{var}(\delta p)$ is to be maximized for given $\sigma_0^2, \sigma_1^2, \dots$, the dividend process must be a moving average process in terms of its own innovations.⁹ I have thus shown, rather than assumed, that if the variance of δp is to be maximized, the forecast of d_{t+k} will have the usual ARIMA form as in the forecast popularized by Box and Jenkins.

We can now find the maximum possible variance for δp for given variance of d . Since the innovations in (5) are perfectly positively correlated, $\text{var}(\delta p) = (\sum_{k=0}^{\infty} \bar{\gamma}^{k+1} \sigma_k)^2$. To maximize this subject to the constraint $\text{var}(d) = \sum_{k=0}^{\infty} \sigma_k^2$ with respect to $\sigma_0, \sigma_1, \dots$, one may set up the Lagrangean:

$$(9) \quad L = \left(\sum_{k=0}^{\infty} \bar{\gamma}^{k+1} \sigma_k \right)^2 + \nu \left(\text{var}(d) - \sum_{k=0}^{\infty} \sigma_k^2 \right)$$

⁹Of course, all indeterministic stationary processes can be given linear moving average representations, as Hermann Wold showed. However, it does not follow that the process can be given a moving average representation in terms of its own innovations. The true process may be generated nonlinearly or other information besides its own lagged values may be used in forecasting. These will generally result in a less than perfect correlation of the terms in (5).

where ν is the Lagrangean multiplier. The first-order conditions for $\sigma_j, j=0, \dots, \infty$ are

$$(10) \quad \frac{\partial L}{\partial \sigma_j} = 2 \left(\sum_{k=0}^{\infty} \bar{\gamma}^{k+1} \sigma_k \right) \bar{\gamma}^{j+1} - 2\nu \sigma_j = 0$$

which in turn means that σ_j is proportional to $\bar{\gamma}^j$. The second-order conditions for a maximum are satisfied, and the maximum can be viewed as a tangency of an isoquant for $\text{var}(\delta p)$, which is a hyperplane in $\sigma_0, \sigma_1, \sigma_2, \dots$ space, with the hypersphere represented by the constraint. At the maximum $\sigma_k^2 = (1 - \bar{\gamma}^2) \text{var}(d) \bar{\gamma}^{2k}$ and $\text{var}(\delta p) = \bar{\gamma}^2 \text{var}(d) / (1 - \bar{\gamma}^2)$ and so, converting to standard deviations for ease of interpretation, we have

$$(11) \quad \sigma(\delta p) \leq \sigma(d) / \sqrt{\bar{r}_2}$$

$$\text{where } \bar{r}_2 = (1 + \bar{r})^2 - 1$$

Here, \bar{r}_2 is the two-period interest rate, which is roughly twice the one-period rate. The maximum occurs, then, when d_t is a first-order autoregressive process, $\hat{d}_t = \bar{\gamma} \hat{d}_{t-1} + \epsilon_t$, and $E_t \hat{d}_{t+k} = \bar{\gamma}^k \hat{d}_t$, where $\hat{d} \equiv d - E(d)$ as before.

The variance of the innovation in price is thus maximized when information about dividends is revealed in a smooth fashion so that the standard deviation of the new information at time t about a future dividend d_{t+k} is proportional to its weight in the present value formula in the model (5). In contrast, suppose all dividends somehow became known years before they were paid. Then the innovations in dividends would be so heavily discounted in (5) that they would contribute little to the standard deviation of the innovation in price. Alternatively, suppose nothing were known about dividends until the year they are paid. Here, although the innovation would not be heavily discounted in (5), the impact of the innovation would be confined to only one term in (5), and the standard deviation in the innovation in price would be limited to the standard deviation in the single dividend.

Other inequalities analogous to (11) can also be derived in the same way. For exam-

ple, we can put an upper bound to the standard deviation of the change in price (rather than the innovation in price) for given standard deviation in dividend. The only difference induced in the above procedure is that Δp_t is a different linear combination of innovations in dividends. Using the fact that $\Delta p_t = \delta_t p_t + \bar{r} p_{t-1} - d_{t-1}$ we find

$$(12) \quad \Delta p_t = \sum_{k=0}^{\infty} \bar{\gamma}^{k+1} \delta_t d_{t+k} \\ + \bar{r} \sum_{j=1}^{\infty} \delta_{t-j} \sum_{k=0}^{\infty} \bar{\gamma}^{k+1} d_{t+k-1} - \sum_{j=1}^{\infty} \delta_{t-j} d_{t-1}$$

As above, the maximization of the variance of Δp for given variance of d requires that the time t innovations in d be perfectly correlated (innovations at different times are necessarily uncorrelated) so that again the dividend process must be forecasted as an ARIMA process. However, the parameters of the ARIMA process for d which maximize the variance of Δp will be different. One finds, after maximizing the Lagrangean expression (analogous to (9)) an inequality slightly different from (11),

$$(13) \quad \sigma(\Delta p) \leq \sigma(d) / \sqrt{2\bar{r}}$$

The upper bound is attained if the optimal dividend forecast is first-order autoregressive, but with an autoregressive coefficient slightly different from that which induced the upper bound to (11). The upper bound to (13) is attained if $\hat{d}_t = (1 - \bar{r}) \hat{d}_{t-1} + \varepsilon_t$ and $E_t d_{t+k} = (1 - \bar{r})^k \hat{d}_t$, where, as before, $\hat{d}_t \equiv d_t - E(d)$.

II. High Kurtosis and Infrequent Important Breaks in Information

It has been repeatedly noted that stock price change distributions show high kurtosis or "fat tails." This means that, if one looks at a time-series of observations on δp or Δp , one sees long stretches of time when their (absolute) values are all rather small and then an occasional extremely large (absolute)

value. This phenomenon is commonly attributed to a tendency for new information to come in big lumps infrequently. There seems to be a common presumption that this information lumping might cause stock price changes to have high or infinite variance, which would seem to contradict the conclusion in the preceding section that the variance of price is limited and is maximized if forecasts have a simple autoregressive structure.

High sample kurtosis does not indicate infinite variance if we do not assume, as did Fama (1965) and others, that price changes are drawn from the stable Paretian class of distributions.¹⁰ The model does not suggest that price changes have a distribution in this class. The model instead suggests that the existence of moments for the price series is implied by the existence of moments for the dividends series.

As long as d is jointly stationary with information and has a finite variance, then p , p^* , δp , and Δp will be stationary and have a finite variance.¹¹ If d is normally distributed, however, it does not follow that the price variables will be normally distributed. In fact, they may yet show high kurtosis.

To see this possibility, suppose the dividends are serially independent and identically normally distributed. The kurtosis of the price series is defined by $K = E(\hat{p})^4 / (E(\hat{p})^2)^2$, where $p \equiv \hat{p} - E(p)$. Suppose, as an example, that with a probability of $1/n$

¹⁰The empirical fact about the unconditional distribution of stock price changes is not that they have infinite variance (which can never be demonstrated with any finite sample), but that they have high kurtosis in the sample.

¹¹With any stationary process X_t , the existence of a finite $var(X_t)$ implies, by Schwartz's inequality, a finite value of $cov(X_t, X_{t+k})$ for any k , and hence the entire autocovariance function of X_t , and the spectrum, exists. Moreover, the variance of $E_t(X_t)$ must also be finite, since the variance of X equals the variance of $E_t(X_t)$ plus the variance of the forecast error. While we may regard real dividends as having finite variance, innovations in dividends may show high kurtosis. The residuals in a second-order autoregression for d_t have a studentized range of 6.29 for the Standard and Poor series and 5.37 for the Dow series. According to the David-Hartley-Pearson test, normality can be rejected at the 5 percent level (but not at the 1 percent level) with a one-tailed test for both data sets.

the public is told d_t at the beginning of time t , but with probability $(n-1)/n$ has no information about current or future dividends.¹² In time periods when they are told d_t , \hat{p}_t equals $\bar{\gamma}d_t$, otherwise $\hat{p}_t=0$. Then $E(\hat{p}_t^4)=E((\bar{\gamma}d_t)^4)/n$ and $E(\hat{p}_t^2)=E((\bar{\gamma}d_t)^2)/n$ so that kurtosis equals $nE(\bar{\gamma}d_t^4)/E((\bar{\gamma}d_t)^2)$ which equals n times the kurtosis of the normal distribution. Hence, by choosing n high enough one can achieve an arbitrarily high kurtosis, and yet the variance of price will always exist. Moreover, the distribution of \hat{p}_t conditional on the information that the dividend has been revealed is also normal, in spite of high kurtosis of the unconditional distribution.

If information is revealed in big lumps occasionally (so as to induce high kurtosis as suggested in the above example) $\text{var}(\delta p)$ or $\text{var}(\Delta p)$ are not especially large. The variance loses more from the long interval of time when information is not revealed than it gains from the infrequent events when it is. The highest possible variance for given variance of d indeed comes when information is revealed smoothly as noted in the previous section. In the above example, where information about dividends is revealed one time in n , $\sigma(\delta p)=\bar{\gamma}n^{1/2}\sigma(d)$ and $\sigma(\Delta p)=\bar{\gamma}(2/n)^{1/2}\sigma(d)$. The values of $\sigma(\delta p)$ and $\sigma(\Delta p)$ implied by this example are for all n strictly below the upper bounds of the inequalities (11) and (13).¹³

III. Dividends or Earnings?

It has been argued that the model (2) does not capture what is generally meant by efficient markets, and that the model should be replaced by a model which makes price the present value of expected earnings rather than dividends. In the model (2) earnings

¹²For simplicity, in this example, the assumption elsewhere in this article that d_t is always known at time t has been dropped. It follows that in this example $\delta_t p_t \neq \Delta p_t + d_{t-1} - r p_{t-1}$ but instead $\delta_t p_t = p_t$.

¹³For another illustrative example, consider $\hat{d}_t=\bar{\gamma}\hat{d}_{t-1}+\epsilon_t$ as with the upper bound for the inequality (11) but where the dividends are announced for the next n years every $1/n$ years. Here, even though \hat{d}_t has the autoregressive structure, ϵ_t is not the innovation in d_t . As n goes to infinity, $\sigma(\delta p)$ approaches zero.

may be relevant to the pricing of shares but only insofar as earnings are indicators of future dividends. Earnings are thus no different from any other economic variable which may indicate future dividends. The model (2) is consistent with the usual notion in finance that individuals are concerned with returns, that is, capital gains plus dividends. The model implies that expected total returns are constant and that the capital gains component of returns is just a reflection of information about future dividends. Earnings, in contrast, are statistics conceived by accountants which are supposed to provide an indicator of how well a company is doing, and there is a great deal of latitude for the definition of earnings, as the recent literature on inflation accounting will attest.

There is no reason why price per share ought to be the present value of expected earnings per share if some earnings are retained. In fact, as Merton Miller and Franco Modigliani argued, such a present value formula would entail a fundamental sort of double counting. It is incorrect to include in the present value formula both earnings at time t and the later earnings that accrue when time t earnings are reinvested.¹⁴ Miller and Modigliani showed a formula by which price might be regarded as the present value of earnings corrected for investments, but that formula can be shown, using an accounting identity to be identical to (2).

Some people seem to feel that one cannot claim price as present value of expected dividends since firms routinely pay out only a fraction of earnings and also attempt somewhat to stabilize dividends. They are right in the case where firms paid out no dividends, for then the price p_t would have to grow at the discount rate \bar{r} , and the model (2) would not be the solution to the difference equation implied by the condition $E_t(H_t)=r$. On the other hand, if firms pay out a fraction of dividends or smooth short-run fluctuations in dividends, then the price of the firm will grow at a rate less than the

¹⁴LeRoy and Porter do assume price as present value of earnings but employ a correction to the price and earnings series which is, under additional theoretical assumptions not employed by Miller and Modigliani, a correction for the double counting.

discount rate and (2) is the solution to the difference equation.¹⁵ With our Standard and Poor data, the growth rate of real price is only about 1.5 percent, while the discount rate is about $4.8\% + 1.5\% = 6.3\%$. At these rates, the value of the firm a few decades hence is so heavily discounted relative to its size that it contributes very little to the value of the stock today; by far the most of the value comes from the intervening dividends. Hence (2) and the implied p^* ought to be useful characterizations of the value of the firm.

The crucial thing to recognize in this context is that once we know the terminal price and intervening dividends, we have specified all that investors care about. It would not make sense to define an *ex post* rational price from a terminal condition on price, using the same formula with earnings in place of dividends.

IV. Time-Varying Real Discount Rates

If we modify the model (2) to allow real discount rates to vary without restriction through time, then the model becomes untestable. We do not observe real discount rates directly. Regardless of the behavior of P_t and D_t , there will always be a discount rate series which makes (2) hold identically. We might ask, though, whether the movements in the real discount rate that would be required aren't larger than we might have expected. Or is it possible that small movements in the current one-period discount rate coupled with new information about such movements in future discount rates could account for high stock price volatility?¹⁶

¹⁵To understand this point, it helps to consider a traditional continuous time growth model, so instead of (2) we have $P_0 = \int_0^\infty D_t e^{-rt} dt$. In such a model, a firm has a constant earnings stream I . If it pays out all earnings, then $D=I$ and $P_0 = \int_0^\infty I e^{-rt} dt = I/r$. If it pays out only s of its earnings, then the firm grows at rate $(1-s)r$, $D_t = sIe^{(1-s)r t}$ which is less than I at $t=0$, but higher than I later on. Then $P_0 = \int_0^\infty sIe^{(1-s)r t} e^{-rt} dt = \int_0^\infty sIe^{-srt} dt = sI/(rs)$. If $s \neq 0$ (so that we're not dividing by zero) $P_0 = I/r$.

¹⁶James Pesando has discussed the analogous question: how large must the variance in liquidity premia be in order to justify the volatility of long-term interest rates?

The natural extension of (2) to the case of time varying real discount rates is

$$(14) \quad P_t = E_t \left(\sum_{k=0}^{\infty} D_{t+k} \prod_{j=0}^k \frac{1}{1+r_{t+j}} \right)$$

which has the property that $E_t((1+H_t)/(1+r_t)) = 1$. If we set $1+r_t = (\partial U/\partial C_t)/(\partial U/\partial C_{t+1})$, i.e., to the marginal rate of substitution between present and future consumption where U is the additively separable utility of consumption, then this property is the first-order condition for a maximum of expected utility subject to a stock market budget constraint, and equation (14) is consistent with such expected utility maximization at all times. Note that while r_t is a sort of *ex post* real interest rate not necessarily known until time $t+1$, only the conditional distribution at time t or earlier influences price in the formula (14).

As before, we can rewrite the model in terms of detrended series:

$$(15) \quad p_t = E_t(p_t^*)$$

$$\text{where } p_t^* \equiv \sum_{k=0}^{\infty} d_{t+k} \prod_{j=0}^k \frac{1}{1+\bar{r}_{t+j}}$$

$$1+\bar{r}_{t+j} \equiv (1+r_t)/\lambda$$

This model then implies that $\sigma(p_t) \leq \sigma(p_t^*)$ as before. Since the model is nonlinear, however, it does not allow us to derive inequalities like (11) or (13). On the other hand, if movements in real interest rates are not too large, then we can use the linearization of p_t^* (i.e., Taylor expansion truncated after the linear term) around $d=E(d)$ and $\bar{r}=E(\bar{r})$; i.e.,

$$(16) \quad \hat{p}_t^* \equiv \sum_{k=0}^{\infty} \bar{y}^{k+1} \hat{d}_{t+k} - \frac{E(d)}{E(\bar{r})} \sum_{k=0}^{\infty} \bar{y}^{k+1} \hat{\bar{r}}_{t+k}$$

where $\bar{y} = 1/(1+E(\bar{r}))$, and a hat over a variable denotes the variable minus its mean. The first term in the above expression is just the expression for p_t^* in (4) (demeaned). The second term represents the effect on p_t^* of

movements in real discount rates. This second term is identical to the expression for p_t^* in (4) except that d_{t+k} is replaced by \hat{r}_{t+k} and the expression is premultiplied by $-E(d)/E(\bar{r})$.

It is possible to offer a simple intuitive interpretation for this linearization. First note that the derivative of $1/(1+\bar{r}_{t+k})$, with respect to \bar{r} evaluated at $E(\bar{r})$ is $-\bar{\gamma}^2$. Thus, a one percentage point increase in \bar{r}_{t+k} causes $1/(1+\bar{r}_{t+k})$ to drop by $\bar{\gamma}^2$ times 1 percent, or slightly less than 1 percent. Note that all terms in (15) dated $t+k$ or higher are premultiplied by $1/(1+\bar{r}_{t+k})$. Thus, if \bar{r}_{t+k} is increased by one percentage point, all else constant, then all of these terms will be reduced by about $\bar{\gamma}^2$ times 1 percent. We can approximate the sum of all these terms as $\bar{\gamma}^{k-1}E(d)/E(\bar{r})$, where $E(d)/E(\bar{r})$ is the value at the beginning of time $t+k$ of a constant dividend stream $E(d)$ discounted by $E(\bar{r})$, and $\bar{\gamma}^{k-1}$ discounts it to the present. So, we see that a one percentage point increase in \bar{r}_{t+k} , all else constant, decreases p_t^* by about $\bar{\gamma}^{k+1}E(d)/E(\bar{r})$, which corresponds to the k th term in expression (16). There are two sources of inaccuracy with this linearization. First, the present value of all future dividends starting with time $t+k$ is not exactly $\bar{\gamma}^{k-1}E(d)/E(\bar{r})$. Second, increasing \bar{r}_{t+k} by one percentage point does not cause $1/(1+\bar{r}_{t+k})$ to fall by exactly $\bar{\gamma}^2$ times 1 percent. To some extent, however, these errors in the effects on p_t^* of $\bar{r}_t, \bar{r}_{t+1}, \bar{r}_{t+2}, \dots$ should average out, and one can use (16) to get an idea of the effects of changes in discount rates.

To give an impression as to the accuracy of the linearization (16), I computed p_t^* for data set 2 in two ways: first using (15) and then using (16), with the same terminal condition p_{1979}^* . In place of the unobserved \bar{r}_t series, I used the actual four-six-month prime commercial paper rate plus a constant to give it the mean \bar{r} of Table 2. The commercial paper rate is a *nominal* interest rate, and thus one would expect its fluctuations represent changes in inflationary expectations as well as real interest rate movements. I chose it nonetheless, rather arbitrarily, as a series which shows much more fluctuation than one would normally expect to see in an

TABLE 2—SAMPLE STATISTICS FOR PRICE AND DIVIDEND SERIES

	Sample Period:	Data Set 1: Standard and Poor's	Data Set 2: Modified Dow Industrial
		1871–1979	1928–1979
1)	$E(p)$	145.5	982.6
	$E(d)$	6.989	44.76
2)	\bar{r}	.0480	0.456
	\bar{r}_2	.0984	.0932
3)	$b = \ln \lambda$.0148	.0188
	$\hat{\sigma}(b)$	(.0011)	(1.0035)
4)	$\text{cor}(p, p^*)$.3918	.1626
	$\sigma(d)$	1.481	9.828
Elements of Inequalities:			
Inequality (1)			
5)	$\sigma(p)$	50.12	355.9
6)	$\sigma(p^*)$	8.968	26.80
Inequality (11)			
7)	$\sigma(\Delta p + d_{-1} - \bar{r}p_{-1})$	25.57	242.1
	$\min(\sigma)$	23.01	209.0
8)	$\sigma(d)/\sqrt{\bar{r}_2}$	4.721	32.20
Inequality (13)			
9)	$\sigma(\Delta p)$	25.24	239.5
	$\min(\sigma)$	22.71	206.4
10)	$\sigma(d)/\sqrt{2\bar{r}}$	4.777	32.56

Note: In this table, E denotes sample mean, σ denotes standard deviation and $\hat{\sigma}$ denotes standard error. $\min(\sigma)$ is the lower bound on σ computed as a one-sided χ^2 95 percent confidence interval. The symbols p , d , \bar{r} , \bar{r}_2 , b , and p^* are defined in the text. Data sets are described in the Appendix. Inequality (1) in the text asserts that the standard deviation in row 5 should be less than or equal to that in row 6, inequality (11) that σ in row 7 should be less than or equal to that in row 8, and inequality (13) that σ in row 9 should be less than that in row 10.

expected *real* rate. The commercial paper rate ranges, in this sample, from 0.53 to 9.87 percent. It stayed below 1 percent for over a decade (1935–46) and, at the end of the sample, stayed generally well above 5 percent for over a decade. In spite of this erratic behavior, the correlation coefficient between p^* computed from (15) and p^* computed from (16) was .996, and $\sigma(p_t^*)$ was 250.5 and 268.0 by (15) and (16), respectively. Thus the linearization (16) can be quite accurate. Note also that while these large movements in \bar{r}_t cause p_t^* to move much more than was observed in Figure 2, $\sigma(p^*)$ is still less than half of $\sigma(p)$. This suggests that the variability \bar{r}_t that is needed to save the efficient

markets model is much larger yet, as we shall see.

To put a formal lower bound on $\sigma(\bar{r})$ given the variability of Δp , note that (16) makes \hat{p}_t^* the present value of z_t, z_{t+1}, \dots where $z_t \equiv \hat{d}_t - \hat{r}_t E(d)/E(\bar{r})$. We thus know from (13) that $2E(\bar{r})\text{var}(\Delta p) \leq \text{var}(z)$. Moreover, from the definition of z we know that $\text{var}(z) \leq \text{var}(d) + 2\sigma(d)\sigma(\bar{r})E(d)/E(\bar{r}) + \text{var}(\bar{r})E(d)^2/E(\bar{r})^2$ where the equality holds if d_t and \bar{r}_t are perfectly negatively correlated. Combining these two inequalities and solving for $\sigma(\bar{r})$ one finds

(17)

$$\sigma(\bar{r}) \geq (\sqrt{2E(\bar{r})\text{var}(\Delta p)} - \sigma(d))E(\bar{r})/E(d)$$

This inequality puts a lower bound on $\sigma(\bar{r})$ proportional to the discrepancy between the left-hand side and right-hand side of the inequality (13).¹⁷ It will be used to examine the data in the next section.

V. Empirical Evidence

The elements of the inequalities (1), (11), and (13) are displayed for the two data sets (described in the Appendix) in Table 2. In both data sets, the long-run exponential growth path was estimated by regressing $\ln(P_t)$ on a constant and time. Then λ in (3) was set equal to e^b where b is the coefficient of time (Table 2). The discount rate \bar{r} used to compute p^* from (4) is estimated as the average d divided by the average p .¹⁸ The terminal value of p^* is taken as average p .

With data set 1, the nominal price and dividend series are the real Standard and Poor's Composite Stock Price Index and the associated dividend series. The earlier observations for this series are due to Alfred

¹⁷In deriving the inequality (13) it was assumed that d_t was known at time t , so by analogy this inequality would be based on the assumption that r_t is known at time t . However, without this assumption the same inequality could be derived anyway. The maximum contribution of \bar{r}_t to the variance of ΔP occurs when \bar{r}_t is known at time t .

¹⁸This is not equivalent to the average dividend price ratio, which was slightly higher (.0514 for data set 1, .0484 for data set 2).

Cowles who said that the index is

intended to represent, ignoring the elements of brokerage charges and taxes, what would have happened to an investor's funds if he had bought, at the beginning of 1871, all stocks quoted on the New York Stock Exchange, allocating his purchases among the individual stocks in proportion to their total monetary value and each month up to 1937 had by the same criterion redistributed his holdings among all quoted stocks.

[p. 2]

In updating his series, Standard and Poor later restricted the sample to 500 stocks, but the series continues to be value weighted. The advantage to this series is its comprehensiveness. The disadvantage is that the dividends accruing to the portfolio at one point of time may not correspond to the dividends forecasted by holders of the Standard and Poor's portfolio at an earlier time, due to the change in weighting of the stocks. There is no way to correct this disadvantage without losing comprehensiveness. The original portfolio of 1871 is bound to become a relatively smaller and smaller sample of U.S. common stocks as time goes on.

With data set 2, the nominal series are a modified Dow Jones Industrial Average and associated dividend series. With this data set, the advantages and disadvantages of data set 1 are reversed. My modifications in the Dow Jones Industrial Average assure that this series reflects the performance of a single unchanging portfolio. The disadvantage is that the performance of only 30 stocks is recorded.

Table 2 reveals that all inequalities are dramatically violated by the sample statistics for both data sets. The left-hand side of the inequality is always at least five times as great as the right-hand side, and as much as thirteen times as great.

The violation of the inequalities implies that "innovations" in price as we measure them can be forecasted. In fact, if we regress $\delta_{t+1}p_{t+1}$ onto (a constant and) p_t , we get significant results: a coefficient of p_t of $-.1521$ ($t = -3.218$, $R^2 = .0890$) for data set 1 and a coefficient of $-.2421$ ($t = -2.631$, $R^2 = .1238$) for data set 2. These results are

not due to the representation of the data as a proportion of the long-run growth path. In fact, if the holding period return H_t is regressed on a constant and the dividend price ratio D_t/P_t , we get results that are only slightly less significant: a coefficient of 3.533 ($t=2.672$, $R^2=.0631$) for data set 1 and a coefficient of 4.491 ($t=1.795$, $R^2=.0617$) for data set 2.

These regression tests, while technically valid, may not be as generally useful for appraising the validity of the model as are the simple volatility comparisons. First, as noted above, the regression tests are not insensitive to data misalignment. Such low R^2 might be the result of dividend or commodity price index data errors. Second, although the model is rejected in these very long samples, the tests may not be powerful if we confined ourselves to shorter samples, for which the data are more accurate, as do most researchers in finance, while volatility comparisons may be much more revealing. To see this, consider a stylized world in which (for the sake of argument) the dividend series d_t is absolutely constant while the price series behaves as in our data set. Since the actual dividend series is fairly smooth, our stylized world is not too remote from our own. If dividends d_t are absolutely constant, however, it should be obvious to the most casual and unsophisticated observer by volatility arguments like those made here that the efficient markets model must be wrong. Price movements cannot reflect new information about dividends if dividends never change. Yet regressions like those run above will have limited power to reject the model. If the alternative hypothesis is, say, that $\hat{p}_t = \rho \hat{p}_{t-1} + \epsilon_t$, where ρ is close to but less than one, then the power of the test in short samples will be very low. In this stylized world we are testing for the stationarity of the p_t series, for which, as we know, power is low in short samples.¹⁹ For example, if post-

war data from, say, 1950–65 were chosen (a period often used in recent financial markets studies) when the stock market was drifting up, then clearly the regression tests will not reject. Even in periods showing a reversal of upward drift the rejection may not be significant.

Using inequality (17), we can compute how big the standard deviation of real discount rates would have to be to possibly account for the discrepancy $\sigma(\Delta p) - \sigma(d)/(2\bar{r})^{1/2}$ between Table 2 results (rows 9 and 10) and the inequality (13). Assuming Table 2 \bar{r} (row 2) equals $E(\bar{r})$ and that sample variances equal population variances, we find that the standard deviation of \bar{r}_t would have to be at least 4.36 percentage points for data set 1 and 7.36 percentage points for data set 2. These are very large numbers. If we take, as a normal range for \bar{r}_t implied by these figures, ± 2 standard deviation range around the real interest rate \bar{r} given in Table 2, then the real interest rate \bar{r}_t would have to range from –3.91 to 13.52 percent for data set 1 and –8.16 to 17.27 percent for data set 2! And these ranges reflect lowest possible standard deviations which are consistent with the model only if the real rate has the first-order autoregressive structure and perfect negative correlation with dividends!

These estimated standard deviations of *ex ante* real interest rates are roughly consistent with the results of the simple regressions noted above. In a regression of H_t on D_t/P_t and a constant, the standard deviation of the fitted value of H_t is 4.42 and 5.71 percent for data sets 1 and 2, respectively. These large standard deviations are consistent with the low R^2 because the standard deviation of H_t is so much higher (17.60 and 23.00 percent, respectively). The regressions of $\delta_t p_t$ on p_t suggest higher standard deviations of expected real interest rates. The standard deviation of the fitted value divided by the average detrended price is 5.24 and 8.67 percent for data sets 1 and 2, respectively.

¹⁹If dividends are constant (let us say $d_t=0$) then a test of the model by a regression of $\delta_{t+1} p_{t+1}$ on p_t amounts to a regression of p_{t+1} on p_t with the null hypothesis that the coefficient of p_t is $(1+\bar{r})$. This appears to be an explosive model for which t -statistics are not valid yet our true model, which in effect assumes $\sigma(d)=0$, is nonexplosive.

VI. Summary and Conclusions

We have seen that measures of stock price volatility over the past century appear to be far too high—five to thirteen times too

high—to be attributed to new information about future real dividends if uncertainty about future dividends is measured by the sample standard deviations of real dividends around their long-run exponential growth path. The lower bound of a 95 percent one-sided χ^2 confidence interval for the standard deviation of annual changes in real stock prices is over five times higher than the upper bound allowed by our measure of the observed variability of real dividends. The failure of the efficient markets model is thus so dramatic that it would seem impossible to attribute the failure to such things as data errors, price index problems, or changes in tax laws.

One way of saving the general notion of efficient markets would be to attribute the movements in stock prices to changes in expected real interest rates. Since expected real interest rates are not directly observed, such a theory can not be evaluated statistically unless some other indicator of real rates is found. I have shown, however, that the movements in expected real interest rates that would justify the variability in stock prices are very large—much larger than the movements in nominal interest rates over the sample period.

Another way of saving the general notion of efficient markets is to say that our measure of the uncertainty regarding future dividends—the sample standard deviation of the movements of real dividends around their long-run exponential growth path—understates the true uncertainty about future dividends. Perhaps the market was rightfully fearful of much larger movements than actually materialized. One is led to doubt this, if after a century of observations nothing happened which could remotely justify the stock price movements. The movements in real dividends the market feared must have been many times larger than those observed in the Great Depression of the 1930's, as was noted above. Since the market did not know in advance with certainty the growth path and distribution of dividends that was ultimately observed, however, one cannot be sure that they were wrong to consider possible major events which did not occur. Such an explanation of the volatility of stock prices, however,

is "academic," in that it relies fundamentally on unobservables and cannot be evaluated statistically.

APPENDIX

A. Data Set 1: Standard and Poor Series

Annual 1871–1979. The price series P_t is Standard and Poor's Monthly Composite Stock Price index for January divided by the Bureau of Labor Statistics wholesale price index (January *WPI* starting in 1900, annual average *WPI* before 1900 scaled to 1.00 in the base year 1979). Standard and Poor's Monthly Composite Stock Price index is a continuation of the Cowles Commission Common Stock index developed by Alfred Cowles and Associates and currently is based on 500 stocks.

The Dividend Series D_t is total dividends for the calendar year accruing to the portfolio represented by the stocks in the index divided by the average wholesale price index for the year (annual average *WPI* scaled to 1.00 in the base year 1979). Starting in 1926 these total dividends are the series "Dividends per share... 12 months moving total adjusted to index" from Standard and Poor's statistical service. For 1871 to 1925, total dividends are Cowles series Da-1 multiplied by .1264 to correct for change in base year.

B. Data Set 2: Modified Dow Jones Industrial Average

Annual 1928–1979. Here P_t and D_t refer to real price and dividends of the portfolio of 30 stocks comprising the sample for the Dow Jones Industrial Average when it was created in 1928. Dow Jones averages before 1928 exist, but the 30 industrials series was begun in that year. The published Dow Jones Industrial Average, however, is not ideal in that stocks are dropped and replaced and in that the weighting given an individual stock is affected by splits. Of the original 30 stocks, only 17 were still included in the Dow Jones Industrial Average at the end of our sample. The published Dow Jones Industrial Average is the simple sum of the price per share of the 30 companies divided by a divisor which

changes through time. Thus, if a stock splits two for one, then Dow Jones continues to include only one share but changes the divisor to prevent a sudden drop in the Dow Jones average.

To produce the series used in this paper, the *Capital Changes Reporter* was used to trace changes in the companies from 1928 to 1979. Of the original 30 companies of the Dow Jones Industrial Average, at the end of our sample (1979), 9 had the identical names, 12 had changed only their names, and 9 had been acquired, merged or consolidated. For these latter 9, the price and dividend series are continued as the price and dividend of the shares exchanged by the acquiring corporation. In only one case was a cash payment, along with shares of the acquiring corporation, exchanged for the shares of the acquired corporation. In this case, the price and dividend series were continued as the price and dividend of the shares exchanged by the acquiring corporation. In four cases, preferred shares of the acquiring corporation were among shares exchanged. Common shares of equal value were substituted for these in our series. The number of shares of each firm included in the total is determined by the splits, and effective splits effected by stock dividends and merger. The price series is the value of all these shares on the last trading day of the preceding year, as shown on the Wharton School's Rodney White Center Common Stock tape. The dividend series is the total for the year of dividends and the cash value of other distributions for all these shares. The price and dividend series were deflated using the same wholesale price indexes as in data set 1.

REFERENCES

- C. Amsler, "An American Consol: A Reexamination of the Expectations Theory of the Term Structure of Interest Rates," unpublished manuscript, Michigan State Univ. 1980.
- S. Basu, "The Investment Performance of Common Stocks in Relation to their Price-Earnings Ratios: A Test of the Efficient Markets Hypothesis," *J. Finance*, June 1977, 32, 663-82.
- G. E. P. Box and G. M. Jenkins, *Time Series Analysis for Forecasting and Control*, San Francisco: Holden-Day 1970.
- W. C. Brainard, J. B. Shoven, and L. Weiss, "The Financial Valuation of the Return to Capital," *Brookings Papers*, Washington 1980, 2, 453-502.
- Paul H. Cootner, *The Random Character of Stock Market Prices*, Cambridge: MIT Press 1964.
- Alfred Cowles and Associates, *Common Stock Indexes, 1871-1937*, Cowles Commission for Research in Economics, Monograph No. 3, Bloomington: Principia Press 1938.
- E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *J. Finance*, May 1970, 25, 383-420.
- _____, "The Behavior of Stock Market Prices," *J. Bus., Univ. Chicago*, Jan. 1965, 38, 34-105.
- C. W. J. Granger, "Some Consequences of the Valuation Model when Expectations are Taken to be Optimum Forecasts," *J. Finance*, Mar. 1975, 30, 135-45.
- M. C. Jensen et al., "Symposium on Some Anomalous Evidence Regarding Market Efficiency," *J. Financ. Econ.*, June/Sept. 1978, 6, 93-330.
- S. LeRoy and R. Porter, "The Present Value Relation: Tests Based on Implied Variance Bounds," *Econometrica*, forthcoming.
- M. H. Miller and F. Modigliani, "Dividend Policy, Growth and the Valuation of Shares," *J. Bus., Univ. Chicago*, Oct. 1961, 34, 411-33.
- F. Modigliani and R. Cohn, "Inflation, Rational Valuation and the Market," *Financ. Anal. J.*, Mar./Apr. 1979, 35, 24-44.
- J. Pesando, "Time Varying Term Premiums and the Volatility of Long-Term Interest Rates," unpublished paper, Univ. Toronto, July 1979.
- P. A. Samuelson, "Proof that Properly Discounted Present Values of Assets Vibrate Randomly," in Hiroaki Nagatani and Kate Crowley, eds., *Collected Scientific Papers of Paul A. Samuelson*, Vol. IV, Cambridge: MIT Press 1977.
- R. J. Shiller, "The Volatility of Long-Term Interest Rates and Expectations Models of the Term Structure," *J. Polit. Econ.*, Dec. 1979, 87, 1190-219.

- ____ and J. J. Siegel, "The Gibson Paradox and Historical Movements in Real Interest Rates," *J. Polit. Econ.*, Oct. 1979, 85, 891-907.
- H. Wold, "On Prediction in Stationary Time Series," *Annals Math. Statist.* 1948, 19, 558-67.
- Commerce Clearing House, *Capital Changes Reporter*, New Jersey 1977.
- Dow Jones & Co., *The Dow Jones Averages 1855-1970*, New York: Dow Jones Books 1972.
- Standard and Poor's *Security Price Index Record*, New York 1978.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/243675049>

Capital Theory and Investment Behavior

Article in American Economic Review · November 1962

CITATIONS
911

READS
5,016

1 author:



Dale W. Jorgenson
Harvard University
82 PUBLICATIONS 5,564 CITATIONS

[SEE PROFILE](#)



Capital Theory and Investment Behavior

Dale W. Jorgenson

The American Economic Review, Vol. 53, No. 2, Papers and Proceedings of the Seventy-Fifth Annual Meeting of the American Economic Association. (May, 1963), pp. 247-259.

Stable URL:

<http://links.jstor.org/sici?&sici=0002-8282%28196305%2953%3A2%3C247%3ACTAIB%3E2.0.CO%3B2-J>

The American Economic Review is currently published by American Economic Association.

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/aea.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.

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact support@jstor.org.

CAPITAL THEORY AND INVESTMENT BEHAVIOR*

By DALE W. JORGENSEN
University of California, Berkeley

Introduction

There is no greater gap between economic theory and econometric practice than that which characterizes the literature on business investment in fixed capital. According to the neoclassical theory of capital, as expounded for example by Irving Fisher, a production plan for the firm is chosen so as to maximize utility over time. Under certain well-known conditions this leads to maximization of the net worth of the enterprise as the criterion for optimal capital accumulation. Capital is accumulated to provide capital services, which are inputs to the productive process. For convenience the relationship between inputs, including the input of capital services, and output is summarized in a production function. Although this theory has been known for at least fifty years, it is currently undergoing a great revival in interest. The theory appears to be gaining increasing currency and more widespread understanding.

By contrast, the econometric literature on business investment consists of *ad hoc* descriptive generalizations such as the "capacity principle," the "profit principle," and the like. Given sufficient imprecision, one can rationalize any generalization of this type by an appeal to "theory." However, even with the aid of much ambiguity, it is impossible to reconcile the theory of the econometric literature on investment with the neoclassical theory of optimal capital accumulation. The central feature of the neoclassical theory is the response of the demand for capital to changes in relative factor prices or the ratio of factor prices to the price of output. This feature is entirely absent from the econometric literature on investment.

It is difficult to reconcile the steady advance in the acceptance of the neoclassical theory of capital with the steady march of the econometric literature in a direction which appears to be diametrically opposite. It is true that there have been attempts to validate the theory. Both profits and capacity theorists have tried a rate of interest here or a price of investment goods there. By and large these efforts have been unsuccessful; the naïve positivist can only conclude, so much the worse for the theory. I believe that a case can be made that previous attempts to "test" the neoclassical theory of capital have fallen so far

* The research for this paper was completed while the author was Ford Foundation Research Professor of Economics at the University of Chicago. The research was supported by the National Science Foundation.

short of a correct formulation of this theory that the issue of the validity of the neoclassical theory remains undecided. There is not sufficient space to document this point in detail here; but I will try to illustrate what I would regard as a correct formulation of the theory in what follows.

Stated baldly, the purpose of this paper is to present a theory of investment behavior based on the neoclassical theory of optimal accumulation of capital. Of course, demand for capital is not demand for investment. The short-run determination of investment behavior depends on the time form of lagged response to changes in the demand for capital. For simplicity, the time form of lagged response will be assumed to be fixed. At the same time a more general hypothesis about the form of the lag is admitted than that customary in the literature. Finally, it will be assumed that replacement investment is proportional to capital stock. This assumption, while customary, has a deep justification which will be presented below. A number of empirical tests of the theory is presented, along with an analysis of new evidence on the time form of lagged response and changes in the long-run demand for capital resulting from changes in underlying market conditions and in the tax structure.

Summary of the Theory

Demand for capital stock is determined to maximize net worth. Net worth is defined as the integral of discounted net revenues; all prices, including the interest rate, are taken as fixed. Net revenue is defined as current revenue less expenditure on both current and capital account, including taxes. Let revenue before taxes at time t be $R(t)$, direct taxes, $D(t)$, and r the rate of interest. Net worth, say W , is

$$W = \int_0^\infty e^{-rt} [R(t) - D(t)] dt.$$

We will deduce necessary conditions for maximization of net worth for two inputs—one current and one capital—and one output. The approach is easily generalized to any number of inputs and outputs.

Let p be the price of output, s the wage rate, q the price of capital goods, Q the quantity of output, L the quantity of variable input, say labor, and I the rate of investment; net revenue is

$$R = pQ - sL - qI.$$

Let u be the rate of direct taxation, v the proportion of replacement chargeable against income for tax purposes, w the proportion of interest,

and x the proportion of capital losses chargeable against income; where K is capital stock and δ the rate of replacement, direct taxes are

$$D = u[pQ - sL - (v\delta q + wrq - x\dot{q})K]$$

Maximizing net worth subject to a standard neoclassical production function and the constraint that the rate of growth of capital stock is investment less replacement, we obtain the marginal productivity conditions

$$\frac{\partial Q}{\partial L} = \frac{s}{p},$$

$$\frac{\partial Q}{\partial K} = \frac{q \left[\frac{1 - uv}{1 - u} \delta + \frac{1 - uw}{1 - u} r - \frac{1 - ux}{1 - u} \frac{\dot{q}}{q} \right]}{p}.$$

The numerator of the second fraction is the "shadow" price or implicit rental of one unit of capital service per period of time. We will call this price the user cost of capital. We assume that all capital gains are regarded as "transitory," so that the formula for user cost, say c , reduces to

$$c = q \left[\frac{1 - uv}{1 - u} \delta + \frac{1 - uw}{1 - u} r \right].$$

Second, we assume that output and employment on the one hand and capital stock on the other are determined by a kind of iterative process. In each period, production and employment are set at the levels given by the first marginal productivity condition and the production function with capital stock fixed at its current level; demand for capital is set at the level given by the second marginal productivity condition, given output and employment. With stationary market conditions, such a process is easily seen to converge to the desired maximum of net worth. Let K^* represent the desired amount of capital stock, if the production function is Cobb-Douglas with elasticity of output with respect to capital, γ ,

$$K^* = \gamma \frac{pQ}{c}.$$

We suppose that the distribution of times to completion of new investment projects is fixed. Let the proportion of projects completed in time τ be w_τ . If investment in new projects is I_t^E and the level of starts of new projects is I_t^N , investment is a weighted average of past starts:

$$I_t^E = \sum_{\tau=0}^{\infty} w_{\tau} I_{t-\tau}^N = w(L) I_t^N,$$

where $w(L)$ is a power series in the lag operator, L . We assume that in each period new projects are initiated until the backlog of uncompleted projects is equal to the difference between desired capital stock, K_t^* , and actual capital stock, K_t :

$$I_t^N = K_t^* - [K_t + (1 - w_0) I_{t-1}^N + \dots],$$

which implies that:

$$I_t^E = w(L)[K_t^* - K_{t-1}^*].$$

It is easy to incorporate intermediate stages of the investment process into the theory. For concreteness, we consider the case of two intermediate stages, which will turn out to be anticipated investment, two quarters hence, and anticipated investment, one quarter hence. A similar approach can be applied to additional intermediate stages such as appropriations or commitments. The distribution of completions of the first stage, given new project starts, may be described by a sequence, say $\{v_{0\tau}\}$; similarly, the distribution of completions of a second stage, given completion of the first stage, may be described by a sequence $\{v_{1\tau}\}$. Finally, the distribution of investment expenditures, given completion of a second intermediate stage is described by a sequence $\{v_{2\tau}\}$. Where $I_t^{S_1 E}$ represents completions of the first stage, $I_t^{S_2 E}$ completions of the second stage, and I_t^E actual investment, as before, we have:

$$\begin{aligned} I_t^{S_1 E} &= \sum_{\tau=0}^{\infty} v_{0\tau} I_{t-\tau}^N = v_0(L) I_t^N, \\ I_t^{S_2 E} &= \sum_{\tau=0}^{\infty} v_{1\tau} I_{t-\tau}^{S_1 E} = v_1(L) I_t^{S_1 E}, \\ I_t^E &= \sum_{\tau=0}^{\infty} v_{2\tau} I_{t-\tau}^{S_2 E} = v_2(L) I_t^{S_2 E}. \end{aligned}$$

where $v_0(L)$, $v_1(L)$, and $v_2(L)$ are power series in the lag operator.

Up to this point we have discussed investment generated by an increase in desired capital stock. Total investment, say I_t , is the sum of investment for expansion and investment for replacement, say I_t^R :

$$I_t = I_t^E + I_t^R.$$

We assume that replacement investment is proportional to capital

stock. The justification for this assumption is that the appropriate model for replacement is not the distribution of replacements for a single investment over time but rather the infinite stream of replacements generated by a single investment; in the language of probability theory, replacement is a recurrent event. It is a fundamental result of renewal theory that replacements for such an infinite stream approach a constant proportion of capital stock for (almost) any distribution of replacements for a single investment and for any initial age distribution of capital stock. This is true for both constant and growing capital stocks. Representing the replacement proportion by δ , as before,

$$I_t^R = \delta K_t;$$

combining this relationship with the corresponding relationship for investment in new projects, we have:

$$I_t = w(L)[K_t^* - K_{t-1}^*] + \delta K_t.$$

Using the assumption that capital stock is continued in use up to the point at which it is replaced, we obtain the corresponding relationships for gross investment at each of the intermediate stages, say $I_t^{S_1}$ and $I_t^{S_2}$:

$$\begin{aligned} I_t^{S_1} &= v_0(L)[K_t^* - K_{t-1}^*] + \delta K_t, \\ I_t^{S_2} &= v_1(L)v_0(L)[K_t^* - K_{t-1}^*] + \delta K_t; \end{aligned}$$

we can also derive the following;

$$\begin{aligned} I_t^{S_2} &= v_1(L)[I_t^{S_1} - \delta K_t] + \delta K_t, \\ I_t &= v_2(L)[I_t^{S_2} - \delta K_t] + \delta K_t, \\ I_t &= v_2(L)v_1(L)[I_t^{S_1} - \delta K_t] + \delta K_t. \end{aligned}$$

For empirical implementation of the theory of investment behavior, it is essential that each of the power series— $v_0(L)$, $v_1(L)$, $v_2(L)$ —have coefficients generated by a rational function; for example,

$$w(L) = v_2(L)v_1(L)v_0(L) = \frac{s(L)}{t(L)},$$

where $s(L)$ and $t(L)$ are polynomials. We will call the distribution corresponding to the coefficients of such a power series a rational power series distribution. The geometric and Pascal distributions are among the many special instances of the rational power series distribution.

Empirical Results

To test the theory of investment behavior summarized in the preceding section, the corresponding stochastic equations have been fitted to quarterly data for U. S. manufacturing for the period 1948-60. The data on investment are taken from the OBE-SEC Survey; first and second anticipations of investment expenditure as reported in that Survey are taken as intermediate stages.¹ With two intermediate stages, six possible relationships may be fitted. First, for actual investment and both intermediate stages, the level of investment is determined by past changes in desired capital stock. Second, investment is determined by past values at each intermediate stage and the second anticipation is determined by past values of the first anticipation. The first test of the theory is the internal consistency of direct and derived estimates of the coefficients of each of the underlying power series in the lag operator.

The results of the fitting are given in Table 1. For each of the fitted relationships coefficients of the polynomials $s(L)$ and $t(L)$ in the ex-

¹ Data on capital stock were obtained by interpolating the capital stock series for total manufacturing given in the U.S. national accounts between 1949 and 1959, using the formula

$$K_{t+1} = I_t + (1 - \delta)K_t.$$

Given an investment series, a unique value of δ may be determined from the initial and terminal values of capital stock. Investment data from the OBE-SEC Survey were used for the interpolation. For desired capital stock, the quantity pQ was taken to be sales plus changes in inventories, both from the *Survey of Current Business*. User cost depends on a number of separate pieces of data. The quantity q is an investment deflator, δ is, of course, a fixed parameter (taken to be equal to .025), r is the U.S. government long-term bond rate. The tax functions vary with time; as an example, the tax rate u is the ratio between corporate income tax payments and corporate profits before taxes as reported in the U.S. national accounts. A detailed description of the data underlying this study will be reported elsewhere.

² To derive the form of the functions used in the actual fitting, we take $v_2(L)v_1(L)V_0(L)$ as an example. First:

$$I_t = \frac{s(L)}{t(L)} [K_t^* - K_{t-1}^*] + \delta K_t.$$

Secondly,

$$I_t = s(L)[K_t^* - K_{t-1}^*] + [1 - t(L)][I_t - \delta K_t] + \delta K_t.$$

The coefficient t_0 may be normalized at unity so that:

$$1 - t(L) = -t_1 L - t_2 L^2 - \dots$$

The a priori value $\delta = .025$ was used to compute $I_{t-\tau} - \delta K_{t-\tau}$. An estimate of δ is given by the coefficient of K_t . If δ is different from its a priori value, the process of estimation can be reiterated, using a second approximation to the value of δ .

The parameter γ is estimated using the constraint:

$$\sum_{\tau=0}^{\infty} w_{\tau} = 1.$$

TABLE 1

REGRESSION COEFFICIENTS AND GOODNESS OF FIT STATISTICS, UNRESTRICTED ESTIMATES

Regression	γ_{s_0}	γ_{s_1}	γ_{s_2}	b_1	b_2	δ	R^2	s	Δ^2/s^2
t ₂₁₇₀			.00102 (.00049)	-1.51911 (.09945)	.63560 (.10098)	.02556 (.00163)	.94265	.10841	2.14039
t ₁₇₀		.00132 (.00073)		-1.25242 (.12667)	.36656 (.12977)	.02618 (.00240)	.89024	.16229	2.00431
t ₀	.00109 (.00085)			-1.26004 (.13044)	.37281 (.13138)	.02549 (.00278)	.87227	.18974	2.37298
t ₂₁₁			.81357 (.03492)			.01962 (.00175)	.92729	.11955	1.16294
t ₁		.90024 (.02722)				.02295 (.00127)	.96234	.08604	1.47693
t ₁		.89462 (.03145)				.02337 (.00155)	.95121	.10597	1.70179

REGRESSION COEFFICIENTS AND GOODNESS OF FIT STATISTICS, RESTRICTED ESTIMATES

t ₂₁₇₀			.00106 (.00049)	-1.52387 (.09925)	.63100 (.10074)		.94156	.10830	2.10778
t ₁₇₀		.00133 (.00073)		-1.25704 (.12509)	.36769 (.12862)		.88986	.16087	2.00549
t ₀	.00109 (.00084)			-1.26395 (.12942)	.37300 (.13051)		.87128	.18848	2.53442
t ₂₁₁			.82764 (.04037)				.89995	.13883	.87127
t ₂		.91545 (.02933)					.95406	.09409	1.23560
t ₁		.90271 (.03276)					.94538	.11100	1.53759

pression for each power series as a rational function are given.² For example, the power series $v_2(L)v_1(L)v_0(L)$ is expressed as:

$$v_2(L)v_1(L)v_0(L) = \frac{.00106L^2}{1 - 1.52387L + .63100L^2}$$

The value of the replacement proportion δ estimated from data on capital stock is .025. Two sets of regressions were run, one with δ fitted from the data (unrestricted), the other with $\delta = .025$ (restricted). Throughout, the coefficient of multiple determination R^2 , the standard error of estimate for the regression s , and the Von Neumann ratio Δ^2/s^2 are presented as measures of goodness of fit.

The first set of tests of the theory is the comparison of alternative estimates of each of the fundamental power series. As an example, one may take the hypothesis that the direct estimates of the power series $v_2(L)$ and $v_1(L)$, when combined, give an estimate of $v_2(L)v_1(L)$ which is close to that obtained by direct estimation. Using the unrestricted estimates, the result of this comparison is:

$$.91545L \odot .90271L = .82639L^2;$$

the derived estimate, which may be compared with the direct estimate, $.82764L^2$. The difference between the two estimates is slightly over .03 standard errors. A similar test of the hypothesis that the direct estimates of the power series $v_1(L)$ and $v_0(L)$, when combined, yield an estimate of $v_1(L)v_0(L)$ which is close to that obtained by direct estimation results in

$$.00109 \cdot \frac{.90271L}{1 - 1.26395L + .37300L^2} = \frac{.00098L}{1 - 1.26395L + .37300L^2},$$

which may be compared with the direct estimate,

$$\frac{.00133L}{1 - 1.25704L + .36769L^2}.$$

The coefficient of the numerator is within half a standard error of the derived estimate. The coefficients of the denominator are within .06 and .04 standard errors of the derived estimates. The similarity of derived and direct estimates for the power series $v_2(L)v_1(L)v_0(L)$ is less striking. The three possible derived estimates are extremely similar to each other, but they differ considerably from the direct estimate. Nevertheless, using any of the derived estimates as the null hypothesis for a test of the direct estimates would probably lead to acceptance of the null hypothesis. In general, the theory of investment behavior is strongly confirmed by the set of tests of internal consistency. Of course, given the internal consistency of the alternative estimates, it is possible to improve efficiency of estimation for the model as a whole by combining information from the various sources.

The tests of internal consistency just described are tests of the theory of investment in new projects. A test of the theory of replacement investment is a test of the consistency of the empirical results with the hypothesis $\delta = .025$. This hypothesis is borne out in two ways. First, for all but one of the regressions, the usual null hypothesis is accepted; a much stronger result is that for the first three regressions, estimates of the relationships under the restriction that $\delta = .025$ results in a reduction in the standard error of estimate for the regression. Finally, each of the standard errors of the estimates of δ is less than one-tenth the size of the corresponding regression coefficient. We conclude that the hypothesis that replacement is a constant fraction of capital stock, specifically, that $\delta = .025$, is strongly validated by the empirical results.

We turn now to comparisons of the fitted regressions with some simple alternatives. First, as alternatives for the first three regressions, we take the naïve models:

$$I_t = I_{t-1},$$

$$I_t^{S_2} = I_{t-1}^{S_2},$$

$$I_t^{S_1} = I_{t-1}^{S_1}.$$

Simple as these models may be, they are quite stringent standards for comparison for seasonally adjusted quarterly data, much more stringent, for example, than the corresponding models for annual data. The appropriate statistics for comparison are the standard errors of estimate and the VonNeumann ratios. Results of this comparison are given separately for the periods 1948–60 and second quarter 1955 to 1960 in Table 2.³ For the period as a whole, each of the regression models has a standard error well below that for the corresponding naïve model. For the later subperiod the advantage of the regression models is even greater. Turning to the VonNeumann ratios, there is practically no evidence of autocorrelated errors for the fitted models and very clear evidence of autocorrelation for the naïve models. Of course, this test is biased in favor of the fitted regressions. Even with this qualification, the fitted regressions are clearly superior in every respect to the corresponding naïve models.

As a standard of comparison for the second three regressions, we take the forecasts actually used by the Department of Commerce in presenting the results of the OBE-SEC Survey. These alternative models take the form:

$$I_t = I_{t-2}^{S_1},$$

$$I_t = I_{t-1}^{S_2},$$

$$I_t^{S_2} = I_{t-1}^{S_1}.$$

Despite the high level of performance of the OBE-SEC anticipations data, the fitted regressions constitute a substantial improvement in both goodness of fit as measured by standard error of estimate and absence of autocorrelation of residuals. The test for autocorrelation is not biased in favor of the fitted regressions, so that the evidence is unequivocal; the fitted relationships are clearly superior to the corresponding forecasting models for the period as a whole and for the subperiod since second quarter 1955.

A further comparison of the fitted regressions with the corresponding naïve and forecasting models is given in the second half of Table 2, where an analysis of the conformity of turning points of each of the

³ Data for both anticipations and actual expenditures on a revised basis are available from the Department of Commerce only since the second quarter of 1955. Anticipations data for the earlier period were revised by multiplying each observation by the ratio of revised to unrevised actual investment for the period in which the observation was made.

TABLE 2
GOODNESS OF FIT STATISTICS: FITTED, NAIVE, AND FORECASTING MODELS

Model	1948I-1960IV			1955III-1960IV			1948I-1960IV			1955III-1960IV		
	R ²	s	Δ ² /s ²	R ²	s	Δ ² /s ²	TP Error	Over-estimate	Under-estimate	TP Error	Over-estimate	Under-estimate
$I_t = f(\Delta K_t^*)$.94156	.10830	2.10778	.94298	.11378	1.95800	29%	47%	24%	23%	41%	36%
$I_t^* = f(\Delta K_t^*)$.88986	.16087	2.00549	.94757	.09368	1.84699	29	43	27	41	32	27
$I_t^* = f(\Delta K_t^*)$.87128	.18848	2.53442	.91921	.14372	2.25394	39	33	27	41	27	32
$I_t = I_{t-1}$.86193	.15950	.66900	.81929	.18410	.52078	22	39	39	23	32	45
$I_t^* = I_{t-1}^{s_2}$.84966	.18058	1.04366	.81297	.19435	.94465	24	37	39	32	27	41
$I_t^* = I_{t-1}^{s_1}$.83854	.20282	1.31901	.81161	.19947	.83580	35	25	39	32	27	41
$I_t = f(I_t^{s_1})$.92729	.11955	1.16294	.92169	.12996	1.09855	16	43	41	23	41	36
$I_t = f(I_t^{s_2})$.96234	.08604	1.47693	.95931	.09368	1.65030	14	45	41	18	45	36
$I_t^* = f(I_t^{s_1})$.95121	.10597	1.70179	.96477	.09045	2.26525	16	45	39	23	41	36
$I_t = I_t^{s_1}$.83673	.17391	.69933	.77138	.20707	.46505	25	47	27	27	50	23
$I_t = I_t^{s_2}$.93504	.10969	.99342	.91854	.12359	.91146	14	51	35	18	50	32
$I_t^* = I_{t-1}^{s_1}$.93380	.11983	1.45737	.92833	.12031	1.25062	20	49	31	18	45	36

NOTE: Total percentages may not add to 100% because of rounding error.

"forecasts" to the turning points of the actual data is presented. In general, the first set of fitted regressions is slightly inferior to the naive models and the second set slightly superior to the forecasting models on the basis of this criterion. A final comparison is between the fitted regression of investment on changes in desired capital stock and the forecast of investment from its second anticipation. The comparison favors the fitted regression; however, the anticipations data used in a fitted relationship between investment and second anticipation provide a model which is superior to the simple forecasting model and to the fitted regression of investment on changes in desired capital stock.

Structure of the Investment Process

In the preceding sections, only those aspects of the theory of investment behavior relevant to testing the theory were presented. In this section certain further implications of the theory are developed. Specifically, we will characterize the long-term response of investment to changes in the underlying market conditions and the tax structure and the time pattern of response of investment to changes in demand for capital.

First, using the facts that gross investment is determined by the relationship:

$$I_t = w(L)[K_t^* - K_{t-1}^*] + \delta K_t$$

and that capital stock is determined by past investments, we obtain:

$$\begin{aligned} I_t &= [1 - (1 - \delta)L]w(L)K_t^*, \\ &= y(L)K_t^*, \end{aligned}$$

where $y(L)$ is a power series in the lag operator. We define the τ -period response of investment to a change in market conditions or tax structure as the change in gross investment resulting from a change in the underlying conditions which persists for τ periods. More precisely, suppose that desired capital remains at a fixed level for τ periods to the present; then,

$$K_t^* = K_{t-\nu}^*, \quad (\nu = 1, 2, \dots, \tau),$$

and

$$\begin{aligned} I_t &= \sum_{\nu=0}^{\infty} y_{\nu} K_{t-\nu}^*, \\ &= z_{\tau} K_t^* + \sum_{\nu=\tau+1}^{\infty} y_{\nu} K_{t-\nu}^*, \end{aligned}$$

where $\{z_r\}$ is the sequence of cumulative sums of the coefficients of $y(L)$. As an example, the response of gross investment to a change in the rate of interest is:

$$\frac{\partial I}{\partial r} = z_r \frac{\partial K^*}{\partial r}.$$

The coefficients $\{z_r\}$ characterize the time pattern of response. Obviously,

$$\lim_{r \rightarrow \infty} z_r = \lim_{r \rightarrow \infty} \sum_{v=0}^r y_v = \delta,$$

so that the long-term response of gross investment to changes in, say, the rate of interest, is

$$\frac{\partial I}{\partial r} = \delta \frac{\partial K^*}{\partial r}.$$

Clearly, the short-term responses approach the long-term response as a limit; the approach is not necessarily monotone, since the coefficients of the power series $y(L)$ are not necessarily non-negative.

Long-term response and elasticities of gross investment with respect to the price of output, price of capital goods, and the rate of interest are given in the top half of Table 3. The corresponding responses and elasticities for the income tax rate, the proportion of replacement and the proportion of interest chargeable against income for tax purposes are given in the bottom half of Table 3. It should be noted that the rate of interest and the tax rate are measured as proportions, not percentages. For example, a decrease in the rate of interest by 1 per cent increases manufacturing gross investment by \$.15178 billions per quarter in the long run, at least to a first approximation.

The time pattern of response is presented in Table 4, where the functions $w(L)$, $y(L)$, and $z(L)$ are derived from the fitted regressions. The

TABLE 3

RESPONSES AND ELASTICITIES OF INVESTMENT WITH RESPECT TO CHANGES IN MARKET CONDITIONS AND TAX STRUCTURE

	RESPONSE		ELASTICITY	
	Average	End of Period	Average	End of Period
Market Conditions				
Price of output.....	.35830	.35299	1.00000	1.00000
Price of capital goods....	-.35273	-.32106	-1.00000	-1.00000
Rate of interest.....	-14.23653	-15.17789	-.29143	-.37866
Tax Structure				
Income tax rate.....	-.37487	-.33016	-.50959	-.42064
Proportion of replacement	.18729	.20502	.39181	.48565
Proportion of interest....	.55656	.79840	.19428	.32659

TABLE 4
TIME FORM OF LAGGED RESPONSE

Lag	$w(L)$	$y(L)$	$z(L)$
0	0	0	0
1	0	0	0
2	.11277	.11277	.11277
3	.14209	.03214	.14491
4	.13700	-.00154	.14337
5	.11965	-.01393	.12944
6	.09969	-.01697	.11247
7	.08101	-.01619	.09628
8	.06491	-.01407	.08221
9	.05159	-.01170	.07051
10	.04081	-.00949	.06102
11	.03219	-.00760	.05342
12	.02535	-.00604	.04738
13	.01994	-.00478	.04260
14	.01567	-.00377	.03883
15	.01231	-.00297	.03586
16	.00967	-.00233	.03353
17	.00760	-.00183	.03170
18	.00597	-.00144	.03026
19	.00469	-.00113	.02913
20	.00368	-.00089	.02824
Remaining.....	.01346	-.00325	
Rate of decline.....	.78531	.78531	

average lag between change in demand for capital stock and the corresponding net investment is, roughly, 6.5 quarters or about a year and a half. Of course, this estimate is affected by the essentially arbitrary decision to set the proportion of the change invested in the same period and period immediately following the change at zero. The coefficients of the power series $z(L)$ are of interest for computation of short-period responses of investment to changes in the demand for capital stock. For example, the 2-period response of manufacturing gross investment to a change in the rate of interest of 1 per cent is:

$$z_r \frac{\partial K^*}{\partial r} = \frac{.11277}{.02500} \cdot .15178 = .68465 \text{ billions/quarter.}$$

By comparison, the corresponding 10-period response is .37046 billions per quarter. The response dies out, almost to its long-term level of .15178 billions/quarter, by twenty periods from the initial change in demand for capital stock. Similar calculations of the response of gross investment to changes in market conditions or the tax structure may be made for any of the six determinants of demand for capital by combining the responses given in Table 3 with the time pattern presented in Table 4.