1992

Assorted Non-Shaikh 5

Anwar Shaikh PhD

ASSORTED
NON SHAIKH J

JPHIL. Vol 18, no 4
S.

JPKE, Vol 18, No 4
Summer 1996

CHRISTOPHER G. FULLER 1996

Elements of a Post Keynesian alternative to "household production"

Recently a prominent Post Keynesian economist pointed out that, with a few exceptions, "Post Keynesians have been relatively silent about the microeconomics of household choice" (Lavoie, 1994, p. 539). Important contributions have been made by, for instance, Earl (1986) and Drakopoulos (1992) in the goods and brand choice-making behavior of individuals, but Post Keynesian economists have certainly been silent in projecting and subsequently modeling a vision of "consumption activity." By this, we mean a vision of the social context and the institutional structure within which time use and expenditure decisions are made by well-defined groupings of individuals (as opposed to the more specific analyses of an individual's goods-brand choice by Earl and Drakopoulos).[1]

The dominant economic vision of consumption activity is Gary Becker's neoclassical "household production" approach. Of course, Post Keynesians have offered important direct and implicit critisisms of household production theory. In this paper a number of these critisisms are considered for their usefulness in the development of a Post Keynesian

The author is in the Department of Economics at the University of East London. A version of this paper was presented at the 1995 Eastern Economics Association Conference in New York. I wish to thank the anonymous referee for helpful comments. The usual disclaimer applies.

[1] Post Keynesians differ in the relative emphasis they give to, first, the idea that irreversible time and "Knightian uncertainty" must be accepted as facts and, second, the idea that socioinstitutional structures should be recognized from the beginning within economic analysis. This has naturally led to systematic differences of emphases in Post Keynesian approaches to macroeconomics and the microeconomics of firms—contrast Post Keynesian macroeconomics with neo-Ricardian growth theories and behavioralist theories of the firm with cost plus/oligopoly-based Post Keynesian approaches. While Post Keynesian microeconomic analyses of consumer behavior should logically reflect a similar difference of emphasis, they do not seem to do so at present.

vision of consumption activity. We then suggest a way of reworking criticisms of Becker's approach, showing how elements of an alternative Post Keynesian vision of consumption activity might emerge.

Post Keynesian responses to household production theory

Elements of household production theory

The application of Gary Becker's "economic" approach to human behavior in consumption activity is best known as "household production theory." We consider in turn three features of this approach in particular:

1. *Global rationality*: All observable human activity can be understood as choice of market goods purchases and time "inputs" in the production of nonmarket "commodities" to maximize utility.
2. *Competition-driven social relations*: Social interactions between consumers involve competition to produce "distinction" as a "scarce" utility-yielding commodity. More distinction generated through social relations for one person implies, *ceteris paribus*, less for others.
3. *Household-centered production*: The institutional base for consumption activity is assumed to be the household, which gives rise to the underlying technology of commodity production functions.

Post Keynesians on global rationality

Becker's household production theory is part of a more general project to attempt to understand all aspects of human behavior using tools of neoclassical price theory and the assumption that observed actions result as if from the solution to a single optimization problem. Post Keynesians consider that as a consequence of human cognitive and time limitations, any such project is misconceived (Nicolaides, 1988; Hodgson, 1988). Rational decision makers must make use of rules that have themselves not been derived from optimizing principles—or indeed from any calculations (Earl, 1986; Hodgson, 1988; Lavoie, 1994). These rules may be subjective guesses or socially observed norms or conventions. Post Keynesians therefore consider that Becker's implicit presumption of global rationality must be rejected.

Post Keynesians are clear in their alternative espousal of bounded/procedural rationality in the spirit of Herbert Simon's original (1955) argument, and this has provided an important justification for the modeling of processes of goods purchase through lexical and "noncompensatory" choice procedures (Earl, 1986; Drakopoulos, 1992). Nevertheless, as Hahn and Hollis (1979, p. 11) have noted in a largely sympathetic discussion of bounded rationality, this assumption "is descriptively plausible but has not so far proved theoretically useful, since the aspiration levels and the search activities are ill-defined."

This criticism is applicable to, but not serious for Earl and Drakopoulos's analyses since such approaches focus largely on analyzing the choice processes of an individual over a specific set of goods. However, the criticism has much more force when the concern is with a view of the structure of time usage and expenditures across groupings of individuals—a vision of "consumption activity." In the latter case, the axiom of procedural rationality must be combined with a view of both the nature of social relations and the institutional context for consumption activity if hypotheses about general patterns of time usage and expenditure decisions across groups of persons are to be developed. In what follows, the extent to which Post Keynesian propositions in these areas can help to guide development of an alternative vision is considered.

Post Keynesians on social relations

Gary Becker's application of household production to "social interactions" (1974) enabled him to claim that neoclassical consumer theory no longer suppressed the "social" dimension to goods purchase behavior—a frequent criticism leveled by Post Keynesians, among others. Becker assumes the individual has a stock or endowment of "social distinction" and can "purchase" aspects of its social environment favorable to the production of distinction via its own efforts.[2]

Becker's approach to social relations in consumption has three features. First, he deliberately avoids the use of behavioral assumptions widely accepted in other disciplines but not by neoclassical price theory.[3] Becker's model accounts for apparently "socialized" behavior (such as conformism and "keeping up with the Joneses") by assuming

[2] For instance, by purchasing more "fashionable clothing," the individual can produce more favorable opinions from others about him or herself and so increase distinction.

[3] If others in person A's social environment have relatively more distinction—because, say, they purchase more fashionable clothing—then A experiences a reduction

that commodities such as "distinction" are scarce.[4] Social relations among humans become a process of competition for utility-yielding distinction. Second, Becker treats social relations as simply another way each person can "purchase" utility. To spend time in social relations has no function for individuals that is not reducible to utility. Third, since "distinction" is assumed to be scarce in Becker's view, each person gains from social relations at the expense of others.

On the first point Post Keynesians reject Becker's deliberate avoidance of "noneconomic" behavioral assumptions. Frequent allusions are made to the way in which expenditure patterns are molded by the effects of socialization processes upon consumers as social class members as the following quotations cited by Lavoie (1994) indicate:

> There is a kind of competition in consumption, induced by the desire to impress the Joneses, which makes each family strive to keep up at least an appearance of being as well off as those that they mix with, so that outlay by one induces outlay by others. [Robinson, 1956, p. 251]
>
> A household's consumption pattern, at any given point in time . . . reflects the lifestyle of the households that constitute its social reference group. [Eichner, 1986, p. 160]
>
> The consumption of each class will be guided by a conception of its appropriate lifestyle, given its place in the social pyramid. . . . Emulation effects normally follow the social hierarchy; the consumption styles of the rich and famous set standards to which the rest aspire (or sometimes, against which they react). [Nell, 1992, pp. 393, 396]

An eclectic theoretical strategy, while typical of Post Keynesian theorizing in general, has particularly characterized Post Keynesian work here. Such eclecticism is justified on the basis that an open approach

---

in his or her "endowment" of distinction. Given "well-behaved" preferences, person A will tend to increase his or her efforts to produce distinction, relative to spending on other commodities in response to this constraint change. Hence, A will spend more on fashionable clothing.

[4] As Becker (1976) argues,

> For economists to rest a large part of their theory of choice on differences in tastes is disturbing since they admittedly have no useful theory of the formation of tastes, nor can they rely on a well-developed theory of tastes from any other discipline in the social sciences, since none exists. [p. 133]

Hence the neoclassical economist should instead "[continue] to search . . . for the subtle forms that prices and incomes take in explaining differences among men and periods" (Becker and Stigler, 1977 p. 76).

sympathetic to interdisciplinary discourse and cooperation over assumptions acknowledges that economists do not have a monopoly of wisdom about human behavior. This is a valuable aspect of Post Keynesian method. However, at some point a specific theoretical position must be taken as to the way in which social relations among persons generate uniformities in time uses, life-styles, and expenditure patterns to justify talk of "social classes." Theoretical abstinence here may be a manifestation of academic modesty but would be unhelpful if further flesh is to be added to the bones of the axiom of "procedural rationality."[5]

On the second and third points, Post Keynesians have been very quiet. A rare exception is Mary McNally (1980), who argues for the idea of a "social process" in consumption and that there is an implied conception of consumption activity as rooted in a process of interpersonal relations within which goods are purchased and used. However, this conception is incorporated within a modified Beckerian framework. This retains the idea that social relations and goods acquisition are essentially alternative ways of generating utility. No conceptual hierarchy is made explicit between the purpose of allocating time and expenditures to participation in interpersonal relations and the purpose of allocating time and goods to produce nonmarket "commodities." More generally, Post Keynesians have not considered the possibility that commitments to spend time and expenditures in participation in social relations may be a greater priority than the expenditure of such resources in other areas because this is a "rule" of procedural rationality. Little attention has also been paid to Becker's exclusion of the possibility that interpersonal associations are not uniformly "competitive" but instead involve aspects of mutual or shared achievement among persons.

Although used to justify an eclectic approach to consumption, the Post Keynesian view of academic relationships could be used in support of a similar theoretical view of social relationships between consumers. If academic activity is viewed as the search for cooperative interdisciplinary communication, the sane perspective might be adopted for the

[5] Some might argue that it is sufficient to use a macroeconomic class-based consumption function approach to account for expenditure patterns. But such an approach alone offers no explanation of the forces linking patterns of time uses in social relations to expenditure patterns that can be identified with particular "life-styles" or "classes." Such a macro-oriented approach excludes the possibility of an analysis of how such patterns might emerge as a result of material constraints (as well as normative and capability constraints) upon patterns of time spent in social relations among persons in consumption.

analysis of consumer relations. An emphasis upon interpersonal communication in economically functional social relations would also have an impeccable lineage. Consider a neglected aspect of a frequently quoted observation from Adam Smith:

> Whether . . . the propensity to truck, barter and exchange one thing for another . . . be one of those original principles in human nature of which no further account can be given, or whether, as seems more probable, it be the necessary consequences of the faculties of reason and speech, it belongs not to our present object to inquire. It is common to all men, and to be found in no other race of animals, which seem to know neither this, nor any other species of contracts." [Smith (1776), 1970, vol. 1, pp. 117–118]

Neoclassicals emphasize Smith's suggestion that the mainspring to economically functional human relations is "human nature." Smith's second, "more probable" possibility, that human "faculties of reason and speech" are the stimulus to such relations, is rarely mentioned. Yet the role of word-of-mouth information usage by consumers is accepted as a commonplace in the marketing research literature (Brown and Reingen, 1987). Furthermore, in their remarks, if not in their formal models, neoclassical economists have often conceded the importance of the acquisition of information through word-of-mouth sources and personal contacts in the labor market (Rees, 1966), the insurance market (Kunreuther, 1978), and in voting decisions (Downs, 1957). The argument that rational human action is "embedded" within a structure of ongoing relations with specific others is a theme of sociologists (Granovetter, 1985; Anderson et al., 1994) and some institutionalist-oriented social scientists (Thompson et al., 1991). Furthermore, the work of Douglas and Isherwood (1980) into an anthropological understanding of consumption activity as communication centered and recent sociological work into consumer "life-styles" and the projection of "self-identity" through the usage of goods (Tomlinson, 1990; Featherstone, 1991; Keat et al., 1994, ch. 3) are also consistent with a view of consumption as a process of goods usage in order to facilitate communication within personal associations.

We therefore suggest that Post Keynesians adopt a stronger position relating to the role of social relations in consumption activity and advance the following four propositions: First, that consumption activity be viewed principally as a process of social relations among persons in which physical goods are used. Second, that social relations in consumption activity involve the active pursuit and maintenance of cooperative nonmonetary relations or "personal encounters" with other identifiable persons face to face. Third, that in such relations the joint activity of mutual word-of-mouth interpersonal communication and mutual transfer of certain types of nonmarket "services" occurs.[6] Fourth, that physical market goods are used as a means of facilitating such functional social relations.

Post Keynesians on the household as the center of consumption activity

The household is seen as the physical base for consumption activity in Becker's theory in an analogy to the factory as the base for the firm. Peter Earl (1986) has observed that

> it is disappointing to see that household production theory has hitherto only explored the analogy in terms of the neoclassical theory of the firm. [Yet] . . . further significant insights are to be obtained by looking at household choices from the standpoint of the behavioral theory of the firm. [Earl, 1986, p. 40]

The issue of whether the household-firm analogy is appropriate—however the analogy is pursued—has not been raised by Post Keynesians. The reluctance to question this analogy is to a degree understandable. The household has been used as the institutional unit because it is a fixed location or base for the biological family unit; the family itself retains a degree of ongoing stability due to the majority acceptance of the legal institution of marriage and statistics relating to consumer characteristics are based on an assumed household unit. In any case, denial of the relevance of the household as one institutional base in which consumption activity occurs would be an untenable position. Nevertheless, the exclusive focus of attention upon the household should at least be questioned. First, the household is not the only institutional location for consumption: there are many widely used institutional environments in which goods are used by consumers in their nonwork time—public

[6] The "services"—physical performance of "favors" on the spot and supply of verbally communicated data—we suggest are transferred mutually between personal contacts have a strong similarity to those provided by health and education services. The central point about such services is that "the acts of production and consumption of the benefits of the service are simultaneous, i.e., no intermediate consumption good is produced" (Brown and Jackson, 1990, p. 129). The benefits of such relations are specific to participation in such relations. Such services cannot be produced by a supplier, then "purchased" by a consumer in anonymous monetary exchanges, and then "consumed" at some other time and place.

houses, restaurants, theaters, wine bars, places of worship, retail outlets, sports, health and fitness clubs, public libraries, museums, parks, and so on. These have no obvious place within frameworks centered upon the household, relegated instead to aspects of the theory of public goods and clubs. Second, marketeers and advertisers have shifted their emphasis in the last twenty years from simple demographics and the family life cycle to "life-style" profiles (emphasizing activities, interests, and opinions of individuals) and "psychographics"—classifying individuals by personality types—as a means of consumer segmentation. This increased usage of psychographic data suggests a recognition that the household-centered demographic approaches

> do not provide a sufficiently broad view of consumers in the sense of explaining how their life patterns influence purchasing decisions. [Williams, 1981, p. 91]

Furthermore, there are at least two adverse consequences of locating consumption activity within the household. The first is a neglect of the spatial dimension to consumption activity. The importance of mobility within a consumption infrastructure to consumption activity is downgraded, leaving the costs involved in both visits to and involvement within this infrastructure as something to be considered separately by specialists in transport and urban economics. Time allocation, service usage, and expenditures involved in "circulation" within the consumer's environment are not integrated from the outset into the structure of consumer expenditures and time usage in household production theory. This also means that the *access* differing degrees of mobility external to the household bring to the range of direct consumption experiences a given consumer systematically enjoys cannot be recognized or analyzed. Socialization—even if recognized and integrated into household production theory—would be confined to the household and involve family influences only.[7]

The second consequence of a household-centered approach is a treatment of all extra-household institutions as social details, with no economic function for individuals. We have noted the infrastructure of commercially, publicly, and voluntarily provided instititions for consumption that individuals choose to use goods within. It is common to see defended the economic function of "social overhead capital" (communication and transport linkages) for producers in underdeveloped countries. It is less common to see an explicit function given to the parallel "consumer social overhead capital" that enables individuals more effectively to undertake consumption activity.

In view of our argument for a cooperative, communicative vision of social relations in consumption, the parallel "infrastructure" we identify may indeed have a clear economic function as is suggested by Boland below:

> One of the roles that institutions play is to create knowledge and information for the individual decision-maker. In particular, institutions provide social knowledge which may be needed for interaction with other individual decision-makers. [Boland, 1979, p. 963]

This infrastructure may function to inform consumers of behavior "appropriate" to effective interpersonal communication during their planned or accidental encounters with others within it. A concept of such an infrastructure would raise the issue of differences in the number of the environments[8] and "quality" of the environments each consumer has access to. Aspects such as relative air cleanliness, absence of noise, heating, seating, and lighting facilities become relevant to a conception of the "quality" of such environments given that their function is to facilitate association and interpersonal communication among persons.[9]

Toward a "monetary" paradigm for consumption activity?

So where should Post Keynesian consumption theory go from here? We have so far argued that first, Post Keynesians should adopt essentially the view of consumption activity as a process of cooperation-seeking behavior through interpersonal communication in which goods have a

[7] Despite the importance of motor car possession and access to its use for consumption activity in the twentieth century, household production theory offers no way of integrating into its analysis the fact that a person has access to the use of such a good. No acknowledgment can be made of the increased range of consumption opportunities and experiences created by access to use of a car.

[8] In particular, household production offers no way of analyzing the effects upon car and non-car users of changes in the "consumption infrastructure"—particularly the shopping environment—arising from commercial incentives to cater to an increasing proportion of consumers as motor car users. Yet there may be significant adverse consequences for non-car users of changes in this infrastructure introduced to complement car users' consumption activity patterns.

[9] For instance, the spatial, environmental, and safety channels through which different dimensions of the quality of consumption activity are affected by motor car parking, emissions, noise, and movement could become much more transparent if the perspective advocated here were adopted.

facilitating role. Second, we have suggested that the household-firm analogy must be dispensed with, not modified. It should instead be assumed that individuals circulate—among other individuals—within an array of functional physical environments beyond the household. Given these points of departure, it will be necessary to mesh the concept of "procedural rationality" with this view of social relations and a functional consumption infrastructure. For instance, the observation—based upon actual consumer behavior—that individuals are "embedded" in ongoing social relations implies an assumption about the way in which individuals prioritize their nonwork time use. How a priority investment of time in a process of interpersonal communication with others is a rule of procedural rationality must be explained. Moreover, given such embeddedness, conduct in consumption will involve behavioral rules designed to achieve the individual's aims in such relations. What aims do individuals generate through their embeddedness in such relations? What conduct is involved in achieving such aims? Finally, this theory must clarify what are perceived to be the major determinants of time usage, goods purchases, and expenditures on other services. The nature of the facilitating role goods play as part of individual "conduct" in ongoing social relations requires clarification. Furthermore, since we assume consumption involves circulation among persons in different physical environments, the determinants of patterns of time usage and associated expenditures upon circulation must be clarified.

The above suggestions dispense with the household-firm analogy, but not with the use of analogies per se. Indeed, an alternative analogy—to the "neo-Marshallian" analysis of monetary microfoundations (see, for instance, Clower in Walker, 1984)—might be exploited.[10] In that literature, the emphasis is upon economically functional and costly to organize exchange relations among persons who "fumble and grope rather than optimise" (Clower, 1975, in Walker, 1984, p. 194). By analogy, we have posited the existence of a set of economically functional non-money-mediated social relations—"personal encounters"—at the heart of consumption activity. In the monetary literature, "trading posts" are given a rationale as a set of quasi-public goods: an exchange infrastructure that reduces the search costs associated with exchange. Similarly, we have argued that there exists a consumption infrastructure that assists individuals in setting up and participating in personal encounters. Moreover, monetary theory has recognized that in a costly-to-

[10] In Fuller (1995) this analogy is considered in more detail.

exchange world, the social rule of usage and acceptance of identifiable physical notes and coin increases exchange opportunities. We have argued that market goods are used to "facilitate" individuals in some sense during communication in such relations. Thus, both notes and coins and physical goods might be understood as "currencies" to the extent that they facilitate a coexisting and complementary set of functional economic relations among individuals. The prospect of a "monetary theory of consumption" might then be possible.

Conclusion

This paper has considered and reworked Post Keynesian criticisms of neoclassical choice theory with Becker's household production theory particularly in mind. We have argued that the way in which such criticisms have been followed up has restricted the development of a Post Keynesian view of "consumption activity"—as opposed to an approach to individual choice-making processes—that is distinct from household production. It was suggested that a clear adherence to a concept of procedural rationality is a necessary but not sufficient condition for a vision of consumption activity. The strong Post Keynesian sympathy for "openness" to other disciplines could be used to justify a clear alternative to the competitive vision adopted by Becker. Finally, while Post Keynesians rightly decry the way in which household-firm analogies are formalized, this seems to have diverted attention from challenging the analogy itself.

In the light of these points we have indicated instead a paradigm of procedurally rational individuals investing their nonwork time and using goods in communicative, cooperation-seeking relations with specific others, while circulating within a functional consumption infrastructure, of which the household is but one component. Consumer expenditures therefore relate to commitments arising from a multidimensional "participatory" process of communication and circulation by individuals, rather than to a one-dimensional "production" concept of household commodity generation. The household-firm analogy has been a catalyst for innovation in the neoclassical theory of consumption activity. Perhaps a different analogy—to monetary theory—would be appropriate for Post Keynesians.

REFERENCES

Anderson, Michael; Bechhofer, Frank; and Gershuny, Jonathan. *The Social and Political Economy of the Household*. Oxford: Oxford University Press, 1994.

Becker, Gary. "A Theory of Social Interactions." *Journal of Political Economy*, 1974, *82* (6), 1063–1091.

———. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press, 1976.

Becker, Gary, and Stigler, George. "De Gustibus Non Est Disputandum." *American Economic Review*, 1977, *67* (2), 76–91.

Boland, Laurence A. "Knowledge and the Role of Institutions in Economic Theory." *Journal of Economic Issues*, 1979, *13* (4), 957–972.

Brown, C.V., and Jackson, P.J. *Public Sector Economics*, 4th ed. Oxford: Blackwell, 1990.

Brown, J.J., and Reingen, P.H. "Social Ties and Word-of-Mouth Referral Behavior." *Journal of Consumer Research*, 1987, *14*, 350–362.

Clower, Robert W. "Reflections on the Keynesian Perplex." *Zeitschrift für Nationalokonomie*, 1975, *35*, 1–24.

Downs, A. *An Economic Theory of Democracy*. New York: Harper and Row, 1957.

Drakopoulos, Stavros A. "Psychological Thresholds, Demand and Price Rigidity." *Manchester School of Economics and Social Studies*, June 1992, *40*, 152–168.

Earl, Peter. *Lifestyle Economics*. Brighton, UK: Wheatsheaf, 1986.

Eatwell, John; Milgate, Murray; and Newman, P., eds. *The New Palgrave, General Equilibrium*. London: Macmillan, 1989.

Eichner, Alfred. *Towards a New Economics: Essays in Post-Keynesian and Institutionalist Theory*. London: Macmillan/M.E. Sharpe, 1986.

Featherstone, Mike. *Consumer Culture and Postmodernism*. Beverly Hills, CA: Sage Publications, 1991.

Fuller, Christopher G. "A Suggestion for Complicating the Theory of Consumption." 1995 (mimeo).

Granovetter, Mark. "Economic Action and Social Structure: The Problem of Embeddedness." *American Journal of Sociology*, 1985, *91*, 481–510.

Hahn, Frank H., and Hollis, M., eds. *Philosophy and Economic Theory*. Oxford: Oxford University Press, 1979.

Hodgson, Geoffrey, M. *Economics and Institutions: A Manifesto for a Modern Institutional Economics*. Cambridge: Polity Press, 1988.

Isherwood, Baron, and Douglas, Mary. *The World of Goods: Towards an Anthropology of Consumption*. Harmondsworth, UK: Penguin, 1980.

Keat, Russell; Whiteley, Nigel; and Abercrombie, Nicholas. *The Authority of the Consumer*. London: Routledge, 1994.

Kunreuther, H., et al. *Disaster Insurance Protection: Public Policy Lessons*. New York: Wiley, 1978.

Lamberton, Donald, M., ed. *Economics of Information and Knowledge*. Harmondsworth, UK: Penguin, 1970.

Lavoie, Marc. "A Post Keynesian Approach to Consumer Choice." *Journal of Post Keynesian Economics*, Summer 1994, *16*, 539–562.

McNally, Mary. "Consumption, Utility and Social Process." *Journal of Post Keynesian Economics*, 1980, *2*, 381–391.

Nell, Edward. "Demand, Pricing and Investment." In *Transformational Growth and*

Nicolaides, Phedon. "Limits to the Expansion of Neoclassical Economics." *Cambridge Journal of Economics*, 1988, *12*, 313–328.

Rees, A. "Information Networks in Labor Markets." *American Economic Review*, 1966, *56* (2), 559–566.

Robinson, Joan. *The Accumulation of Capital*. London: Macmillan, 1956.

Simon, Herbert A. "A Behavioral Model of Rational Choice." *Quarterly Journal of Economics*, 1955, *69*, 99–118.

Smith, Adam. *The Wealth of Nations* (1776). Vol. 1. Harmondsworth, UK: Pelican 1970.

Thompson, Grahame; Frances, Jennifer; Levacic, Rosalind; and Mitchell, Jeremy. *Markets, Hierarchies and Networks*. Beverly Hills, CA: Sage Publications, 1991.

Tomlinson, Alan, ed. *Consumption, Identity and Style*. London: Routledge, 1990.

Walker, Donald, A., ed. *Money and Markets: Essays by Robert W. Clower*. Cambridge: Cambridge University Press, 1984.

Williams, K.C. *Behavioral Aspects of Marketing*. London: Heinemann, 1981.

and Keynesian interest rate theories, the analysis is fundamentally Keynesian, in the sense that equilibrium can exist with the income below the full employment level.

QUESTIONS

1. Suppose that the transactions demand for money is completely interest-inelastic. That is, suppose that the quantity of money demanded for transactions purposes is not affected at all by changes in the interest rate, but is dependent only on the level of income. How will this affect the following?
   a. Figure 5–5.
   b. Figure 5–6.
   c. The effectiveness of fiscal policy in the revised version of figure 5–11.
2. Consider an initial equilibrium with the interest rate at a minimum in Fig. 5–12, with the LM curve initially in position LM_3.
   a. If the money supply is increased by 10 percent, what will be the effects on (i) the equilibrium level of national income? and (ii) the velocity of money?
   b. Suppose one wishes to reject answer (i) in *a* above, and argue that expansive monetary policy is effective in stimulating the level of aggregate demand. What must be argued about the shape or movability of the IS or LM curves?

SUGGESTED READING

Alvin Hansen, *A Guide to Keynes* (New York: McGraw-Hill, 1953), *Chap. 7.*

Macroeconomics
Paul Wonnacott, 1974
Richard D. Irwin, Illinois
USA

6

A classical rebuttal

Hey Diddle Diddle
Distribute the Middle
The Premise controls the Conclusion. . . .[1]
Frederick Winsor

THE previous chapter presented a theoretical system more comprehensive than either the simple Keynesian or simple classical systems, but with elements of each. The IS/LM analysis may therefore be considered a partial integration of Keynesian and classical economics. But in one important respect, the Keynesian and classical systems cannot be integrated or compromised, since they lead to flatly contradictory conclusions: Keynesian theory establishes the possibility of an equilibrium at less than full employment even with wage and price flexibility, while classical economics holds that equilibrium *must* involve full employment unless wages and/or prices are inflexible in the face of inadequate demand. These two conclusions are clearly oil and water: they cannot be mixed or compromised. Nor can the related conflict over money be mixed or compromised. In the Keynesian system, it may be impossible to raise aggregate demand to the full-employment level by increasing the money stock. In the classical system, it is possible to stimulate aggregate demand to the full-employment level by a sufficient increase in the quantity of money.

The IS/LM analysis holds out the possibility of an equilibrum with less than full employment. It is therefore fundamentally a Keynesian analysis, although it does contain some of the elements of the classical theory of interest. The IS/LM analysis must therefore either be subject to a rebuttal from the classical viewpoint, or Keynesian economics must be granted a sweeping victory in the theoretical debate.

[1] From Frederick Winsor and Marian Parry *The Space Child's Mother Goose* (New York: Simon and Schuster, 1958).

The first part of this chapter presents a classical rebuttal. This rebuttal is *highly theoretical, and has little direct policy significance.* In particular, it investigates the *equilibrium* effects of general price and wage reductions—omitting the dynamic consequences of deflation (especially in terms of expectations) which would have to be put at the center of any discussion of deflation as a practical policy alternative.[2] However, general deflation is a particular means of increasing the real quantity of money in a static equilibrium framework. Therefore, the theoretical points raised in the first part of this chapter do have some relevance for the evaluation of monetary policy—which involves increasing the real quantity of money in an alternative way, through increases in the nominal quantity rather than through a general deflation.

The latter part of this chapter will elaborate on the classical theme, presenting criticisms of the Keynesian theoretical framework, and especially the Keynesian concept of an equilibrium multiplier.

Keynes attacked classical theory on the ground that the capital markets might prove inadequate; specifically, he argued that, if the full-employment savings schedule intersected the full-employment investment schedule at a negative rate of interest, there would be unemployment, since the nature of money and the capital markets rule out the possibility of a negative interest rate (Fig. 5–2). However, the equalization of full-employment savings and investment through changes in the rate of interest represented *only one of two* classical mechanisms for ensuring full employment in equilibrium. The other, which Keynes ignored, provided the basis for the theoretical classical counterattack.

THE QUANTITY THEORY AND THE REAL-BALANCE EFFECT

Suppose we go back to the simplest classical quantity theory ($MV = PT$, with V reasonably stable). Aggregate demand is an increasing function of the quantity of money, and therefore full employment will result if the real money supply is increased enough either through

[2] The purely theoretical basis of the classical deflation argument was at times explicitly stressed by classical writers. For example, in the preface to *Lapses from Full Employment* (London: Macmillan, 1945), p. v, classical standard-bearer A. C. Pigou cautioned: "Professor Dennis Robertson, who has very kindly read my proofs, has warned me that the form of this book may suggest that I am in favour of attacking the problem of unemployment by manipulating wages rather than by manipulating aggregate demand. I wish, therefore, to say clearly that this is not so." Nevertheless, Pigou and other classical writers were not always so careful, and their writings at times suggested that wage and price cuts were, indeed, an appropriate antidepression measure.

general deflation or through an increase in the nominal money supply while prices remain constant. What is the logic of this argument?

What the classicists held, simply, was that if people have more real money, they will spend more. If their wealth in the form of money increases, they will be in a position to increase their levels of consumption, and they will do so. In other words, classical economists argued that something is left out of the Keynesian consumption function, namely, the stimulative effect of increases in individuals' money holding on their consumption behavior. Classical economists believed that, if people are provided with *enough* money, there is *no limit* to the amount they are willing to consume, since consumer wants are unlimited; full employment can therefore be assured by a sufficiently large increase in the quantity of money.

The effect of a change in the real stock of money on the level of consumption is known by a variety of names: the real-balance effect, the wealth effect, or the Pigou effect.[3] The theoretical significance of the real-balance effect is fundamental: If increases in the (real) money supply do indeed stimulate consumption, then the Keynesian liquidity trap is invalid. Additional quantities of money are not caught in this trap and prevented from affecting the level of aggregate demand; rather they stimulate aggregate demand by shifting up the consumption function.

The real-balance effect and the IS/LM analysis

The IS/LM analysis divides the economy into two markets: the product market, with the IS function showing the locus of equilibrium points for this market; and the financial market, with the LM curve showing the equilibrium points for this second market. If the real-balance argument is accepted as plausible, *this dichotomy must be rejected. The product market and the monetary market are not separable.* Rather, an increase in the (real) quantity of money will cause consumption to shift upward as a function of income, and will therefore cause the IS curve to shift upward to the right, as illustrated in Figure 6-1 (an elaboration of Figures 5–3 and 5–4). An increase in consumption out of any given level of income is the same thing as a decrease in savings. Thus, as the real-money supply increases from m_1 to m_2, the savings function for income level Y_1 shifts from $S_{m_1}Y_1$ to $S_{m_2}Y_1$. Thus, equilibrium point A_1 moves to

[3] Because of A. C. Pigou's early attack on Keynesian theory, "The Classical Stationary State," *Economic Journal*, 1943, pp. 343–51.

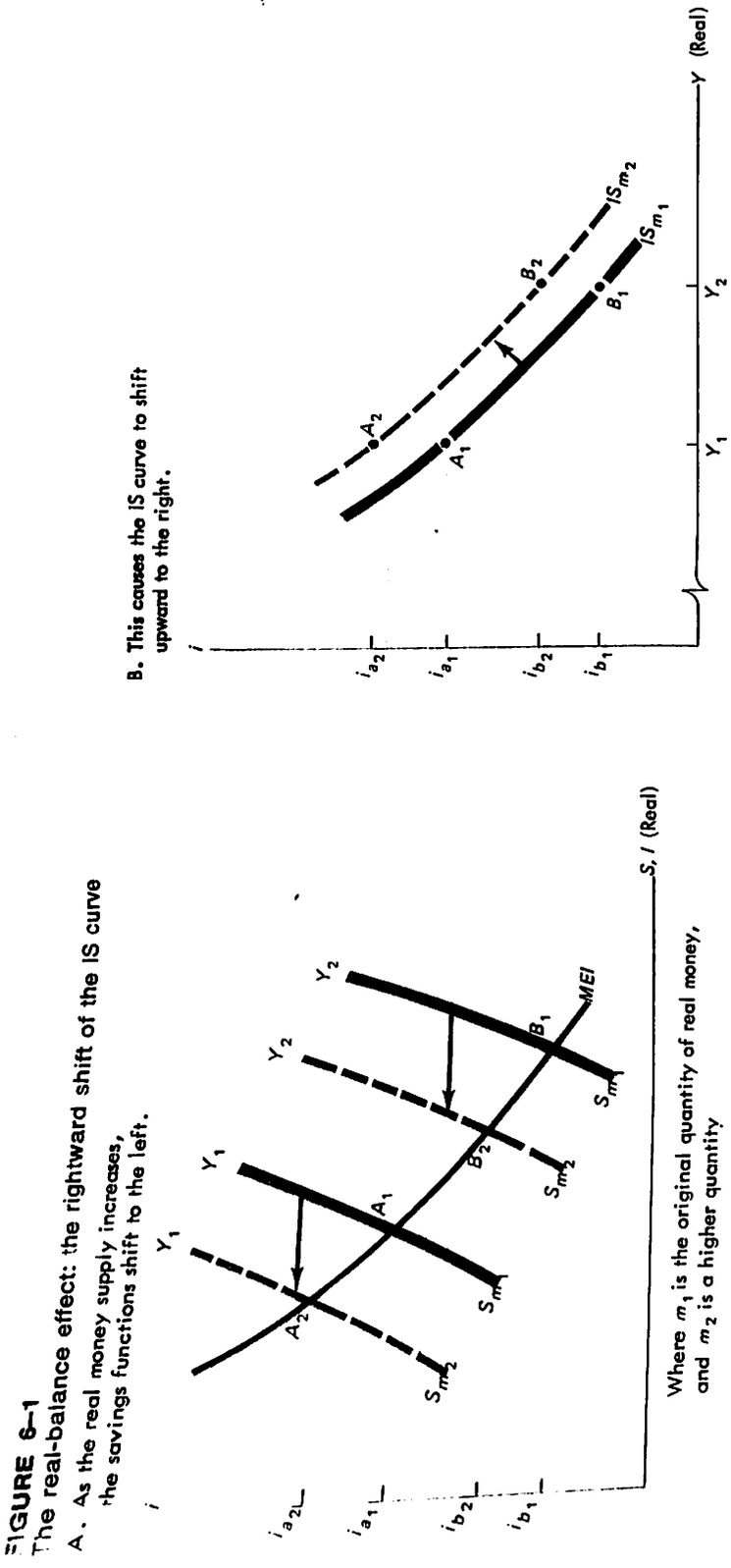
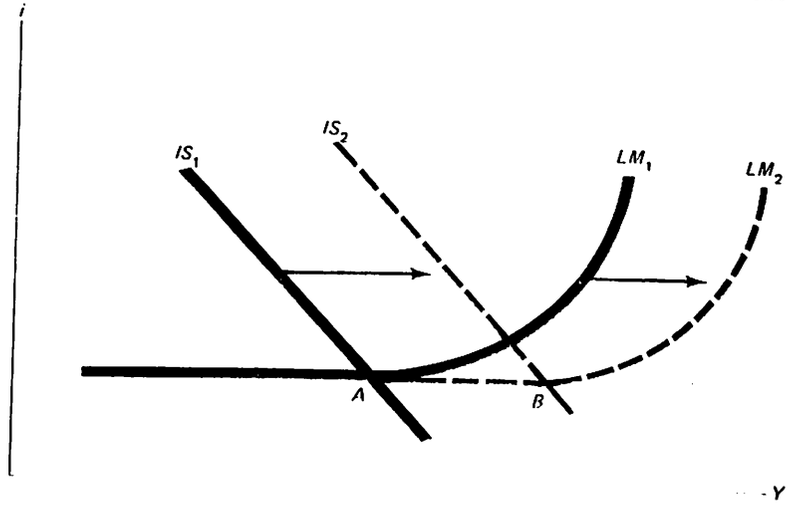


FIGURE 6–1
The real-balance effect: the rightward shift of the IS curve

A_2. The *IS* curve moves from IS_{m_1} to IS_{m_2}, as shown in part B of Figure 6–1.

Once the IS function is permitted to shift in response to changes in the money supply, the "Keynesian" range of the LM function ceases to act as a trap preventing any increase in the money stock from increasing aggregate demand. Rather, an increase in the money stock will cause *both* the LM and the IS functions to move to the right: the LM function because the money supply is used directly in the derivation of this function; and the IS function because of the real-balance effect on the savings function (Fig. 6 2).[1] In the classical theoretical system, wants are

FIGURE 6–2
IS/LM analysis: Effect of increase in money supply with real-balance effect



unlimited, and there is therefore no limit to how far the IS curve can be shifted to the right if there is a sufficient increase in the quantity of money. Unemployment cannot exist in equilibrium if the real money supply is increased enough, either through an increase in the nominal quantity of money or through an increase in the real quantity of money as a result of a fall in prices. Classical economists have a powerful theoretical rebuttal to Keynes' demonstration of an unemployment equilibrium.

From the viewpoint of practical real-world policies, the IS/LM analysis can no longer be used to conclude that monetary policy becomes ineffective during a depression, when interest rates are at or close to

[1] On the rightward movement of the IS curve as a result of the real-balance effect, see Don Patinkin, "Rejoinder to J. R. Hicks," *Economic Journal*, September 1959.

their minimums. (The reader is warned against jumping to the opposite conclusion, that monetary policy *is* effective in such circumstances. This is a complicated question, which will be deferred until Chapter 8.)

Say's Law and the inconsistency in classical theory[5]

Keynes was a brilliant economist, steeped in the classical tradition which he attacked in his *General Theory*. It is therefore somewhat paradoxical—and, indeed, may at first glance seem downright astounding—that he should look at only one half of the classical mechanism when he launched his attack on the classical proposition that equilibrium could exist only at full employment. (He attacked the classical argument that the rate of interest would change so as to equate full-employment savings and desired investment (Fig. 5–1), but he ignored the classical argument that changes in the real quantity of money would affect savings.)[6] Keynes' omission of the real-balance effect becomes more understandable, however, when it is recognized that the classical economists were not consistent regarding the real-balance effect and, indeed, held views which were contradictory.

The difficulty apparently grew out of a natural confusion between propositions which were true *only in equilibrium*, and those which were *invariably* true whether the economy was in equilibrium or not.

Classical economists looked on money as a veil, behind which real goods and services were *ultimately* exchanged for other real goods and services. The butcher, the baker, the candlestick maker, it is true, sold their goods for money in the first instance, but the reason they did so was in order to have the money to buy the shoes and sealing wax which they wanted. Money, of course, was important in oiling the wheels of commerce, making it possible to sell in large batches and to engage in complex transactions involving many individuals which would have been

[5] This section may be skipped without loss of continuity.

[6] Actually, it is not *precisely* correct to say that Keynes completely ignored the classical argument that an increase in the quantity of money can ensure full employment. On page 235 of the *General Theory*, he explicitly attributed unemployment to an inadequate supply of money:

> Unemployment develops, that is to say, because people want the moon;—men cannot be employed when the object of desire (i.e. money) is something which cannot be produced and the demand for which cannot be readily choked off. There is no remedy but to persuade the public that green cheese is practically the same thing and to have a green cheese factory (i.e. a central bank) under public control.

This "green cheese" quotation is difficult to reconcile with the rest of the *General Theory* and, in particular, with Keynes' argument that the economy might reach an equilibrium at less than full employment.

cumbersome or impossible in a barter economy. Money made the system work smoothly, but it was not what the game was all about: People were selling goods in order *ultimately* to get other goods.

In careless hands, this line of argument was extended until it got some classical economists into trouble. In particular, J. B. Say (who became one of Keynes' favorite whipping boys) argued along the following lines in the early part of the 19th century. People sell goods to get other goods. (Note that the word "ultimately" has been dropped from this statement; this is important, as will be seen shortly.) Therefore, the supply of one good involves the demand for some other good. Therefore, for the economy as a whole, the supply of all goods must be equal to the demand for all goods. *Supply creates its own demand,* and there can *never* be a general oversupply of goods. It is true that there may be an oversupply, say, of meat, and therefore distress in the meat industry. But what this meant, according to Say, is that there was a corresponding excess demand for some other product—bread, shoes, or whatever.

This conclusion, *Say's Law*, may be put formally, thus:

$$\sum_{i=1}^{n} S_i \equiv \sum_{i=1}^{n} D_i \qquad (6\text{–}1)$$

where S and D stand for supply and demand, respectively;

there are n goods in the economy, 1, 2, . . . n. (Incidentally, services are included throughout this argument, and the word "goods" should be taken as meaning "goods and services"); and

$\equiv$ means "is always equal to."

The important point of identity (6–1): Say was arguing that the supply of all goods was of necessity always equal to the demand for all goods, *regardless of the general price level.*

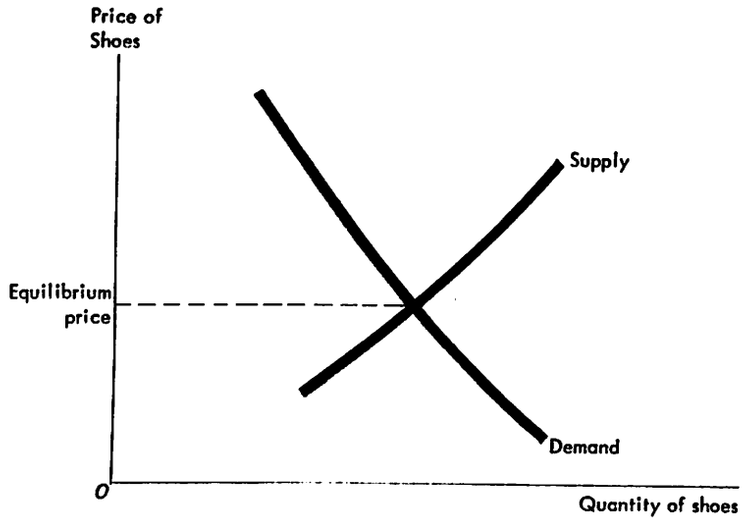
As implied above, this formulation leads to trouble. In order to explain why this is so, it is necessary to digress briefly to the work of 19th century economist Leon Walras. In looking at individual markets, Walras noted that there were two sides—in the ordinary transaction involving money, there was an offer (supply) of money in exchange for goods. In other words, demand does indeed involve supply, and therefore the total demand and supply in the economy must invariably be equal *if the demand and supply of money are included in total demand and supply.* Formally, where money is identified as the $n+1$*st* item, *Walras' Law* may be written:

$$\sum_{i=1}^{n+1} S_i \equiv \sum_{i=1}^{n+1} D_i \qquad (6\text{–}2)$$

Walras Law is correct; there are indeed two sides to every transaction, and therefore total demand is indeed equal to total supply, provided that the demand and supply of money are included.[7]

However, the demand and supply for *each* item need not invariably be equal. If the price is higher than the equilibrium level for a particular good, then the quantity demanded will be less than the quantity supplied. Demand and supply for a specific good are equal, but only at the equilibrium price. This may be put in the familiar diagram of microeconomic textbooks (Fig. 6–3).

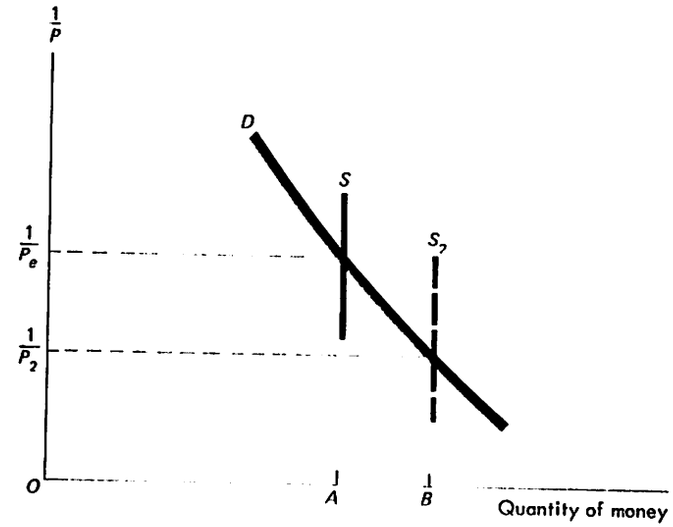
FIGURE 6–3
Demand and supply: An individual good



Now, according to the Walrasian argument, exactly the same type of proposition applies to the $n + 1st$ item, money. If the supply of money is taken as given, then the demand and supply of money look like the functions in Figure 6–4. The "price" or value of money is equal to its purchasing power: The higher the general price index, the lower the value or "price" of money in terms of goods. Thus, the reciprocal of the general price index, $1/p$, is put on the vertical axis when illustrating the demand and supply curves for money. Because the *reciprocal* of the

[7] For simplicity, this exposition is in terms of goods and money. To be complete, it would have to include goods, services, money, and *other financial assets*. For an extensive elaboration of the Walrasian model to include other financial assets, see Don Patinkin, *Money, Interest, and Prices*, 2d. ed. (New York: Harper and Row, 1965).

FIGURE 6–4
Demand and supply of money



price index appears on the vertical axis, *lower* points are observed as prices *rise*.

The formulation of Figure 6–4 is consistent with the quantity theory of money. The demand and supply of money are equal, but only at the equilibrium general price level. If prices are above this level (that is, if prices are at p_2, giving an observation *below* the equilibrium height of $1/p_e$ in Figure 6–4), then the demand for money exceeds the supply; or, put another way, the supply of goods exceeds their demand, and there will be unemployment. The indicated solutions: Increase the quantity of money (to S_2), or allow the general price level to fall to its equilibrium level (p_e). These, of course, represent the standard classical responses to unemployment.

But let us return to Say's Law, and the problem which it raises. Say's Law implies that the demand and supply of money are *invariably* equal, *regardless of the general price level or of the existing quantity of money.*[8] This may be seen by subtracting Say's Law (eq. 6–1) from the correct formulation, namely, Walras' Law (eq. 6–2). This gives:

$$S_m \equiv D_m \tag{6–3}$$

where the subscript m stands for money.

[8] This was pointed out explicitly by Oscar Lange, "Say's Law: A Restatement and Criticism," in Lange, Francis McIntyre, and Theodore O. Yntema, eds., *Studies in Mathematical Economics and Econometrics* (Chicago: University of Chicago Press, 1942), pp. 52–53.

The problem which this raises: It is inconsistent with the quantity theory which states that, if we begin at an equilibrium with full employment and with equilibrium prices (p_e in Figure 6–4), and then we increase the quantity of money (from S to S_2), there will be an *excess supply* of money at the existing price level (p_e), and when consumers attempt to spend this money, the general price level will be bid up to its new equilibrium (p_2). With the quantity of money at S_2 and with prices at p_e, there will be an excess supply of money: The demand for money (OA) will be less than the supply (OB). So states the quantity theory. In contrast, with Say's Law (and consequently with identity 6–3), there can be no excess demand or supply of money; the quantity theory cannot apply.

Say's Law thus introduces a dichotomy into classical theory, a dichotomy between equilibrium in the real goods sector and what is happening in the monetary sector. Say's writings, therefore, blurred the logic behind the real-balance effect; Say's Law by implication denied the effect of the money supply on the demand for goods and services. It is not surprising that, in attacking Say's Law, Keynes propounded a theory which ignored the real-balance effect. The division between the goods sector and the financial sector was continued into the IS/LM analysis presented in the last chapter (but was amended in the IS/LM presentation in Figure 6–2). This dichotomy is, however, inconsistent with the quantity theory and with the real-balance effect; in order to be sure of full employment in equilibrium, a classical economist must argue that an increase in the supply of money will affect the aggregate demand for goods. He must argue that there is a demand and supply for money which are *not necessarily* equal, but which have the general characteristics shown in Figure 6–4, and are equal only when the price level is at equilibrium.[9]

It was at first glance surprising that Keynes focused only on one half of the classical mechanism, and ignored the real-balance effect in putting forward his case that unemployment might exist in equilibrium. The explanation, as we have seen, was that classical economists, and particularly J. B. Say, had laid the basis for Keynes' theoretical approach by splitting the theory of the demand and supply of goods and services from the theory of the demand and supply of money. It may seem even more astounding that classical economists should create this split, and at times ignore the effect of the real quantity of money when investigating

[9] Because a quantity theorist must reject the Say-Keynes dichotomy between the real and monetary sector, and because he must look at the (microeconomic) demand and supply of money, modern classical economists object to the sharp separation between macroeconomic theory and microeconomic theory. It is also why Patinkin picked the subtitle he did for his *Money, Interest, and Prices*, namely: *An Integration of Monetary and Value Theory.*

the markets for goods and services. After all, the quantity theory of money—which lay at the heart of their analysis of aggregate demand—required that the money supply affect the aggregate demand for goods and services. But once again there is an answer; the classical economists were not simply stupid. Rather, they were frequently concerned with a problem quite different from the unemployment problem addressed by Keynes. Specifically, they were concerned with government policies which restricted international trade with tariffs and other devices. In part, trade restrictions were based on a crude mercantilist proposition that the wealth of nations depended on their holdings of gold, and that, therefore, trade restrictions were desirable as a means for increasing the quantity of gold in a country. In arguing against this crude mercantilist thesis, Say and many other classical economists[10] insisted that the wealth of nations lay, not in their holdings of precious metal, but in their ability to produce goods and services to satisfy the wants of their people. In so doing, some of them tended to simplify—a procedure fully justified by the very elementary level of much of the tariff debate. In the process of simplification, they argued that money did not represent real wealth—an accurate statement at a high level of simplification. The fine points of monetary theory got lost; a dichotomy was put forward which laid the groundwork for Keynes' skipping over the real-balance effect in his demonstration of an unemployment equilibrium.

THE MULTIPLIER AND THE REAL-BALANCE EFFECT

The argument of this chapter thus far has been highly theoretical and abstract, dealing with the nature of equilibrium in the economy. As yet, it seems to have little practical application, although it does point to one very important conclusion, namely that Keynesian theory may lead us to an overly hasty rejection of the importance of monetary policy. (See the discussion of Figure 6–2.) It is important that we get back to something with a clearer relationship to the real world. To this we turn: a closer look at the argument behind the Keynesian multiplier.

In the third chapter, a number of alternative derivations of the Keynesian multiplier were given with the use of diagrams, algebra, and a period analysis. In this section, attention will be focused on the period analysis. It is the most useful for going through the blow-by-blow causal argument behind the multiplier theory, and for drawing out the be-

[10] Including the most famous of them all, Adam Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations*, 1776. Available in the Modern Library Series, Edwin Cannan, ed. (New York: Random House, 1937).

havioral assumptions. In going through the step-by-step process, we will throw into question the "normal psychological law" which Keynes saw determining consumer behavior, and which was the basis of the multiplier theory.[11]

Suppose we look more closely at the multiplier model. Suppose, further, that we introduce an initial spending in the simplest possible way—the government engages in a one-time expenditure of $100 million for road building, acquiring the money by borrowing from the central bank. (That is, the road building is financed with newly printed money.) Then, with an MPC of 0.8, the standard illustration of the simple multiplier is shown in Table 6–1. The questions which arise are these: Why do

TABLE 6–1
The multiplier: Single government expenditure

Time period	*Assumed government spending G*	*Effects on spending for:*		*Total*
		C	*I*	
1	100	0	0	100
2	0	80	0	80
3	0	64	0	64
4	0	51.2	0	51.2
. . . . .	.	.	.	.
. . . . .	.	.	.	.
. . . . .	.	.	.	.
n	0	$100 \times 0.8^{n-1}$	0	$100 \times 0.8^{n-1}$
. . . . .	.	.	.	.
. . . . .	.	.	.	.
. . . . .	.	.	.	.
$\rightarrow\infty$	0	$\rightarrow 0$	0	$\rightarrow 0$
Sum	100	400	0	500

income earners spend only $80 million of their additional $100 million in period 2? What do they do with the remainder, and why? The first answer, of course, is that the remaining $20 million is saved. In what form? As currency, bank deposits, or bonds. (With money created in the first round, *someone* will be holding additional money. The individual saver need not, however, hold his savings in money; he can exchange money for bonds with some other member of the economy.) Good enough; why is it saved? To have something on hand for a rainy day (for contingencies, to buy a large-ticket item, etc.). Fine. When is the rainy day (when do the unforeseen contingencies arise, etc.)?

At this point, the logic of the simple multiplier becomes clear; the savings of period 2 are assumed *never* to influence future consumption

at all. What happens in this simple multiplier analysis is that the $20 million of monetary wealth created in the first round but not actually spent in the second round ceases to have any affect on the level of aggregate demand; similarly, the $16 million saved during the third period is assumed to have no further effect on aggregate demand; and so on. To put it glibly, what happens in the Keynesian system is that, when the period ends, a whistle blows, a trap door opens, and through it fall all wealth accumulations from the savings process;[12] they are assumed never to influence consumers' behavior again.

It is difficult to find people who say that their own consumption is utterly unaffected by past savings. The questions raised in this section, therefore, throw doubts on whether the standard multiplier analysis is based on a reasonable "normal psychological law." On the contrary, it seems reasonable to argue that consumption depends in part on wealth, and that, in other words, past savings will tend to leak back into the spending stream.

It is, however, very important to recognize what this criticism does and does not mean. Most important, it does *not* mean that no multiplier-type process is at work. Rather, it means that, as accumulated past savings (wealth) stimulate consumption, then the multiplier process is stronger than indicated by the simple Keynesian analysis. Any tendency of consumers to spend their savings of period 2 during period 3, 5, 10, or whenever will increase the total level of aggregate spending ultimately resulting from the initial government spending. This illustration has suggested that the inclusion of the effect of wealth on the level of consumption requires an *upward* revision of the multiplier.

The event which began the multiplier process—a single-shot government expenditure financed with the creation of new money in our illustration—involved two simultaneous changes in the economy. It involved a flow of income to the public, in the form of additional government expenditures. And, as a part of the same process, the money wealth of the public was increased. The multiplier analysis focuses on the first of these changes, an income increase whose effect tends to peter out with the passage of time. But, if attention is focused on the second change—the increase in money wealth—then there is no similar reason to expect a vanishing effect; the money, once created, stays in the economic system to influence spending behavior indefinitely into the future. In Keynesian theory, attention is focused on the change in income; a single injection of spending is seen to peter out over time, returning the economy to its initial equilibrium (Fig. 3–8). The classical quantity theory, however,

[12] Or, in the words of my colleague Lloyd Atkinson, "A penny saved is a penny burned."

focuses on the increase in the quantity of money. A *single* increase in the money stock will influence the level of aggregate demand indefinitely into the future; a single injection of money into the system will *permanently* increase the level of aggregate demand; a single injection will increase the *equilibrium* level of aggregate demand.

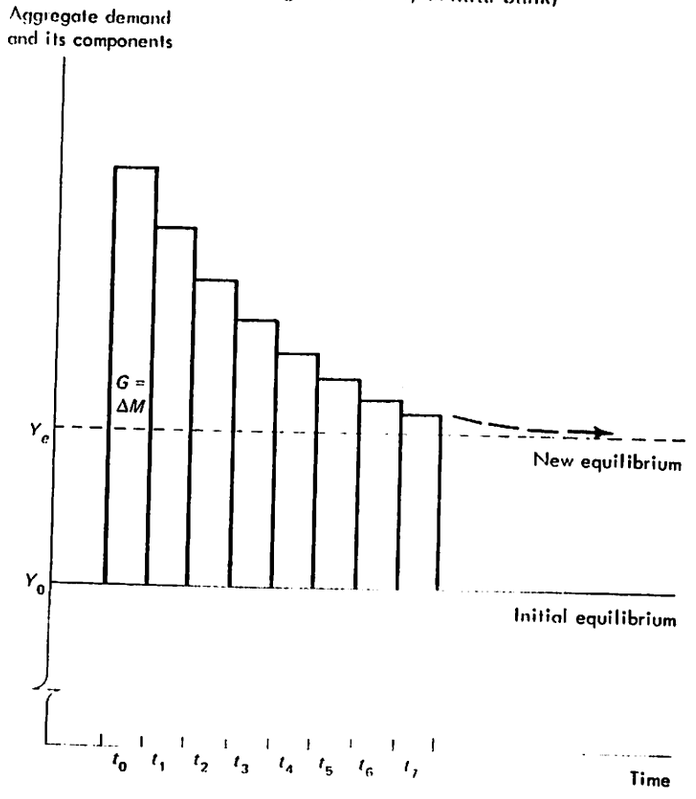
This illustration was chosen in order to draw the line most sharply between Keynesian and classical theory; the government expenditures stressed by Keynesian theory were financed by the money balances stressed in classical theory. What we have done is introduce real money into the system in a much more realistic manner than the price-wage reduction method which dominated the debate between Keynes and the classical economists. In doing so, we have put the real balance effect into a context in which it has a relevance of real-world issues: Will a *single* program of government spending financed with newly created money lead to a *permanent* change in the level of demand? Our answer: Yes.

The above illustration—with government spending financed with newly created money—drew the line most sharply between multiplier analysis and classical thinking. The same general question can, however, also be raised for other types of injections into the spending stream which initiate a multiplier process. Consider a government expenditure financed by selling bonds to the public. As a result of the expenditures, the total wealth holdings of the public will be permanently increased: They will now hold the government bonds in addition to their previous assets. As a result, future consumption should be expected to rise. Or consider investment expenditures. As investment spending proceeds, the real capital stock of the nation will rise. Someone owns that capital stock. Society as a whole owns greater wealth. A plausible result: a stimulus to consumption into the indefinite future.

Wealth and the size of the multiplier: A classical interpretation

Suppose that we return to the initial example, of a government spending financed with newly created money. According to classical economists, the real money balances of the public should influence their consumption expenditure. As a result of the money created to finance the single-shot government expenditure, the money supply is permanently higher than it otherwise would have been. As a result, the level of demand is permanently higher than the initial level. Thus, the effects of the government expenditure (originally shown in Fig. 3–8) must be modified, as illustrated in Figure 6–5. The height of the equilibrium

FIGURE 6–5
The multiplier process: Period analysis with real-balance effect (single government spending financed by central bank)



change in demand will be related to the velocity of money. If, for example, the time period shown in Figure 6–5 is one month, and the equilibrium annual income velocity of money is four, then the equilibrium monthly change in aggregate demand will be one third the change in the money stock. (One third is found by dividing the annual velocity, 4, by the number of months in the year, 12.) Thus, with an initial government expenditure of $100 million financed with the creation of new money, equilibrium monthly aggregate demand will rise by $33.3 million.[13] If one follows a rather rigid quantity theory of money, with velocity being highly stable, then the equilibrium change in aggregate demand will be quickly approached.

For more sophisticated classical treatments, the possibility of a change in the velocity of money must be considered. This complication arises

[13] In this very simple illustration, the possible multiple increase in the money supply as a result of the initial money creation has been ignored. The multiple expansion of the money supply is investigated in Chapter 7.

if the initial government spending is financed with the issue of bonds. In this case, the wealth of the public also increases with the initial government spending, but the money stock does not. As consumers' wealth has permanently increased, their consumption (as a function of current income) should be permanently higher than it would have been in the absence of the government spending. Again, the income increase attributable to the initial government spending will not completely evaporate with the passage of time; there will be a higher equilibrium level of income—although the increase in this case will be less than in the event of an initial money creation.[14]

With the initial injection permanently increasing the level of demand (Fig. 6–5), the size of the multiplier calculated in the standard way—as a series summed over an infinite number of periods—will become infinite; infinity times any minimum constant equals infinity. The question of the size of the multiplier in the classical system therefore becomes uninteresting, and attention in the classical theoretical system is rather focused on the *amount* by which the equilibrium level of income changes (that is, the distance between Y_0 and Y_e in Figure 6–5). For public policy, the relevant question is the change in aggregate demand over some specified finite time period—such as three years—as a result of an initial injection into the system.

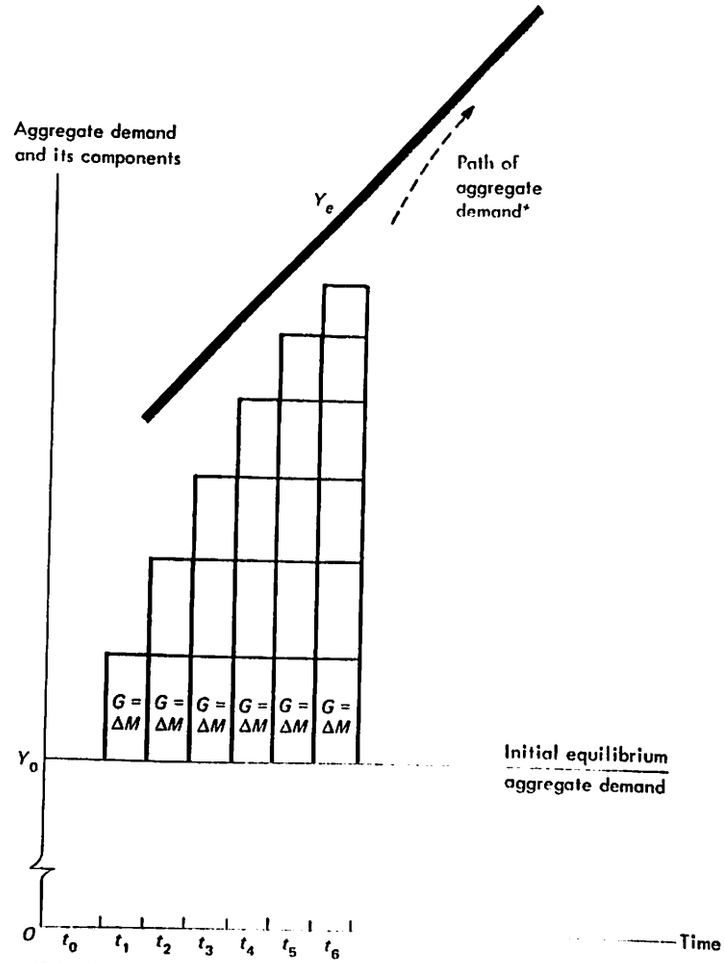
The previous illustrations assumed a *single* government expenditure. Assume, alternatively, a *continuing* government expenditure financed with continuing money creation. Then, if consumption is a function of wealth, aggregate demand will continue to grow without limit, as shown in Figure 6–6.[15] (Compare to Figure 3–9.) This is, of course, what one would expect from a simple application of the quantity theory: If the money supply increases indefinitely, so will aggregate demand. Such a continuing injection will not lead to any new, stable "equilibrium" level of aggregate demand, although there will be an equilibrium *growth path* of aggregate demand.

Following the assumptions of the simple multiplier discussion of Chapter 3, it has been assumed throughout this analysis that there are unem-

[14] With consumption spending higher, there will be a greater demand for the given stock of money balances. Interest rates as a consequence will be higher, with the consumption spending increase therefore being partially offset with pressures on investment. With interest rates higher, the average money balances which individuals and corporations wish to hold at any level of income will be reduced; the velocity of money will rise. (This rise is obviously necessary if income is to rise with a fixed stock of money.) The relationship between the demand for money and interest rates is investigated in Chapter 8.

[15] Figure 6-6 is similar to the curve illustrating the wealth effect in Franco Modigliani, "Monetary Policy and Consumption," *Consumer Spending and Monetary Policy: The Linkages* (Federal Reserve Bank of Boston, Conference Series No. 5, 1971), p. 28.

FIGURE 6–6
The multiplier process: Continuing government spending financed by central bank



* Note: Aggregate demand asymptotically approaches an equilibrium growth path (with slope equal to ΔM times equilibrium income velocity during the time period, in the simple case where equilibrium income velocity is a constant).

ployed resources in the economy, and that real output changes in response to changes in demand, with prices stable. Where the economy is at full employment, the previous analysis must clearly take that into account: A rise in prices will result if aggregate demand increases (or, more precisely, if aggregate demand increases at a rate faster than the increase in productive capacity of the economy). In such cases, increases in demand financed by money creation need not involve an increase in the *real* value of the total quantity of money; rather, they may simply involve increases in prices.

Stocks and flows and the problem of the Keynesian equilibrium

The real-balance or wealth effect objection to the Keynesian concept of an equilibrium level of demand may be reiterated in slightly different terms, concentrating on the distinction between *stocks* and *flows*. A stock is a quantity of something which exists at a point in time: We can speak of the money stock (the amount of money which exists in the economy); the capital stock (the quantity of equipment and other forms of capital which exist); and so on. Expenditure flows—consumption, investment, and so on—involve a time element; they represent a "quantity per time period." It makes no sense to speak of consumption at noon on January 1. Rather, a time dimension must be specified: We can talk of consumption during 1974; or during the first quarter of that year; or during the month of January. The quantity of water in Lake Erie at noon today is a stock; the amount of water which goes over Niagara Falls per month is a flow.

Keynesian economics concentrates on flows, particularly aggregate demand and its components. The difficulty with this, as seen from the classical viewpoint, is that stocks and flows are interrelated. The flow of water going over Niagara Falls will be affected by the stock of water in Lake Erie. In turn, the flow over Niagara Falls will affect the stock of water in Lake Ontario. In macroeconomics, the flow of savings involves a change in the stock of wealth as time passes.

For a general equilibrium to exist, there must be an equilibrium of both stocks and flows, since dissatisfaction with either can lead to a change in the other. If people are dissatisfied with their total wealth, they may try to build up their wealth by saving more. If they have a high level of wealth, they may consume more and save less; that is, the *stock* of wealth may affect the *flow* of saving, and hence the equilibrium aggregate demand (a flow). This last observation, of course, is a restatement of the real-balance effect. Keynesian economics, charge the classicists, concentrates on the equilibrium of flows without reference to what is happening to stocks; this shows up in the emphasis on income as a determinant of consumption, and a tendency to dismiss the importance of wealth. The Keynesian equilibrium consists of flows. Out of any level of income, people save, adding to their stock of wealth. The implicit assumption in the simple Keynesian system is that this accumulation of wealth does not affect future flows.

(From the very early period, however, Keynesian economists did generally recognize and integrate into their thinking a second stock problem. Specifically, in an economy in which investment is taking place, the stock of capital will be rising, and therefore the full-employment level of production of the economy will also be increasing. This means that, for a continuing condition of full employment, aggregate demand must rise as time passes.)

Whatever its failings, classical economics does handle the stock-flow interconnection in a satisfactory manner. Money is a stock. If it is increased from an initial point of equilibrium, people will have an undesirably large stock, and will respond by spending it. Thus, the aggregate demand flow will adjust to a disequilibrium in a stock.

PROBLEMS WITH THE KEYNESIAN EQUILIBRIUM AND WITH THE MULTIPLIER ANALYSIS:[16] THEIR SIGNIFICANCE

On the basis of the real-balance effect, a strong classical counterattack may be made on the Keynesian theory of an unemployment *equilibrium*, in general, and on the Keynesian concept of an equilibrium multiplier, in particular. Multiplier expressions such as 1/MPS are based on the fundamental assumption that at the end of a period a whistle blows and wealth acquired in the savings process disappears. Since the accumulation of wealth is an inherent part of the savings process, and since the savings process is fundamental to the multiplier, this elimination of wealth from consideration may be viewed as an internal contradiction within the multiplier theory.

To adhere to Keynesian equilibrium theory taken to its ultimate logical consequences, it is necessary to argue by implication that individuals are completely unresponsive to accumulations of real money wealth. A diehard Keynesian believes in fat mattresses: People are willing to squirrel away unlimited quantities of idle money without modifying their consumption patterns. That is the implication of the liquidity trap, which is necessary to demonstrate the theoretical possibility of an unemployment equilibrium with flexible wages and prices, or, what is the same thing in static theoretical terms, with stable prices and increasing quantities of nominal money.

This theoretical attack on the multiplier, showing the logical difficulties of a process extended to an infinite number of periods into the future,

[16] While the criticisms of the multiplier in the earlier section are intimately related to the key theoretical issue between classical and Keynesian economists (namely, the possible existence of an unemployment equilibrium), they by no means represent a comprehensive or complete criticism of the multiplier theory. For other criticisms, see, for example, Gottfried Haberler, "Mr. Keynes' Theory of the 'Multiplier': A Methodological Criticism," in American Economic Association, *Readings in Business Cycle Theory* (Homewood, Ill.: Richard D. Irwin, 1951), pp. 193–202; and Milton Friedman, *Capitalism and Freedom* (Chicago: University of Chicago Press, 1962), Chap. 5. For a policy-oriented exposition of the multiplier, see Council of Economic Advisers, *Annual Report*, January 1963, pp. 45–52.

does not, however, give us much clue as to the practical limitations of Keynesian theory as a guide to policy. If Keynesian theory runs into difficulties in its conclusions regarding the long-run equilibrium, why should this be of concern in establishing short-run policy?

Keynesian economics: The long run and the short

The simplest answer, which will for the moment be conditionally acceptable, is that it does not really make much difference. Keynes was concerned with short-run policy making, and, in particular, with the policies necessary to stabilize the economy at full employment. This concern was nowhere more clearly demonstrated than in his dismissal of long-run equilibrium theorizing by economists: "This *long run* is a misleading guide to current affairs. *In the long run* we are all dead."[17] The elaborate—and questionable—theoretical structure which Keynes presented to demonstrate the existence of an unemployment equilibrium was not necessary to make his case for a policy of short-term fiscal expansion; but it was most helpful in convincing economists steeped in equilibrium theory that action was necessary. The Keynesian recommendations of increases in government spending and tax cuts during depressions do not fall as a result of theoretical problems with the nature of the long-term equilibrium.

In a situation such as the depressed 30s, the objections which were made above to the simple multiplier do not throw doubt on the advisability of fiscal expansion, although they may indicate that, in cases where only mild stimulation is needed to restore full employment, some adjustment should be made in the degree of fiscal expansion in order to take account of feedbacks of savings into the spending stream. The only immediately obvious implication of the above discussion is that it is inadvisable, in a situation of depression or recession, to make sizable *long*-term spending commitments in the belief that they will clearly be required for a long-term achievement of full employment. Keynesian theory, although appearing in the guise of long-term equilibrium, should be applied only to short-term problems. It is no less important for this limitation.

For the development of practical economic policies, two aspects of the Keynesian-classical controversy are important. First, classical theory suggests that consumption should be significantly influenced by wealth, and that wealth effects should therefore be taken into account when estimating the future course of demand. More will be said about wealth

[17] J. M. Keynes, *Monetary Reform* (New York: Harcourt, Brace, 1924), p. 88. (Italics in original.)

effects in detail in Chapter 13; suffice it for the moment to note that wealth effects are not easily identified[18] in spite of their logical importance in the theoretical controversy between Keynes and the classical economists.

Second, the discussion of the classical counterattack casts doubt on the precedence given by Keynesian theory to fiscal policy as contrasted to monetary policy, particularly as a cure for depression. The theoretical Keynesian structure dismisses the importance of money in a manner which is not altogether plausible—particularly with the liquidity trap. There is something intrinsic in the Keynesian intellectual framework which reiterates to the unwary economist that "money is not very important; money is not very important." (Just as, on the other side, the quantity theory suggests to the unwary economist that "money is the

[18] Suppose we wish to statistically investigate the Keynesian consumption function. There is no necessary connection between the "periods" of multiplier theory and the periods for which statistics are gathered. Within the Keynesian framework, it is therefore reasonable to specify consumption as a function of "recent" income—that is, income of this period and the previous period. But it is also reasonable to assume that people are also influenced by habit, and that therefore previous consumption (in period $t-1$) should also be used to explain current consumption (in period t). Thus, the consumption function is readily extended within the Keynesian framework to include the following variables:

$$C_t = f(Y_t, Y_{t-1}, C_{t-1}) \tag{6-4}$$

Now, what is the classical objection to the Keynesian consumption function? It is that Keynesian theory ignores the effect of wealth on consumption. To return to the very simple criticisms of the multiplier presented above, the classical objection to Keynesian theory is that people in future periods should be affected by the savings or the changes in wealth during the present period. In other words, the logic of the classical position is that equation (6-1) must be amended to include S_{t-1}, thus:

$$C_t = f(Y_t, Y_{t-1}, C_{t-1}, S_{t-1}) \tag{6-5}$$

where S_{t-1}, the savings of the previous period, is put in as a measure of the change in wealth.

The problem is that the consumption function has now become a statistical monstrosity. Specifically, from the view of the fitting of statistical functions, it makes no sense to include Y_{t-1}, C_{t-1}, and S_{t-1} all as independent variables in the equation, since they are not only closely related, but S_{t-1} may be *directly* derived from the *definitional* relationship among the three variables:

$$S_{t-1} \equiv Y_{t-1} - C_{t-1} \tag{6-6}$$

We get no statistical mileage by putting S_{t-1} into equation (6-5); the central *theoretical* issue between Keynes and the classics is not amenable to such a simple statistical test.

This does not mean, however, that wealth concepts are irrelevant in the empirical investigation of the consumption function. Something will be said on this subject in Chapter 13. For the moment, it suffices to note that, when wealth is introduced statistically into the consumption function, measures *not* directly related to savings of the previous period (as commonly defined) are used. For example, in the econometric model of the Federal Reserve Board and the Massachusetts Institute of Technology (the FR-MIT model), capital gains accrued in the stock market are included as an explanatory variable in the consumption function.

key; money is the key.") If a judgment is to be made about the *relative* merits of fiscal and monetary policy, a much closer look will have to be taken at the manner in which money affects the economy. In the beauty contest between fiscal and monetary policy, the prize cannot be awarded after looking only at the first contestant. In the coming chapters, the second contestant—monetary policy—will be viewed.

KEY POINTS

1. According to classical theory, the dichotomy between the product market (the IS curve) and the financial market (the LM curve) must be rejected. An increase in the quantity of money held by the public will increase their level of consumption out of any given amount of income. Thus, it will shift the savings function down, and shift the IS curve to the right. An increase in the quantity of money directly affects *both* the LM and IS curves.
2. Consequently, the "liquidity trap" argument becomes invalid if consumption responds to additions in real money holdings of the public (the "real balance effect"). Increases in the real money stock will increase aggregate demand. The economy does not remain in an unemployment equilibrium regardless of increases in real money balances.
3. Once we recognize the potential effect of increases in the quantity of money (or other forms of wealth) on consumption, then the multiplier process becomes more complicated. For example, a one-shot government spending financed by newly created money will involve a *permanent* increase in the equilibrium level of national income; the effects do not peter out asymptotically towards zero as foreseen in the simple multiplier theory.

QUESTION

1. In Chap. 3, consumption was introduced as a function of income. We have now introduced a second determinant of consumption demand; namely, money wealth (and other forms of wealth). What other variables might be expected to influence consumption demand?

SUGGESTED READING

Don Patinkin, *Money, Interest, and Prices; An Integration of Monetary and Value Theory*, 2d ed. (New York: Harper & Row, 1965), Chaps. 1-4.

part III

Money and the banking system

Highbrow opinion is like a hunted hare; if you stand still long enough, it will come back to the place it started from.

Dennis Robertson

AN OLD KEYNESIAN COUNTERATTACKS

James Tobin
Yale University

THE CENTRAL MACROECONOMIC ISSUE

The crucial issue of macroeconomic theory today is the same as it was sixty years ago when John Maynard Keynes revolted against what he called the "classical" orthodoxy of his day. It is a shame that there are still "schools" of economic doctrine, but perhaps controversies are inevitable when the issues involve policy, politics, and ideology and elude decisive controlled experiments. As a lifelong Keynesian, I am quite dismayed by the prevalence in my profession today, in a particularly virulent form, of the macroeconomic doctrines against which I as a student enlisted in the Keynesian revolution. Their high priests call themselves New Classicals and refer to their explanation of fluctuations in economic activity as Real Business Cycle Theory. I guess "Real" is intended to mean "not monetary" rather than "not false," but maybe both.

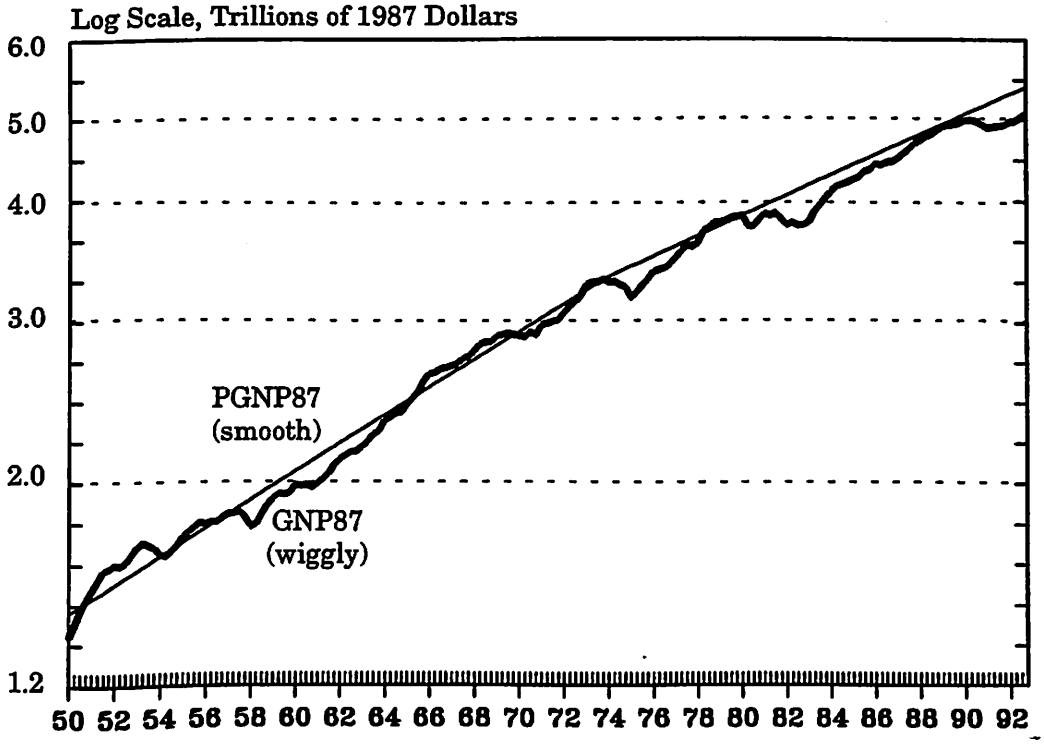
I am going to discuss the issues of theory, Keynesian versus Classical, both then and now. Since the main purpose and preoccupation of macroeconomic theory is to guide fiscal and monetary policies, the theoretical differences imply important differences in policy. Moreover, prevailing doctrines seep gradually into the ways the world is viewed not only by economists but also by students, pundits, politicians, and the general public. It is in this sense but only in this sense that I shall be talking about current events.

The doctrinal differences stand out most clearly in opposing diagnoses of the fluctuations in output and employment to which democratic capitalist societies like our own are subject, and in what remedies, if any, are prescribed. Keynesian theory regards recessions as lapses from full-employment equilibrium, massive economy-wide market failures resulting from shortages of aggregate demand for goods and services and for the labor to produce them. Modern "real business cycle theory" interprets fluctuations as moving equilibrium, individually and socially rational responses to unavoidable exogenous shocks. The Keynesian logic leads its adherents to advocate active fiscal and monetary policies to restore and maintain full employment. From real business cycle models, and other theories in the New Classical spirit, the logical implication is that no policy interventions are necessary or desirable.

Should we describe the macro-economy by two regimes or one? The old Keynesian view favors two regimes. In one, the Keynesian regime, aggregate economic activity is constrained by demand but not by supply. If there were additional effective demands for goods and services, they could be and would be satisfied. "Demand creates its own supply." The necessary inputs of labor, capital capacity, and other factors are available, ready to be employed at prices, wages, and rents that their productivity would earn. Only customers are missing.

The second regime, which Keynes called classical, is supply-constrained. Extra demand could not be satisfied at the economy's existing capacity to produce. The needed workers or other inputs are not available at affordable wages and rents. The supply limits bring about prices and incomes that restrict aggregate demand to capacity output. Should capacity increase, those prices and incomes will automatically generate just

FIGURE 1
Real GNP: Actual and Potential
Quarterly, 1950 - 1992

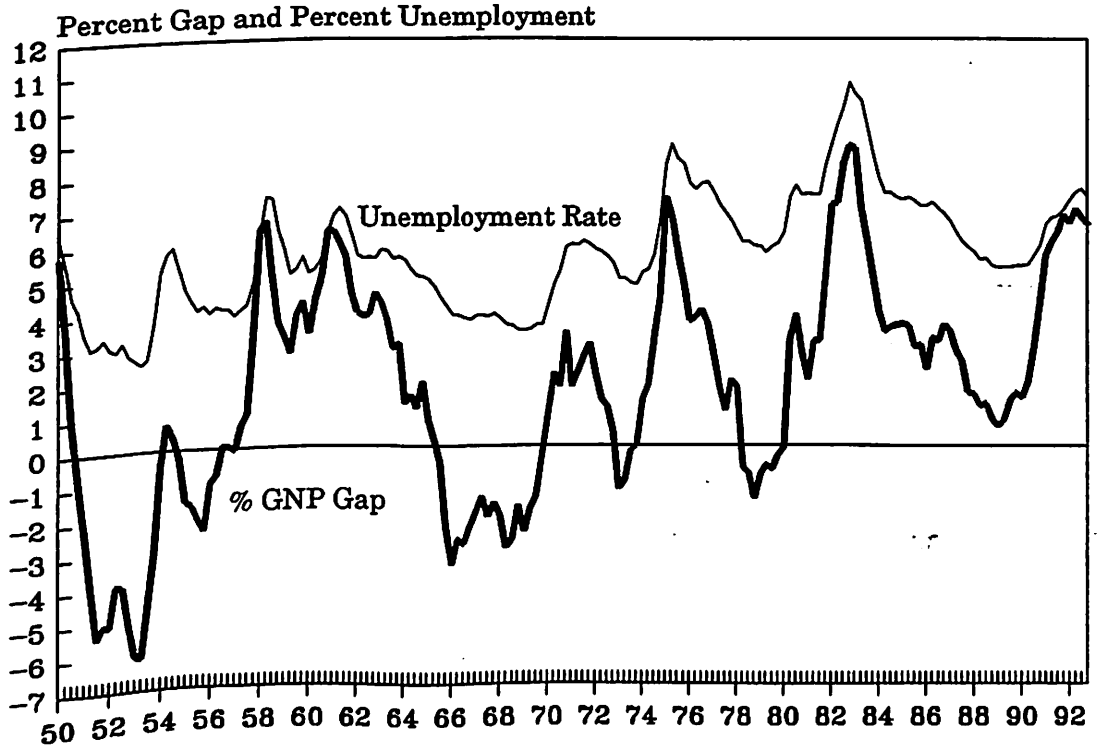


enough additional purchasing power to buy the extra output. "Supply creates its own demand."

New Keynesians believe that the economy is sometimes in one regime, sometimes in the other. New Classicals model the economy as always supply-constrained and in supply-equals-demand equilibrium. In their real business cycle models, the shocks that move economic activity up and down are essentially supply shocks, changes in technology and productivity or in the bounty of nature or in the costs and supplies of imported products. Although external forces of those kinds, for example weather, harvests, natural catastrophes, have been the main sources of fluctuating fortunes for most of human history, and although events continually remind us that they still occur, Keynesians do not agree that they are the main source of fluctuations in business activity in modern capitalist societies.

The distinction between the two views can be concretely illustrated by reference to Figures 1 and 2. Charts of this kind were originated by President Kennedy's Council of Economic Advisers in 1961. They were meant to depict a Keynesian view of the U.S. economy. In Figure 1 the wiggly track is the reported real (i.e., inflation-corrected, measured in 1987 prices) Gross National Product (GNP). The smooth track is Potential GNP (PGNP), a hypothetical estimate of the growing capacity of the economy to produce goods and services. PGNP approximates the supply constraint on GNP. This cannot, of course, be taken literally. "Capacity" means what can be produced by the normal

FIGURE 2
GNP Gap and Unemployment Rate
Quarterly, 1950 - 1992



peacetime operations of a market economy, not what can be done in an emergency mobilization like that of World War II. Sometimes, Figure 1 shows, actual GNP exceeds PGNP. These are situations of unsustainably low unemployment and labor shortage; the economy is overheated and inflation is increasing.

Conceptually PGNP is meant to correspond to full employment, indicated by balance between unemployment and vacancies and by stable rates of change of money wages and prices. In practice, in Figure 1 when GNP coincides with PGNP, the unemployment rates rise gradually from 4 to 5 1/2 percent. The proximate determinants of the growth of PGNP are the growth of employment — which is, since the unemployment rate is held constant, essentially that of the labor force — and the growth of the productivity of labor. Both of these growth rates slowed down around 1973; in Figure 1 the slope of PGNP on logarithmic scale is reduced from 3.5 to 2.5 percent in that year.

The sources of PGNP growth are *supply* phenomena. They are the consequences of demographic and technological trends, which by their very nature change slowly. Actual GNP wanders around PGNP. The Keynesian interpretation of the volatile gap between the two series is that it reflects fluctuations in *demand*. Spending can and does go up and down more quickly than capacity. When actual GNP falls below PGNP, the economy is in the Keynesian demand-constrained regime. When it is above or equal — or even, say, 1 or 2 percent below — PGNP, the economy could be viewed as supply-constrained.

Figure 2 charts the percentage GAP between PGNP and GNP, together with the overall unemployment rate. Clearly the two series go up and down synchronously. However, the amplitude of GAP is much the greater. A one point increase or decrease in the unemployment rate is associated with a 2 1/2 or 3 percent change in the same direction in the GAP. This phenomenon is one of the most important and reliable empirical regularities of macroeconomics. It is known to economists as Okun's Law, because the late Arthur Okun quantified the GAP and its relationship to unemployment for the Council of Economic Advisers to President Kennedy in 1961. The Council wanted to demonstrate to the President and Congress that the economic payoffs of fiscal and monetary stimuli to reduce unemployment went far beyond the direct benefits to the unemployed themselves.

It may seem paradoxical that a one percentage point reduction of unemployment, which might be expected to mean approximately a one percent increase in employment, would raise output by more than one percent, indeed a great deal more. The answer is that the same spending that reduces unemployment rates raises labor inputs to production in other ways: increased hours of work, movement of discouraged workers into the labor force, and more efficient use of overhead workers and of other redundant workers kept on payrolls in hard times.

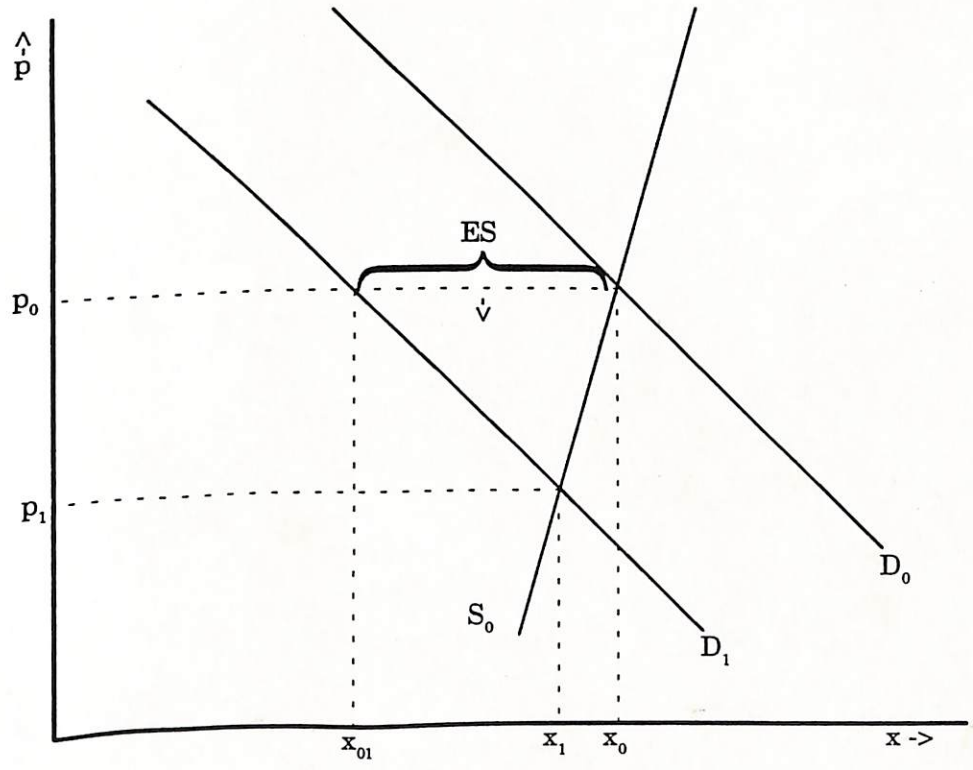
Apostles of New Classical Macroeconomics and Real Business Cycle Theory reject the Keynesian interpretation. For them, there is no PGNP path distinct from actual GNP. The fluctuations of actual GNP are also fluctuations of PGNP, caused by shocks to the economy's productive capacity. One could of course draw a trend, a moving average, through the GNP path. But it would be purely descriptive. It would have no macroeconomic significance. There is only one regime. The economy is always against its supply constraint. It is never demand-constrained in the sense that demand falls short of the normal capacity of a market economy. The economy is continuously at full employment, but the unemployment rate corresponding to full employment fluctuates from one quarter to the next.

Keynesians interpret the quarter-to-quarter and year-to-year fluctuations of unemployment as largely involuntary: workers whose marginal productivities are no less than existing real wages are willing to take such jobs, but the jobs don't exist. New Classicals, in contrast, regard all unemployment as voluntary; workers choose to withdraw from or enter or re-enter the labor force as the advantages of employment change relative to other uses of time. The supply shocks that drive the economy also change those advantages and those choices.

Practical people — forecasters of business conditions, business managers, politicians, workers, even bankers and central bankers — are instinctively Keynesians, especially during recessions. They realize that companies lay off workers and even shut down when their sales fall off. They blame cutbacks in defense for unemployment in Groton, Connecticut, where submarines are built, and blame declines in air travel for hard times in St. Louis and Seattle, where aircraft are made. But the dominant theory in academic macroeconomics today has no room for economy-wide demand shocks and demand-side recessions.

How come? It all has to do with *market-clearing*, specifically the role of prices in clearing markets, that is, in equating demand and supply. The favorite assumption of orthodox economic theory, Classical or Neo-Classical or New Classical, is that the price in any market is determined by the condition that supply equal demand. That is pictured in economists' favorite diagram for beginning students, and it is the unques-

FIGURE 3
Supply, Demand, Market-Clearing



tioned assumption of Ph.Ds. Figure 3 is such a diagram, for a single commodity and its market. If the demand curve is D_0 and the supply curve is S_0, the price is p_0 and the quantity is x_0. Should demand shift to D_1 while supply remains at S_0, price moves to clear the market at p_1, x_1.

Does such a price adjustment occur instantaneously, so that there is no real time during which the markets fail to clear? Is there no real time during which price stays at p_0 and sellers are able to sell only x_{01} even though they would like to sell x_0, so that there is excess supply of *ES*? The arrow pointing downward reflects what we tell our introductory students. If there is excess supply in a market, the price falls. The question is "How fast?" Should we model the whole economy as if all markets, labor markets as well as product markets, are cleared by price adjustments at every moment of time? If so we are altogether ruling out excess demands and excess supplies — in particular, involuntary unemployment — and assuming that all the prices and quantities we observe reflect demand/supply equalities, in other words that no non-price rationing of sales among buyers or sellers occurs. This is the essence of the Keynesian-New Classical dispute.

DEJA VU: THE SAME MACROECONOMIC CONTROVERSY SIXTY YEARS AGO

It's nothing new. The same controversy occurred in the 1930s. It was pretty hard to maintain classical orthodoxy during the Great Depression. But in the absence of any intellectually respectable alternative, the classical supply-constrained, market-clearing model was used by economists in diagnosing, misdiagnosing, the depression and by policy-makers in resisting demand-creating remedies. What came to be known as the "Treasury View" in Britain was echoed in the United States by the Hoover Administration, the Federal Reserve, and initially by the Roosevelt Administration too, and in Germany by the Bruening government, the last government of the Weimar Republic before Hitler.

John Maynard Keynes started revolting against orthodox theories and policies in 1925, when the depression was beginning in Britain. But it was not until he wrote *The General Theory of Employment, Interest and Money* [1936] that he could present a coherent theoretical alternative. The intention of the word "General" in the title was precisely to distinguish his theory from the "classical" supply-constrained market-clearing model. He did so by arguing that economies like those of the U.K., Western Europe, and the U.S. are usually in demand-constrained regimes. In the next 20 or 25 years, the Keynesian Revolution swept the profession and became generally accepted mainstream wisdom. Twenty years later a classical counter-revolution had reopened the debate of the 1930s and put Keynesian economics on the defensive.

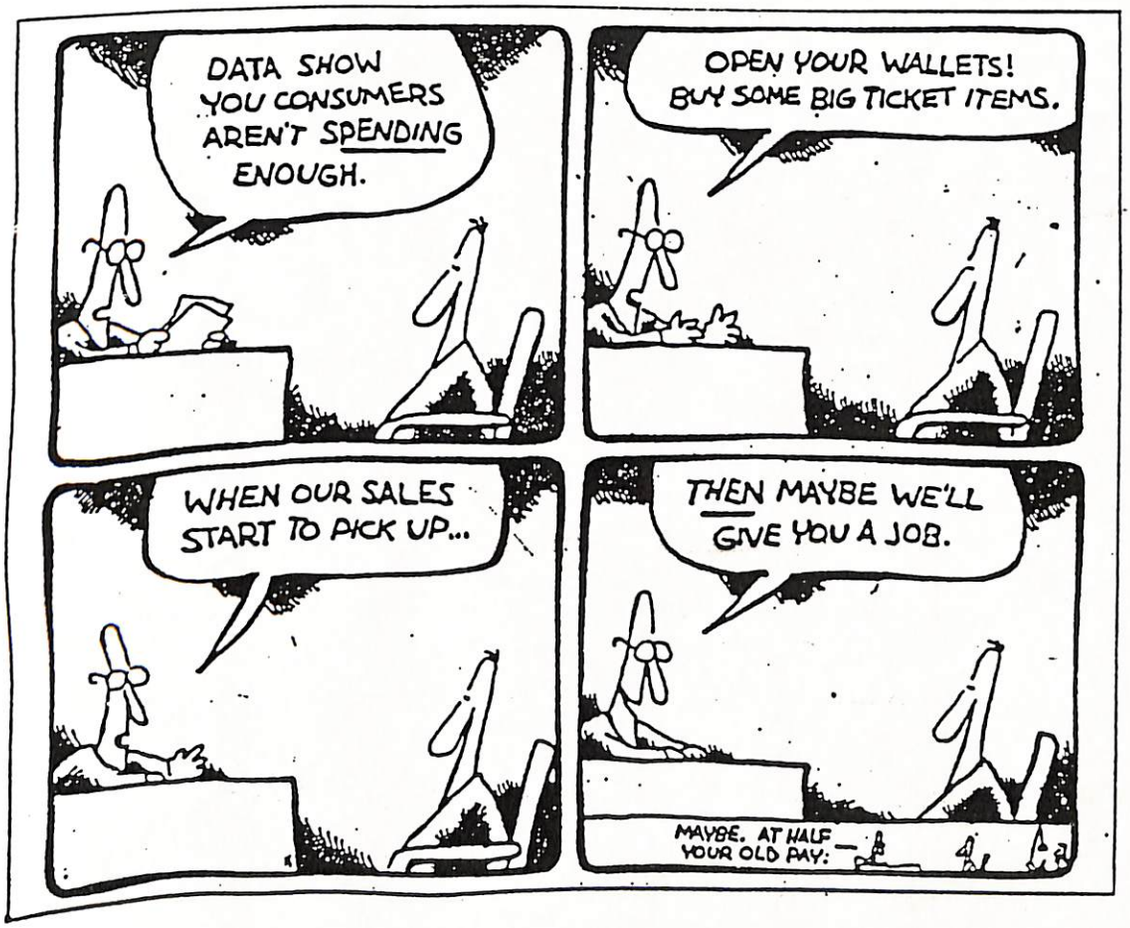
According to the synthesis of classical and Keynesian macroeconomics reached by 1960, Keynesian macroeconomics is short-run. It does not pretend to apply to long-run growth and development. It does not tell poor countries how to lift themselves out of poverty or rich countries how to be richer fifty years hence. In the long run — perhaps with the help of Keynesian policies — markets will somehow clear, new workers will get jobs, and the fruits of technological progress will be realized.

In the 1930s Keynes suspected that involuntary unemployment was not just a transient cyclical phenomenon but a chronic defect of advanced capitalism. In New England's Cambridge, Alvin Hansen [1938] warned of secular stagnation. Those views were natural enough in the 1930s. Both Keynes and Hansen were depicting outcomes to be feared in the absence of the remedial policies and institutions of demand stabilization they were recommending. It can be argued that habitual application of those remedies after World War II, reinforced by the expectation that they would be used, moderated the severity of cyclical departures from the full-employment path.

The most important innovation of the *General Theory*, according to its author, is what he called *the principle of effective demand*. This is his term for the demand constraint I described above. The word "effective" captures the idea that workers can spend on goods and services only the wages they actually earn from employment, not the amounts they would spend if they had all the jobs they would like to have at existing wages. Likewise, employers can hire workers only to the extent they are needed to produce the goods and services they can sell. During the recent recession this impasse was nicely captured by a cartoonist with economic intuition. (Figure 4).

Keynes's "classical" opponents in the 1930s were much more moderate than their descendants today. In the *General Theory,* Keynes's foil was his long-time friend and Cambridge colleague, Professor A. C. Pigou. Neither Pigou nor other orthodox economists of the day were arguing that a model in which prices cleared all markets at every

FIGURE 4
Keynes's Macroeconomic Impasse
The Principle of Effective Demand



Tom Toles
The Buffalo News
October 1991

instant of time was a reliable approximation to actual economies or a practical guide to government policies. The debate was about the efficacy and speed of the economy's natural recuperative mechanisms. If shocks occur that bring about unemployment, will they set in motion corrective adjustments that restore full-employment equilibrium? Specifically, will deflation (or disinflation), the wage and price declines that naturally result from excess supplies (like *ES* in Figure 3), do the job? Will they do it without help from countercyclical fiscal and monetary policies? Keynes said "No, or anyway not always, and if ever, not soon enough." Pigou said "Yes, surely yes, eventually anyway." As a theorist, his main concern was to deny that Keynes's demand-constrained outcomes deserved the status of *equilibria* in the sense that they would repeat themselves

indefinitely in the absence of external shocks. Pigou resented that word General. But as a practical matter he agreed with Keynes on public works spending as a means of reducing unemployment.

In contrast, New Classical theorists today do not allow excess supplies or demands ever to arise in the first place. Thus they finesse the Keynes-Pigou issue, the speed and efficacy of natural adjustment mechanisms in eliminating discrepancies between demand and supply.

THE PRICE FLEXIBILITY CONTROVERSY

What young theorists today describe as Keynesian economics is a caricature of the true thing. Here is a description by authors who, by labeling themselves New Keynesians, evidently intend to convey sympathy.

> According to the Keynesian view, fluctuations in output arise largely from fluctuations in *nominal* aggregate demand. These fluctuations have real effects because nominal wages and prices are rigid...[T]he crucial nominal rigidities were assumed rather than explained, [although] it was clearly in the interests of agents to eliminate the rigidities they were assumed to create...Thus the 1970s saw many economists turn away from Keynesian theories and toward new classical models with flexible wages and prices. [Ball, Mankiw, and Romer 1988, 1, emphasis added]

Those New Classical models are market-clearing models, and they have not just flexible prices but *perfectly* and *instantaneously* flexible prices, an assumption that is surely more extreme, more arbitrary, and more devoid of foundations in individual rational behavior than the imperfect flexibility assumed in Keynesian models. There is a great deal of semantic double-talk in the assertion that the macroeconomic market failures described by Keynesian models vanish if money wages and prices are assumed flexible rather than rigid.

Price flexibility is not a yes-or-no circumstance. Consider instead a spectrum of the degree of price flexibility, from complete flexibility at one extreme to complete rigidity at the other. Complete flexibility means instantaneous adjustment, so that prices are always clearing markets, jumping sufficiently to absorb all demand or supply shocks. Complete rigidity means that nominal prices do not change at all during the period of analysis. In between are various speeds of price adjustment, various lengths of time during which markets are not clearing.

Who owns the middle ground? We Keynesians do, despite common beliefs to the contrary. Keynes and Keynesian economists did not assume complete rigidity, nor did they need to. It is not true that only an arbitrary and gratuitous assumption of complete rigidity, converting nominal demand shocks into real demand shocks, brings into play Keynes's multipliers and other demand-determining processes (including the IS/LM curves taught to generations of college students). Any degree of stickiness that prevents complete instantaneous price adjustment has the same qualitative implications.

In the quotation above, "nominal aggregate demand," means aggregate dollar spending on goods and services. This is not what Keynes meant by "effective demand." He was referring to demands for quantities of goods and services, measured in constant

prices, not in current dollars. He stressed changes in these real demands, not mindless changes in total dollar spending irrespective of what dollars could buy, as the sources of depressions and prosperities. Only people who formed their opinions of Keynesian economics without reading Keynes could make this mistake.

What is true is that Keynes stressed that we live in a monetary economy, as opposed to a frictionless market-clearing barter economy. Prices, including wages and salaries, are quoted in dollars. It is dollar prices that initially respond to excess supplies and demands, not real or relative prices, which value each commodity or service in terms of other commodities. In insisting on this fact, Keynes was deviating from a cherished principle of classical theory, the proposition that "money is a veil" behind which everything works out as it would in a miraculously efficient barter economy. Money is neutral. It affects nominal prices but not real variables. According to this proposition, which Don Patinkin [1956] called "the classical dichotomy," people do not value money for its own sake and therefore they behave in ways that produce the same real outcomes regardless of how much money is circulating. Real prices, the terms of trade between commodities, are the same whether dollar prices are high or low and whether they are inflating, deflating, or stable. Dudley Dillard [1988] called this the "barter illusion" of classical economics.

In any single small market of a large economy, the distinction between money price and real price may be negligible. If a fall in the demand for bagels leads to a decline in their prices in dollars, that is also a decline relative to prices of gasoline, videotapes, plumbers, and everything else. If Figure 3 applies to bagels, we would not have to specify whether the price on the vertical axis is cents per bagel or fractions of a standard shopping-cart package per bagel, and we could assume that the demand and supply curves stay in place as the bagel price moves.

In attacking the classical assumption that markets are continuously cleared by price adjustments, Keynes stressed labor markets in particular, asserting that wages do not move fast enough to avoid excess supplies of labor — involuntary unemployment — at prevailing wages. The difference between money price and real price, negligible for a local bagels market, is crucial for an economy-wide labor market. It is the real wage — the value of wages in goods produced and consumed — that should equate employers' demands for labor with workers' willing supplies. When shocks throw this market out of equilibrium, these real-wage demand and supply schedules may well stay put as wages and other prices adjust. But if in Figure 3 the money wage is the price on the vertical axis, we cannot assume that the demand and supply schedules stay in place as the money wage declines. The demand for labor will certainly depend on the wages that the workers are paid and spend on the products they themselves make, as the intuitive cartoonist-economist understood.

Therefore, if an economy-wide excess supply of labor arises and leads to a fall in money wages throughout the economy, it is by no means obvious that *real* wages fall as much — or at all. Quite possibly, employers just reduce proportionately the dollar prices of the goods they produce. Keynes argued that workers could be quite willing to take jobs at lower real wages but have no way to communicate this willingness.

The question boils down to whether proportionate deflation of all nominal prices, both money wages and product prices, will or will not increase aggregate effective real demand. This is a complicated matter, and I cannot do it justice here. Two issues in this debate need to be distinguished. The first concerns the relation of real aggregate demand to the nominal *price level*. The second concerns its relation to the expected *rate of change* of nominal prices.

Keynes in Book I of *The General Theory* denied that real aggregate demand was related at all to the price and money wage level. In effect, he turned the classical neutrality proposition against the classicals. If all money wages and prices are lowered in the same proportion, how can real quantities demanded be any different? Thus, if real demand is deficient, how can a purely nominal price adjustment undo the damage?

Actually Keynes himself provided an answer in a later chapter. If the nominal quantity of money remains the same, its real quantity increases, interest rates fall, and real demand increases. This mechanism would fail if demand for money became perfectly elastic with respect to interest rates — the "liquidity trap" — or if demand for goods for consumption and investment were perfectly inelastic.

Pigou [1943; 1947], Patinkin [1948], and other authors provided another scenario, the "Pigou effect" or "real balance effect," which alleges a direct positive effect on spending resulting from households' increased wealth, in the case at hand taking the form of the increased real value of their holdings of dollar-denominated assets. This effect does not depend on reduction of interest rates.

To an astonishing degree, the theoretical fraternity has taken the real balance effect to be a conclusive refutation of Keynes. Yet this effect is of dubious strength, and even of uncertain sign. Most nominal assets in a modern economy are "inside" assets, that is the debts of private agents to other private agents. They wash out in accounting aggregation, leaving only the government's nominal debt to the private sector as net wealth. Some, though probably not all, of the interest-bearing debt is internalized by taxpayers who feel poorer because of the taxes they expect they or their heirs to have to pay to finance the interest payments. The base of the real balance effect is therefore quite small relative to the economy. In the United States today the monetary base, the non-interest-bearing federal debt, is only 6 percent of GNP.

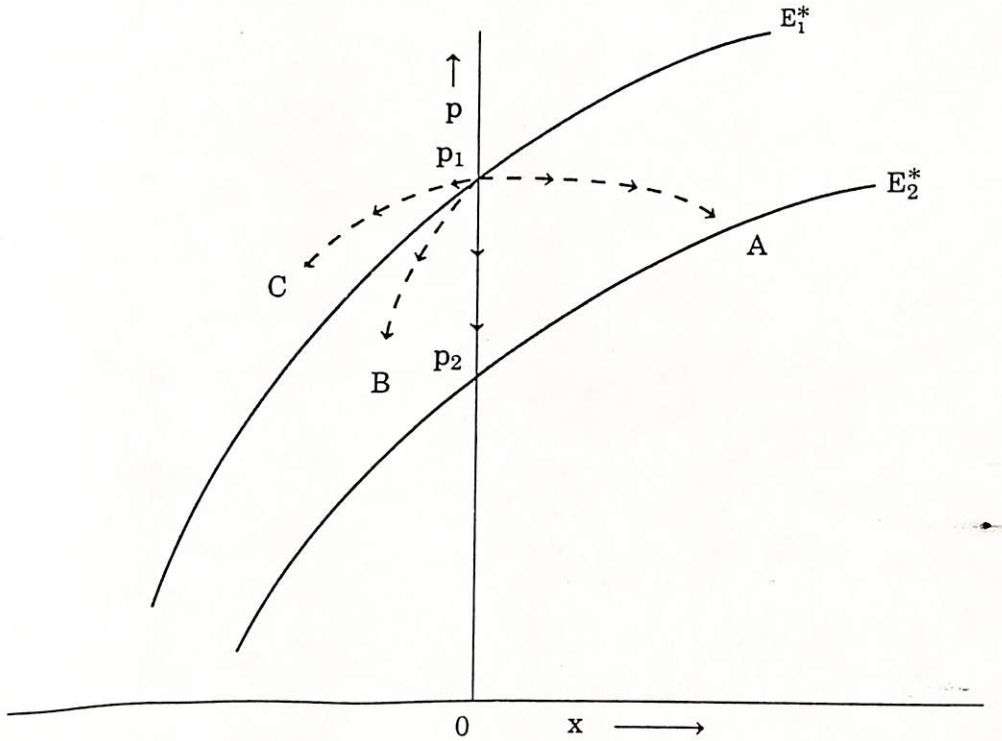
While Don Patinkin [1948] stressed the theoretical importance of the real balance effect, he disclaimed belief in its practical significance. In the Great Depression, he pointed out, the real value of net private balances rose 46 percent from 1929 to 1932, but real national income *fell* 40 percent.

That inside assets and debts wash out in accounting aggregation does not mean that the consequences of price changes on their real values wash out. Price declines make creditors better off and debtors poorer. Their marginal propensities to spend from wealth need not be the same. Common sense suggests that debtors have the higher spending propensities — that is why they are in debt! Even a small differential could easily swamp the Pigou effect — gross dollar-denominated assets are 200 percent of United States GNP.

Irving Fisher [1933] emphasized the increased burden of debt resulting from unanticipated deflation as a major factor in depressions in general and in the Great Depression in particular. Fisher's wealth redistribution effect is quite possibly stronger than the Pigou and Keynes effects combined, particularly when output and employment are low relative to capacity. This may be one reason for the weakness of demand in world economies the past four years.

An even more important argument refers to *rates of change* of nominal prices. The process of change works on aggregate demand in just the wrong direction. Greater expected deflation, or expected disinflation, makes people want to hold money rather than buy goods. It is an increase in the real rate of interest, necessarily so when nominal interest rates are constrained by the zero floor of the interest on money. This is another

FIGURE 5
Aggregate Demand Related to Price Level and Price Change
The Questionable Stability of Price Adjustment



factor Fisher stressed in his explanation of the Great Depression. Keynes stressed it too, as a pragmatic reinforcement of his overall argument.

The process of price change matters when the change takes place in real time, because during the transition it tends to move the demand/supply balance in the wrong direction. After a negative demand shock, an increase in demand associated with a lower price level is required to restore equilibrium; a falling price actually diminishes demand.

Not surprisingly, the New Classicals, and evidently the self-styled New Keynesians too, take the easy way out. The possible instability of the price-adjustment process is an embarrassment. They tacitly avoid it by assuming perfect flexibility, so that after surprise shocks, prices jump to their new equilibria without passage of time.

The problematic stability of real-time price adjustment is evident in Figure 5. Here the horizontal axis represents expected price deflation or inflation, x. The vertical axis represents p the log of the price level. An upward sloping curve like E_1^* plots combinations (x,p) of expected price change and price level that generate the same aggregate real demand E. The slope reflects the assumptions that demand is related negatively to the price level and positively to its expected rate of change. In given circumstances, a higher curve refers to a lower demand E and a lower curve to higher demand. The curvature of the E^* loci reflects the assumption that the "Keynes effect" of increases in real money balances in lowering interest rates declines as those balances rise and interest rates fall.

Suppose that initially the "isoquant" E_1^* makes demand equal to full-employment equilibrium output Y_1^*, here taken to be constant. Points above or left of that isoquant are positions where E is lower than Y_1^*, characterized by Keynesian unemployment. Points below or right of E^*_1 are positions of macroeconomic excess demand. In Figure 5, the equilibrium inflation rate (expected and actual) and price are $(0, p_1)$. Suppose now that a discrete one-time negative shock to real demand shifts the isoquant for $E = Y^*$ down to E^*_2 so that the new equilibrium inflation rate and price are $(0, p_2)$. The old isoquant E^*_1 now implies an E lower than Y^*. To restore equilibrium the price level must fall from p_1 to p_2. How is the price decline to be accomplished? One scenario is the New Classical miracle, an instantaneous precipitous vertical descent, so that there is no time interval during which actual or expected price changes are other than zero. If jumps of that kind in p are excluded, there is no path of actual price changes and rationally expected prices that avoids departure from $E = Y^*$ during the transition. It would take a burst of positive inflation, actual and expected, to offset the negative demand shock, as at point **A**. But this would move the price level in the wrong direction.

The likely scenario is a path like **B** or **C** in Figure 5: The excess supply that now characterizes the initial equilibrium point $(0, p_1)$ and the first isoquant induces prices to decline, and the anticipation of their decline is bad for aggregate demand. Along **B** the real balance effect is strong enough to overcome the negative effects of the deflation; aggregate demand E is increasing as the path hits lower isoquants. The new equilibrium may be attained, though probably by a damped cyclical process. Along **C**, however, the price level effect is too weak to win out, and the gap of E and Y below Y^* is increasing.

Fisher and Keynes both thought that output and employment would be less volatile if money wages and prices were fairly stable, rather than flexible. They were right. Earlier, [Tobin, 1975] I exhibited a simple formal macroeconomic system, classical in the sense that it has only one equilibrium, which is characterized by full employment and a constant price level. It is easy to specify plausible dynamics that make the equilibrium unstable because the price-change effects outweigh the price-level effects. Moreover, the system could be stable locally but unstable for large displacements.

The question whether price flexibility (in any sense short of the perfect-flexibility fairy tale) is stabilizing has begun to receive considerable attention. Delong and Summers [1986] have investigated this question using the Fischer-Taylor staggered-contract model [Fischer, 1977; Taylor, 1980], amended to allow both price-level and price-change effects on demand. Their most interesting simulation has the intuitively desirable property that close to the limit of perfect price flexibility, greater price flexibility means greater output stability, while farther away from it, the reverse is true. Similar results are obtained by Caskey and Fazzari [1988] and Chadha [1989].

EMPIRICAL EVIDENCE

We do not need fancy econometrics to mobilize evidence against the "real business cycle" view that observed fluctuations in output and employment are movements in price-cleared equilibrium. Here are a number of regularities of U.S. business cycles that falsify the implications of the New Classical hypothesis [Okun, 1980].

1. Unemployment itself. If people are voluntarily choosing not to work at prevailing wages, why do they report themselves as unem-

ployed, rather than as "not in labor force"? Real business cycle theory explains fluctuations of unemployment as intertemporal choices between work and leisure. Workers drop out when real wages, the opportunity costs of leisure, are temporarily low relative to what they expect later. This might be an explanation of cyclical movements in employment if real wages were strongly pro-cyclical, but there is no such systematic regularity. Nor is there empirical evidence of high sensitivity of labor supply to current and expected real wages.

2. Unemployment and vacancies. New Classicals ask us to believe that the labor market is in equilibrium at 9 percent unemployment, the same as it is at 5 percent. If so, there would be no reason to expect the balance between unemployment and job vacancies to be any different in the one case than in the other. *Both* unemployment and vacancies would be higher in recession. However, a strong negative association between unemployment and vacancies — as would be expected in Keynesian theory — is obvious in the U.S. and other market capitalist economies.

3. Quits and layoffs. If recessions and prosperities are both equilibria, there is no reason to expect the relative frequency of voluntary quits from jobs and involuntary "separations" to be any different. But of course there are many more layoffs, relative to quits, when unemployment is high and vacancies are scarce. There are many more "job losers" relative to "job leavers" in recessions.

4. Excess capacity. The utilization of plant and equipment varies cyclically parallel to the utilization of labor. Presumably machines are not choosing leisure voluntarily.

5. Unfilled orders and delivery delays. These move pro-cyclically, again suggesting strongly that demand is much higher relative to supply in prosperities than in recessions.

6. Monetary effects on output. According to the "classical dichotomy," monetary events and policies should affect only nominal prices. Real outcomes should be independent of them. The evidence that this is not true is overwhelming.

The list could go on. Why do so many talented economic theorists believe and teach elegant fantasies so obviously refutable by plainly evident facts? Trying to answer that question would take us into a speculative excursion on the sociology of the economics profession, beyond the scope of this paper.

NOTES

This paper is a written version of my lecture at the 1992 annual meetings of the Eastern Economic Association in New York City. The lecture and this paper draw on a longer paper with a similar message [Tobin, 1993]. I would like to express my gratitude for the faithful and valuable research assistance of Mitchell Tobin, Yale College 1992 (no relation).

REFERENCES

Ball, L., Mankiw, N. G., and Romer, D. The New Keynesian Economics and the Output-Inflation Tradeoff. *Brookings Papers on Economic Activity*, 1988:1, 1-65.

Caskey, J. and Fazzari, S. Aggregate Demand Contractions with Nominal Debt Commitments. *Economic Inquiry*, October 1987, 583-97.

_____. Price Flexibility and Macroeconomic Stability: An Empirical Simulation Analysis. Washington University Department of Economics, Working Paper 118, January 1988.

Chadha, B. Is Increased Price Inflexibility Stabilizing? *Journal of Money Credit and Banking*, November 1989, 481-97.

Delong, J. B. and Summers, L. H. Is Increasing Price Flexibility Stabilizing? *American Economic Review*, December 1986, 1031-44.

Dillard, D. The Barter Illusion in Classical and Neoclassical Economics. *Eastern Economic Journal*, October-December 1988, 299-318.

Fischer, S. Long-term Contracts, Rational Expectations, and the Optimal Money Supply Rule. *Journal of Political Economy*, February 1977. 191-205.

Fisher, I. The Debt-Deflation Theory of Great Depressions. *Econometrica*, October 1933, 337-57.

Hansen, A. H. *Full Recovery or Stagnation*. New York: W.W. Norton, 1938.

Keynes, J. M. *The General Theory of Employment, Interest, and Money*. New York: Harcourt Brace, 1936.

Okun, A. M. Rational-Expectations-with-Misperceptions As a Theory of the Business Cycle. *Journal of Money, Credit, and Banking*, November 1980 Part 2, 817-825.

Patinkin, D. Price Flexibility and Full Employment. *American Economic Review*, September 1948, 543-64.

_____. *Money, Interest, and Prices*. New York: Harper and Row, 1956, 2nd, ed., 1965.

Pigou, A. C. The Classical Stationary State. *Economic Journal*, December 1943, 343-51.

_____. Economic Progress in a Stable Environment. *Economica*, August 1947, 180-90.

Taylor, J. Aggregate Dynamics and Staggered Contracts. *Journal of Political Economy*, February 1980, 1-23.

Tobin, J. Keynesian Models of Recession and Depression. *American Economic Review*, May 1975, 195-202.

_____. Price Flexibility and Output Stability: An Old Keynesian View. *Journal of Economic Perspectives*, Spring 1993.

THE THEORY OF PRICE

third edition

George J. Stigler

The University of Chicago

The Macmillan Company, New York
Collier-Macmillan Limited, London

Second Printing, 1966

Earlier editions copyright 1942, 1946, 1952 by The Macmillan Company.

Library of Congress catalog card number: 66-19406

THE MACMILLAN COMPANY, NEW YORK
COLLIER-MACMILLAN CANADA, LTD., TORONTO, ONTARIO

Printed in the United States of America

Preface

Not only a man's ideas, but also his ways of expressing them, have a strong persistence over time, so it is possible for the statisticians to determine disputed authorship (as in the case of the *Federalist Papers*) by the pattern of words and the structure of sentences. I have rewritten the present edition almost completely, but I have no doubt that it is the same book, and by only a slightly different author. Its distinguishing feature continues to be its concentration upon the traditional central core of economic theory—the theory of value. I thank Sam Peltzman for helpful suggestions, Julius Schlotthauer and Richard West for doing much of the graphical work, and Claire Friedland for her assistance at every turn.

G. J. S.

The Theory of Price

2. Calculate the Laspeyres and Paasche indexes:

	YEAR 0	YEAR 1
Quantity of bread	200	170
Quantity of beef	100	120
Price of bread	15¢	12¢
Price of beef	20¢	25¢

Illustrate graphically and explain.

3. Suppose the total utilities of X and Y vary as follows:

$$TU_x = \sqrt{X}$$

$$TU_y = 10Y - Y^2 \qquad (Y < 5)$$

(a) Construct indifference curves between X and Y for a level of satisfaction of 24.

(b) Suppose the utility of each commodity doubles (to $2\sqrt{X}$ and $2[10Y - Y^2]$). Construct the indifference curves for a level of satisfaction of 48.

4. Demonstrate that people are better off with rationing by prices than with rationing by fixed allotments, given the distribution of income.

5. A consumer challenges you to disprove empirically his assertion that his indifference curves intersect. If you have an unlimited number of observations on his actual consumption (at all relative prices and incomes), how would you meet the challenge?

6. If the marginal utility of Y is constant, all indifference curves have the same slope at a given X. Prove.

chapter five

Pricing with Fixed Supplies

Once the demand curve of a commodity is established, we know the price at which each quantity can be sold. But we have begged two questions in constructing this demand curve: what is the market, and is it competitive? After we answer these questions we can analyse the pricing of commodities in fixed supply.

THE MARKET

A market, according to the masters, is the area within which the price of a commodity tends to uniformity, allowance being made for transportation costs.

The price of a commodity "tends to uniformity" for one reason: the buyers at point B refuse to pay more than the price at point A plus transportation, and the buyers at A act similarly. Or the sellers act in this manner. The market area may well differ between buyers and sellers.

As the buyer of an automobile, I will perhaps search only over a circle with a 10-mile radius about my home, so I may readily return to the dealer for services. But this cannot mean that the market area is 314.16 sq. miles, for other buyers are located elsewhere and their circles of search overlap mine. The market area, so far as buyers are involved, is the sum of the areas within which the mobility of consumers is sufficient to ensure the tendency to uniformity in price, allowance being made for transportation costs. For automobiles, this area will probably contain a city and its adjacent suburbs; for the services of gardeners it may be a small

portion of a city; for goods purchased by mail order it may be nation-wide.

The market area from the sellers' viewpoint will usually be larger than from the buyers' viewpoint. There is no important tendency for people in Minneapolis to buy potatoes in Maine. Yet one of the earliest statistical studies of demand revealed that the price of potatoes in Minneapolis depended upon the nation's output of potatoes, but given this output, was not influenced by whether the local output (in Minnesota and Wisconsin) was large or small.

> An investigation was made to determine the effect of variations in the production of Minnesota and Wisconsin taken together on the price of potatoes in Minneapolis and St. Paul. This investigation resulted in the discovery that variations in the production in Minnesota and Wisconsin had no measurable effect on the price of potatoes except to the extent that the production for the entire United States was affected.
>
> Although the fact is surprising, it is very readily explained when once recognized. The explanation will be somewhat clearer if the price situation as shown in [an accompanying figure] is borne in mind. Consider the extreme case of an excess production in Minnesota exactly equaled by a deficiency of production in Maine. In order to take care of the deficiency in the supply for New York City, for example, an unusual quantity is shipped in from New York and Pennsylvania. Large quantities of potatoes having been shipped east instead of west from New York and Pennsylvania, their place is taken by Michigan potatoes. But since Michigan potatoes are being shipped somewhat farther east than usual, Minnesota potatoes can be sold without competition in what is ordinarily Michigan territory. The result is that the Minnesota potatoes sell at practically the same price that would have been obtained if production in both Minnesota and Maine had been normal.[1]

On the other hand, sometimes the market area as defined by sellers is smaller than that of buyers: a cotton farmer will have a relatively small area in which he will sell his crop; the buyers may deal in every cotton-picking state.

Since the market is defined by the uniformity of price, its area will be at least as large as the larger of the areas of sellers' competition and buyers' competition, or the sum of the areas when they partially overlap.

[1] *H. Working, "Factors Determining the Price of Potatoes in St. Paul and Minneapolis," Technical Bulletin 10, University of Minnesota Agricultural Experiment Station (1922), p. 25.*

The size of the market also varies with the time we allow for price adjustments. A perishable good, once it reaches a given city, will be sold there even though it turns out that a higher price could have been fetched elsewhere—but future shipments will iron out the disparity. Once an apartment is built, its rental depends upon the housing demand of the community. But in the long run (meaning a period long enough for the supply of houses to be varied sufficiently), apartments will not be built where rentals are unremunerative, and more will be built where they are remunerative. There is accordingly a tendency for apartments of given quality to have the same rental throughout the country. But this tendency is slow in its workings because the stock of apartments changes very slowly, and it is modified by geographical immobilities of resources (in particular, land) which we shall discuss later. Because of the mobility of entrepreneurs and also of consumers, in the long run most markets are of very large geographical extent.

A perfect market is one characterized by perfect knowledge on the part of the traders. Or stated differently, in a perfect market no buyer ever pays more than any seller will accept, and no seller accepts less than any buyer will pay. These conditions can be met only in a completely centralized market, which is approximated by a few exchanges such as the New York Stock Exchange.

COMPETITION

A competitive market is easily defined only for a perfect market: it is then a market in which the individual buyer or seller does not influence the price by his purchases or sales. Alternately stated, the elasticity of supply facing any buyer is infinite, and the elasticity of demand facing any seller is infinite.

A market may obviously be competitive on only one side: a million buyers can deal with only one seller (monopoly) or a million sellers can deal with one buyer (monopsony). But for the time we shall defer such situations and deal only with competitive situations.

We have defined a perfectly competitive market: what are the conditions under which it will normally arise? The conditions are four:

1. *Perfect knowledge.* If there is not perfect knowledge, there will be an array of prices at which transactions will take place, and almost all real markets display such an array. There will then often be scope for higgling, and to this extent a situation termed bilateral monopoly arises. But if the scope for higgling is small, the departure from competition is small.

2. *Large numbers.* There must be many buyers or sellers if each is to have no appreciable influence upon the price,[2] and they act independently.

3. *Product homogeneity.* If the product is not homogeneous, it is meaningless to speak of large numbers. Hence, if every unit is essentially unique (as in the market for domestic servants), there cannot be large numbers. Yet, if the various units are highly substitutable for one another, the market can approach competition.

4. *Divisibility of the product.*

Perfect competition is a typical example of a concept of everyday life that has been taken over by economists and developed into something almost unrelated to its original form. Originally competition meant a multiplicity of traders, and only that. But when it was discovered that 5 traders might collude, a vast number seemed necessary to guarantee that collusion would not be feasible. When it was realized that even a thousand sellers and buyers were not enough if each pair dealt in ignorance of the others, perfect knowledge was added. The explicit recognition of homogeneity of product came from the fact that even minor differences (a sunny disposition or a fancy container) might lead some people to pay a slightly higher price.

Divisibility has a similar origin. Edgeworth, whom we have met before and shall meet again, was a diabolically clever man. He contrived the following problem: a thousand (or a million) masters hire one servant each—exactly the number available—and no servant can work for two masters. Each master will pay $100; each servant will accept $50—what will the wage rate be? That it will be between $50 and $100, and hence *indeterminate,* is no cause for anxiety. But let it be $50—then a single servant can leave the market and force the wage up to $100—so even perfect knowledge, large numbers, and (let us assume) homogeneity are not enough to de-

[2] *More precisely, the largest buyer or seller must provide only a small fraction of the quantity demanded or supplied, which involves, in addition to large numbers, no extreme inequality of size.*

prive an individual of a large influence over the market price. Hence we assume divisibility, so the departure of one worker leads (say) to about a 30-second lengthening of the working day for other workers, and his power to influence price is destroyed.[3]

If the reader bristles at the acceptance of assumptions such as perfect knowledge and complete product homogeneity, he is both wrong and right. He is wrong in denying the helpfulness of the use of pure, clean concepts in theoretical analysis: they confer clarity and efficiency on the analysis, *without depriving the analysis of empirical relevance.* He is right if he believes these extreme assumptions are not *necessary* to the existence of competition: it is sufficient, for example, if each trader in a market knows a fair number of buyers and sellers, if all traders together have a comprehensive knowledge so only one price rules. The reason for not stating the weakest assumptions (necessary conditions) for competition is that they are difficult to formulate, and in fact are not known precisely. Again, more work for the next generation.

The Demand Curve of the Competitive Firm

Since the competitive firm contributes only a trifling fraction of the total market supply, it has a trifling influence on market price.[4] We may illustrate this influence by considering a market with a unitary demand elasticity ($pq = \$1{,}000$), in which there are already 100 firms. Each supplies 2 units and the price is therefore $\$1{,}000/200 = \$5.00$. An additional supplier now appears. If the 100 firms continue to supply 200 units, the new supplier faces the demand schedule:

QUANTITY	PRICE
0	\$1000/200 = \$5.000
1	1000/201 = 4.975
2	1000/202 = 4.950
3	1000/203 = 4.925

[3] An eight hour day contains 480 minutes—B.C. (before coffee breaks), so if each of the remaining 999 workers works slightly less than half a minute for the employer of the vanished servant, his employer's need will be satisfied. Or, alternatively, each of the thousand masters hires a worker for 30 seconds less.

[4] The use of "trifling" rather than "absolutely no" is a trifling concession to realism. It would be more precise to use the latter phrase, but then some students would believe that the theory is inapplicable where there is even a trifling influence, and this is not true.

If we compute the arc elasticity of the new supplier's demand at an output of 2, it is roughly

$$\frac{3-1}{3+1}\cdot\frac{4.925+4.975}{4.925-4.975}=-\frac{2}{4}\cdot\frac{9.9}{0.05}=-99.$$

It is, in fact, a general rule that under these conditions the elasticity of the demand curve of a firm is equal to the elasticity of the market demand curve *times* the number of sellers.[5]

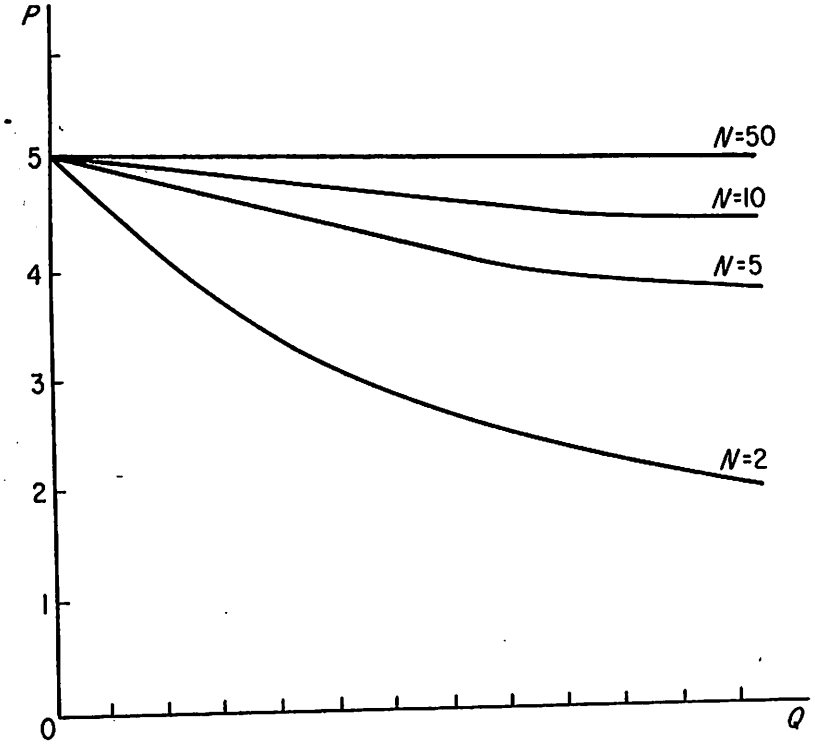


Figure 5-1

We illustrate such individual demand curves for 2, 5, 10, and 50 sellers in Figure 5-1, all on the assumptions that the market demand has unitary elasticity and the output of all firms except one sum to 200. Of course competition connotes fairly numerous sellers, and the demand curves for 2 and 5 sellers are put in only to display how rapidly the individual demand curves become elastic. It would require a very large diagram to make the difference between the demand curves for each of 50 and each of 100 sellers perceptible.

[5] *See mathematical note 7 in Appendix B.*

These demand curves were derived on the condition that all firms sell at the same price. But suppose 100 firms had agreed to fix the price at \$5, and one now contemplated his demand curve if he secretly cut the price to \$4.99 to trustworthy buyers. Assuming that the 99 other firms continued to adhere to \$5; the demand function of this price cutter would be

PRICE	QUANTITY DEMANDED
\$5.01	0
\$5.00	2
\$4.99	$\frac{1000}{4.99}=200.4$

Now his elasticity of demand is approximately

$$\frac{200.4-0}{200.4+0}\div\frac{4.99-5.01}{4.99+5.01}=-\frac{10.00}{0.02}=-500.$$

Of course if he cuts prices secretly and expands sales immensely, the other 99 firms will soon discover their sales are vanishing. But if he is moderate in his sales (perhaps only doubling sales to 4 units) he will reason that the price cutting will not be detected.[6] This reasoning will also be followed by at least 5 or 10 of his rivals, and if 10 double their sales to 4, only 160 (200 — 40) units will be demanded of the other sellers, each of whom will suffer, with rising animosity, a decline of 11 per cent in sales.[7]

This arithmetic portrays the history of a thousand price agreements. We shall discuss monopoly, which is what this is, at a later point, but it seems appropriate to emphasize here that large numbers of sellers not only make the formation of collusive agreements difficult, but also encourage each individual seller to violate the agreement.

PRICE DETERMINATION

Commodities in fixed supply, at least for limited time periods, are very numerous: they include the paintings of Rembrandt, the first editions of Shakespeare, and the number of Fords or Chevrolets

[6] After all, each rival will lose only 2/198 units or 1 per cent of his sales.

[7] Sales of each will be 160/90 = 1.78, a decrease of 0.22 from 2.

five years old. They include also the number of shares of common stock in a large industrial company, and the number of dwelling units in a city—at least for a time. Historically the most important example of all has been the stock of an agricultural product between harvests.

The same apparatus of supply and demand can be used for all these markets. But the details of the apparatus vary in an important respect with one characteristic of commodities and services: can they be stored? Let us call commodities which cannot be stored *perishing*. It is customary to describe all goods which will not survive either time or repeated use as perishable, but those which may be used only once (like a bullet) are often capable of storage for a long period.

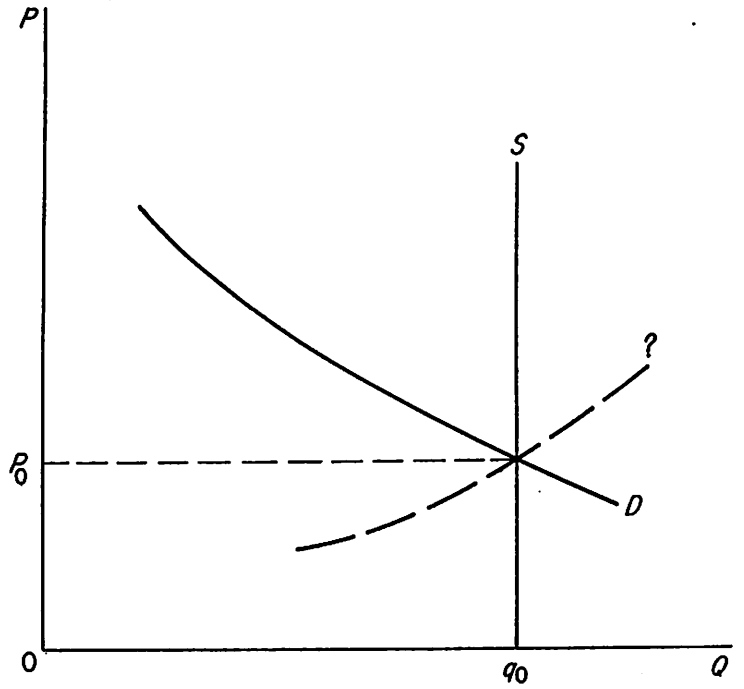


Figure 5-2

Perishing Commodities

The traditional case of a perishing commodity was fresh fish or strawberries brought to market before preservation by freezing was

possible. The stock was naturally thrown on the market (under competition) for what it would fetch, and we may translate this behavior into a vertical supply curve (S in Figure 5-2). The demand curve is determined by tastes, income, and prices of other goods, as described in previous chapters, and the intersection of the two curves (at p_0, q_0) is the equilibrium price. The quantities would be per day or other period for which the perishing commodity remained salable.

The equilibrium price is the price from which there is no tendency to move, so long as the underlying supply and demand conditions do not alter. It is a stable equilibrium, in the sense that if the market is jarred off equilibrium, the dominant forces push it back toward this equilibrium position. For example, if a rumor of a shortage of the commodity drives the price above p_0, the fact that the quantity supplied exceeds the quantity demanded will drive the price down toward equilibrium.

These terms were obviously borrowed from physics—has the economist made sure that they really make any sense in economics? The answer is, let us hope, yes. The stability of equilibrium is indeed the normal state of affairs in a tolerably stable world, and from it we deduce important properties. For example, there is a mysterious dotted line through (p_0, q_0) in Figure 5-2 which I have not had the audacity to label a demand curve. If it were, the intersection with the supply curve would still be an equilibrium point, but it would be highly unstable: the slightest accidental fall in price, for example, would drive price ever lower, because at each lower price the quantity supplied exceeds the quantity demanded. A stable equilibrium, then, implies that an increase in the quantity supplied must lower the price, so it implies (in this case) a negatively sloping demand curve. Stability conditions are a source of information at many points in the subsequent chapters.

Is stability something we can take for granted? Economists have generally argued its acceptance on the intuitive ground that wildly unstable market prices (and quantities traded) are not often observed. This is a relevant consensus, although not a conclusive one. There are in fact some cumulative processes in economic life (one has the name of galloping inflation), but we shall follow the general practice of assuming that the equilibria are stable.

Now that fish and strawberries can be frozen, are there any perishing commodities left? A few commodities like Christmas trees

and cut flowers are perishing, but the important examples are in services. The motel rooms for rent on a given day in a given area are essentially fixed in number and under perfect competition would be thrown on the market each day at an estimated full-occupancy price, so long as the price exceeded any costs of occupancy. There are two reasons why this flexibility of price is not fully attained, although there are seasonal variation in rates, higgling, and so on. The first reason is that some monopoly power may be possessed by the owners, and the second is that the costs of searching are high for the tourist and not negligible for the owner. The symphony concert, the train or plane on a scheduled run, the services of professional men at a given time, the supply of longshoremen on a given day—are all instances of essentially perishing services. Some have prices which do not clear the market because of public or private controls.

But tolerable stability of price is not inconsistent with a price that clears the market. If the demand is steady, the day-to-day fluctuations in price will not be large (unless supply fluctuates). And if demand is postponable (storeable), the same effect can be achieved. Suppose that the supply of cut flowers fluctuates erratically, and that consumers consider flowers tomorrow to be a very good substitute for flowers today. They will then have, on any day, a highly elastic demand for flowers, and the price will be relatively stable.

The usefulness of even the simple graphical analysis of Figure 5-2 is likely to be underestimated by students who have not experienced the ability of men to make mistakes. Consider one of the attacks launched on the "law of supply and demand" by William Thornton just before graphical techniques were introduced in England.[8]

> When a herring or mackeral boat has discharged on the beach, at Hastings or Dover, last night's take of fish, the boatmen, in order to dispose of their cargo, commonly resort to a process called "Dutch Auction." The fish are divided into lots, each of which is set up at a higher price than the salesman expects to get for it, and he then gradually lowers his terms,

[8] *The full attack (not this instance) happens to be famous because it led, or permitted, John Stuart Mill to abandon the wages-fund doctrine. Mill's position in English economics in 1869 was roughly that of Napoleon in the French Army in 1810, so the abandonment was the source of some comment, especially since the criticisms were flimsy. The quotation is from On Labour, pp. 47–48.*

> until he comes to a price which some bystander is willing to pay rather than not have the lot, and to which he accordingly agrees. Suppose on one occasion the lot to have been a hundredweight, and the price agreed to 20 *s*. If, on the same occasion, instead of the Dutch form of auction, the ordinary English mode had been adopted, the result might have been different. The operation would then have commenced by some bystander making a bid, which others might have successively exceeded, until a sum

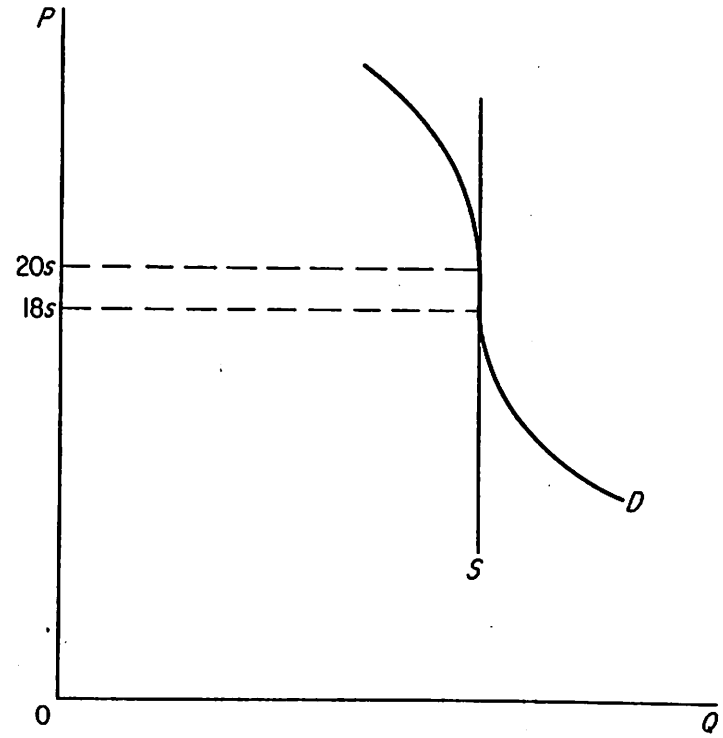


Figure 5-3

> was arrived at beyond which no one but the actual bidder could afford or was disposed to go. That sum would not necessarily be 20 *s*: very possibly it might be only 18 *s*. . . . In the same market, with the same quantity of fish for sale, and with customers in number and in every other respect the same, the same lot of fish might fetch two very different prices.

If we translate Thornton's criticism into a diagram (Figure 5-3), we observe immediately that the result is due to the fact that his demand curve has a vertical branch. This is absurd in a competitive *market* demand curve.

Storeable Goods

Let us turn to the more important case of storeable goods, shares of stock or sheaves of wheat or those first editions. Now the supply curve is no longer a vertical line, denoting the total absence of alternatives for the seller—for he has always the alternative of selling tomorrow, or never.

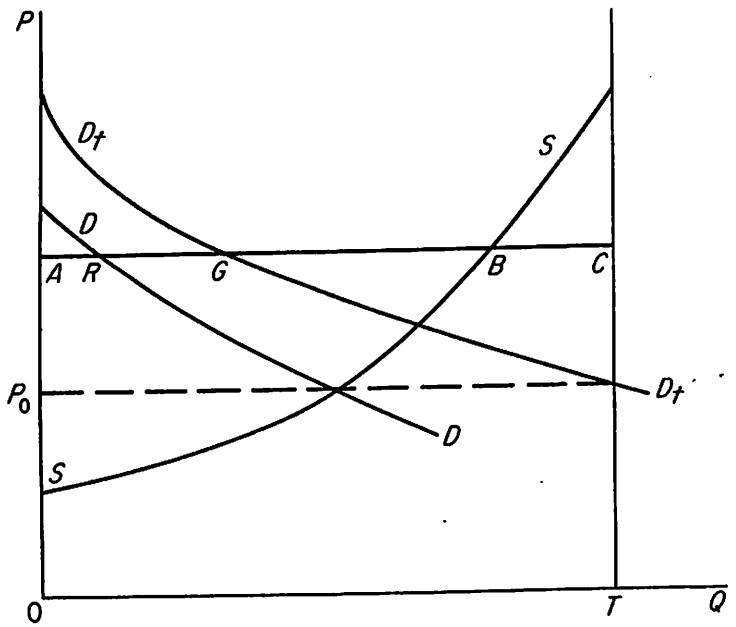


Figure 5-4

In fact the very identity of a seller may be uncertain. Jones may sell his four-year-old car at one price, and buy a second at another. Where this uncertainty arises it can be dealt with by the device of "reversing" the supply curve. Let us arbitrarily divide up traders in a market into "buyers," who have none of the commodity, and "sellers," who posses initial stocks of it (which they may wish either to augment or to sell). The "buyers'" demand curves will be those already discussed in Chapters 3 and 4. The "seller's" supply curve, as we have shown (p. 65), will be constructed in exactly the same way. These curves are shown in Figure 5-4. The total stock is shown

as the vertical line TC. At any price $0A$, a definite quantity is demanded by "buyers," AR. At this price a quantity AB is supplied by sellers, but we can alternately say that sellers wish to hold BC at this price, for this is the portion of the stock which they do not offer for sale. Thus if the total stock ($0T$) is 150, and sellers offer 110 (AB) at price $0A$, they are implicitly demanding 40 units (BC). If we add AR and BC, to get AG, we have the total quantity demanded at this price by "buyers" and "sellers." Applying this procedure at all prices, we obtain the aggregate demand curve D_t, and its intersection with the total stock line sets the equilibrium price P_0.

This construction does more than evade the minor problem of classifying buyers and sellers. It illuminates a common fallacy. Many people have said that if a stock sells for $40 a share on a given day that this is not the "true" price because only a modest number of shares were traded—if a huge block had been thrown on the market, the price would have fallen drastically. Indeed it might have, but if a huge block had been thrown on the market this would have meant that many holders now believed the stock was a poor investment at the price. In effect a large decrease in demand has been implicitly assumed. Since the large block of stock was not thrown on the market, the holders thought it was worth at least this much. The fact that one could not buy a large block of the company's outstanding stock at the market price, similarly, merely means that one cannot double the demand without influencing the price.

The holder of a durable commodity has to take account of two elements of return:

1. The marginal utility (measured in money terms) he derives from holding the commodity. Examples are the pleasure of driving a car or of admiring a painting or of cashing dividend checks paid on a stock.[9]

[9] The condition for maximum satisfaction is

$$\frac{\text{Marginal Utility of } A}{\text{Price of } A} = \frac{\text{Marginal Utility of } B}{\text{Price of } B}$$

This ratio is called the marginal utility of income because it is the amount of utility received per dollar of expenditure at the margin. If we divide the marginal utility of (say) a painting by the marginal utility of income, we obtain the marginal utility of the painting expressed in dollars.

2. The change in the price of the product from now to (say) next year, which may be positive or negative.

The total return from holding the commodity then consists of the sum of the utility (u) and the expected increase in price (Δp). The owner of a first edition of Ricardo's *Principles of Political Economy and Taxation* (1817), of which there are probably 400 copies in the world, expects its price to rise because the number and wealth of potential owners are increasing. But the price cannot on average rise so fast as the interest received on sums of money invested in securities comparable in riskiness to holding Ricardo's *Principles*. If it did, economists would buy the book and have the pleasure of owning it without cost, while receiving the increment of value. In equilibrium, in fact, $u + \Delta p$, the (marginal) return to the holder, must equal the cost of holding the durable good. This cost is composed of the amount that could be earned on the sum elsewhere, ip (where i is the appropriate interest), plus any cost of possession of the good (insurance of a painting, and so forth).

It follows that the greater the utility to be derived from holding a commodity, the lower must be its rate of increase of price. People will not hoard a keg of nails unless its price is expected to rise by i per cent; they will hold the Ricardo if it rises by only $\Delta p = ip - u$, or $\Delta p/p = i - u/p$ per cent, where u/p is the annual utility of possession per dollar invested in the commodity.

Speculation

A more interesting and important pricing problem is posed by the existence of stocks of goods which are periodically produced—agricultural products are of course the leading example. The tasks in rationing a fixed supply until the new crop is harvested are two: to provide supplies throughout the period of fixed supply, so that the entire stock will not be consumed early in the year; and to provide a carry-over as insurance against future crop failures, increases in demand, and the like.

The former task is relatively the easier one: the demand for foodstuffs and textiles is tolerably stable over the period of a year, although there are some fluctuations in demand due to fluctuations in consumer income, seasonal changes in tastes, changes in foreign demands, and so on. If demand (as a schedule or curve) were absolutely identical in every month, and no carry-over was needed, the price would rise each month by the costs of holding the stock. For, if the price were uniform, any holder this month would have the choice of selling his stock now at a given price, p, or of holding it a month and receiving only $(p - c)$, where c is the cost of carrying the good a month. Therefore he would sell now, until the current price was depressed, and the price next month elevated, enough to cover the costs of carrying a stock for a month. This gradual rise in price is in effect the method of charging consumers for the service of holding the stock.

The second task, providing a stock for emergencies, is less simple. As of any time there are an immense array of possible events, each of which will influence the price at any future date. Let our commodity be wheat, with a current price of $2 a bushel; then the possible events may include:

1. A future crop failure, which can be large or small, leading to prices ranging up from $2 to $4, with smaller probabilities of the bigger failures and higher prices.
2. A future bumper crop, also of variable size, with corresponding future prices from $2 down to $1.
3. A business depression, leading to a modest decline in price and quantity demanded.
4. A war, leading (perhaps through conscription of farm workers) to a reduction in output and a higher price.
5. A fair prospect of increased or decreased demand for exports.
6. A possible shift in consumer tastes away from wheat toward meat.

The only thing a holder of wheat can be quite certain of is that something unusual will happen.

The carry-over will be held in warehouses, but who will own it and take the risks of profits or losses? The natural answer is: a group of people who specialize in predicting future demands and supplies. This group, called speculators, develops skill in collecting and assessing current evidence on future conditions, and therefore on average can perform this task more efficiently than, say, the processors of wheat (grain mills).

Each speculator may be described as making a set of estimates of the probabilities of various conditions of supply and demand

at a given future date. These estimates may be assembled into a frequency distribution such as:

PROBABILITY	EXPECTED PRICE
0.05	$3.00
0.10	2.50
0.20	2.25
0.35	2.00
0.20	1.80
0.10	1.60

The average expected price is then simply the sum of the products of the expected prices and their probabilities, which is $2.07 in this case. The confidence with which this estimate is held may be measured by the dispersion about this expected average; obviously the speculator will have more confidence in this price ($2.07) being approached with the above distribution than with

PROBABILITY	EXPECTED PRICE
0.38	$3.00
0.62	1.50

which also has a mean of $2.07. Presumably he will make larger commitments on his prediction the greater his confidence in it.

If the commodity is one that has no futures market—no market in which contracts for future delivery are bought and sold—the trader will buy wheat if the present price *plus* carrying costs is less than $2.07; he will get out of the market if this is not the case.[10] With a futures market, however, he will sell contracts for future delivery if the price he expects is below the futures price currently quoted (and buy futures contracts in the converse case), with the hope of covering the contract (with a "spot" purchase) when it matures at the expected lower price.

Each speculator has a different set of expectations, and a different demand-supply function for futures contracts. We may add them together to get the aggregate demand for (say) May futures in the previous December, as a function of the price of futures contracts; it is denoted D in Figure 5-5. If the futures price is above

[10] *Thus, if real estate prices are expected to fall, it is impossible to sell land "short" because it is not homogeneous and therefore one cannot promise to deliver a particular piece at some future date.*

the price that speculators anticipate, they will supply futures contracts, and at lower futures prices they will demand contracts.

The supply of futures contracts is provided by hedgers—of whom it is sufficient to notice those who buy the wheat from farmers and supply storage. If they do not wish to speculate, they can eliminate their risks by selling futures contracts at prices equal to at least the current price plus carrying costs. Their supply together with the speculators' demand (D) fix the present price of futures contracts.[11]

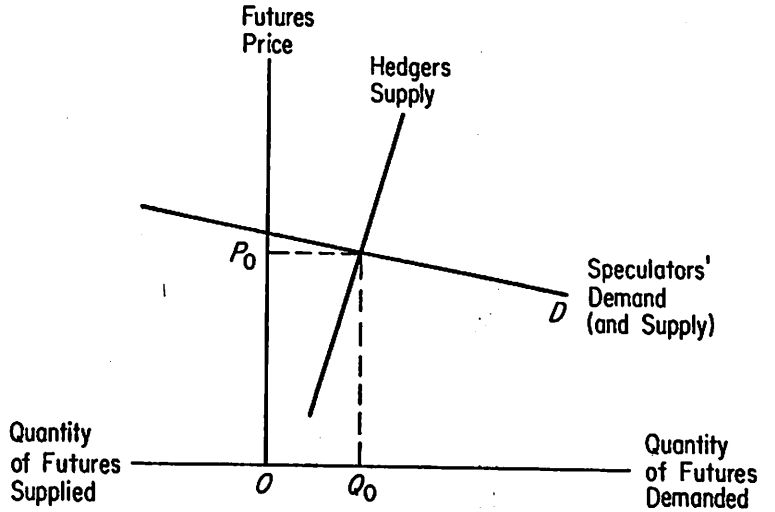


Figure 5-5

The skill with which this delicate task of reading the future is performed is a much-debated point. It is undeniable that if anyone can predict future prices more accurately than the professional speculators, he can make a vast amount of money rather quickly. It is also undeniable that most nonprofessionals (who, by Barnum's law, are constantly being replenished) manage to keep alive by borrowing money from relatives.

The public, and especially farmers, have nevertheless always been hostile toward the speculators, and wave aside the economist's arguments on the need for someone predicting the future and taking

[11] This is a much simplified picture; see L. G. Telser, "Futures Trading and the Storage of Cotton and Wheat," *Journal of Political Economy*, 66 (1958), 233–55.

risks that the predictions are wrong. Ancient laws against forestalling, engrossing, and regrating—buying foodstuffs on the way to market or in a market with a view to resale—are an adequate proof of this popular suspicion. This policy led Adam Smith to say,

> It supposes that there is a certain price at which corn is likely to be forestalled, that is, bought up in order to be sold again soon after in the same market, so as to hurt the people. But if a merchant ever buys up corn, either going to a particular market or in a particular market, in order to sell it again soon after in the same market, it must be because he judges that the market cannot be so liberally supplied through the whole season as upon that particular occasion, and that the price, therefore, must soon rise. If he judges wrong in this, and if the price does not rise, he not only loses the whole profit of the stock which he employs in this manner, but a part of the stock itself, by the expence and loss which necessarily attend the storing and keeping of corn. He hurts himself, therefore, much more essentially than he can hurt even the particular people whom he may hinder from supplying themselves upon that particular market day, because they may afterwards supply themselves just as cheap upon any other market day. If he judges right, instead of hurting the great body of the people, he renders them a most important service.
>
> The popular fear of engrossing and forestalling may be compared to the popular terrors and suspicions of witchcraft."[11]

Smith's comment needs only a minor qualification. The proposition that speculators cannot as a group make money by inducing price fluctuations or withholding supplies is correct if they possess no monopoly power. In actual fact, any large or persistent degree of power to control prices has not been attainable in the large commodity markets, simply because the trade of buying and selling is one which it is easy for anyone to enter. Smith is too kind to those who suspect speculators, however, when he compares their attitudes with those once held toward witches, for there was no proof that the witches had not entered into compacts with the devil.

RECOMMENDED READINGS

Hardy, C. O., *Risk and Risk-Bearing,* Chicago: University of Chicago Press, 1933, Chs. 11, 12.

Working, H., "The Theory of Price of Storage," *American Economic Review,* 39 (1949), 1254–63.

[11] *The Wealth of Nations* (New York: Modern Library ed.), p. 500.

PROBLEMS

1. You are given the information: the total stock of a commodity is 100, the demand function is $q = 80 - p$, and the supply function is $q = 5 + p$. Derive the combined demand curve of buyers and sellers.

2. Mountifort Longfield (1834) argued that the practice by the rich of buying wheat in years of small crops and reselling it to the poor at half price did not reduce the cost of wheat to poor consumers, as compared with having the poor buy directly at market prices. Compare your analysis with his (*Lectures on Political Economy,* London School Reprints, 1931, p. 56).

3. The market demand curve is $p^2q = 1000$ (a constant demand elasticity of -2). Derive the demand curve for one of 2, 10, and 40 firms, all of equal size, with the aggregate industry output of 200 when the firm in question is operating at the same output as other firms. (Thus with 10 firms, each of the other 9 firms has an output of $20 = 200/10$.)

4. On a certain morning you find the following foreign exchange rates quoted:

	PRICES		
	U. S. Dollars	*£ Sterling*	*Francs*
Pound Sterling	2.80	1.	14.0
French francs	0.2025	0.0714	1.0
American dollars	1.	0.36	5.0

(a) What do you do to make money?

(b) Suppose no one bothered with arbitrage, and the rates persisted. What, if any, economic objections are there to these nonequilibrium quotations?

5. The larger the number of traders in a market, and the larger the dollar volume of transactions, the smaller will be the spread between bid and ask prices for a commodity. Explain why.

6. It has been observed that the best grades of products (oranges, apples, and the like) are sent to large cities and are not readily available to consumers in the areas in which they are produced. Explain why.

7. In a market in which carrying costs are negligible (such as common stocks), you are told the price of the commodity at time t—say, $P_t = \$100$. If the market consists of intelligent traders, will it be of any value to a speculator to know what the price was a time unit earlier? (That is, would it be useful to know whether P_{t-1} had been \$50 or \$200?) More generally, can repetitive patterns of prices over time exist?

3. What would happen to the value of men and acres in our example (p. 118) if the labor force increased to 1,200 men?

4. You are told that 100 bushels of wheat can be produced by either 4 man-hours and 2 acre-years or by 3 man-hours and 3 acre-years. Can the marginal products of men and land be determined with this information?

5. Explain or denounce the propositions:

(a) There is no such thing as a free lunch.

(b) There cannot be two expensive lunches.

chapter seven

Production: Diminishing Returns

At a given time there is a set of "technological" possibilities open to any potential producer of any commodity. These possible techniques are commonly labeled "technological" without quote marks, and we shall henceforth dispense with them, but the quote marks should serve to remind us that the methods of converting coffee beans at a port warehouse into coffee ground to specification at a grocery store consist of more than the technical details of the ways of roasting coffee, putting it into bags, and transporting it to buyers. Production involves also the carrying of inventories which are not too large (for they are expensive) or too small (or sales will be lost), the hiring of workers of all descriptions and getting them to work well, borrowing money and collecting debts, advertising and quarrelling with the Federal Trade Commission, detecting changes in consumer tastes, and making out tax returns. The plebian phrase, "know-how," better describes this set of possibilities.

An inventory of all known ways of producing goods—using production in its widest sense to include methods of organizing economic activity—is referred to as the "state of the arts." This inventory contains many methods that no one will use because they are obsolete: they yield goods that are no longer desired; or yield desired goods but require larger amounts of all inputs than other known methods. It contains also many methods that cannot be ranked unambiguously as superior and inferior: process A uses more machinery, process B more labor—so which is more efficient will depend upon the prices of machinery and labor. This inventory

of knowledge grows over time as new discoveries are made. We shall nevertheless assume that it is fixed.

Even in the absence of new discoveries, the "state of the arts" is an immense collection of possibilities, and of the most varied sorts. In fact it contains all published knowledge and the vast empirical experience reposing only in men's heads. It is similarly indescribable in its variety: it contains the methods of making doughnuts (on a large and small scale) and airplanes, of collecting delinquent accounts and recruiting employees, and what not.

The student should therefore be suitably impressed to learn that economists discovered a general law relating the quantities of inputs and the quantity of output for any productive process. The discovery of this law, due to T. R. Malthus (of population fame) and Edward West (who deserves to be as famous) in 1815, was one of the heroic advances in the history of economics.

It turns out that much more can be said about the relationship of output to one of several inputs than about the relationship of output to all inputs, so we begin with this case. This relationship—the law of diminishing returns—answers the question: in what *proportion* should the various inputs be combined?

DIMINISHING RETURNS

The law of diminishing returns may be stated quite briefly:

As equal increments of one input are added, the inputs of other productive services being held constant, beyond a certain point the resulting increments of product will decrease—that is, the marginal products will diminish.

The law is not a tautology, but an assertion about the real world. As such, it must be interpreted in a particular way—even the physical law that freely falling bodies have constant acceleration does not work well if the body is in a tub of molasses. In our case the conditions are:

1. That there be other inputs whose quantities are held constant. If all inputs vary, we have the problem of economies of scale, discussed in the next chapter.

2. The state of technological knowledge is given. The various input-output possibilities are all available at the same time. Obviously if an additional unit of labor is applied to a farm next year, and a new invention makes the product rise more than it did when a man was added this year, this is no contradiction of the law.

3. The proportions in which inputs can effectively combine are variable, or in other words, the coefficients of production are variable (p. 114). The law has relevance even if this condition fails but we shall discuss only the important situation of continuously variable proportions.

Production is a process, not an act, so all of the inputs and outputs are rates of flow per unit of time: man-years, bushels per year, and so on. If economists used completely meticulous language, they would therefore emphasize this flow nature by speaking, not of hiring 7 men, but of hiring the services of 7 men for a year; not of producing 2,000 bushels, but 2,000 bushels per year. They are not this meticulous, and it is customary to refer to productive "factors" rather than their services.

This carelessness has on occasion led to error. For example, it has been said that labor (service) is perishable but capital (a building or machine, say) is not. Yet surely if the services of a man or a machine are not used this year, there is a loss in either case. It will be roughly true that the man's future services are no larger because of this year's unemployment, but machines also rust or become obsolete and in any case a year's services which are postponed 10 years are worth much less than they would be this year.[1] We shall not examine the relationships between services and the capital goods which yield them until we reach the theory of quasi-rents.

Elaboration of the Law

Let us begin with a simple numerical illustration of the law of diminishing returns. In this numerical example (Table 7–1), a series of amounts of labor (M = man years) are used in cooperation with an amount of land (L = acre years) which we hold constant. Diminishing returns sets in with the fifth unit of labor.

[1] Future services must be discounted to obtain their present value, so a dollar of services 10 years hence is worth only $\$1/(1+0.1)^{10} = \$0.39$ if the interest rate is 10 per cent.

It will be noted that the average product of labor begins to diminish only after six units of labor are employed, so average and marginal products begin to diminish at different points and diminish at different rates. Until well into the present century the law of diminishing returns was often stated in terms of both average and marginal products, and they were treated as equivalent. We see that

Table 7–1

MAN YEARS	TOTAL PRODUCT	AVERAGE PRODUCT PER MAN YEAR	MARGINAL PRODUCT OF A MAN YEAR
0	0	0	—
1	5	5	5
2	13	6.5	8
3	23	7.7	10
4	38	9.5	15
5	50	10	12
6	60	10	10
7	68	9.7	8
8	75	9.4	7
9	81	9	6
10	86	8.6	5
11	89	8.1	3
12	91	7.6	2
13	92	7.1	1
14	92	6.6	0
15	91	6.1	−1
16	88	5.5	−3
17	84	4.9	−4

they are not equivalent, and in fact only marginal products are of interest to the economist.

We can demonstrate the importance of marginal products at once by asking the simple question: if the wage rate of labor is 6 units of product, how many laborers should the owner of a plot of ground hire? The arithmetic is performed in Table 7–2, which is based squarely on the data of Table 7–1. The owner will wish to maximize his surplus, which is achieved when he hires 9 men—which is of course where the marginal product of labor equals its cost. Marginal

products are always the guide to maximum profits or minimum cost: wherever a productive service has a different marginal product in two uses, we can increase total product. Thus, if labor had a marginal product of 10 on one farm, and 8 on another, transferring one laborer from the latter to the former farm would increase total product by 2, and the gains continue (at a declining rate) until the marginal products are equal.

Table 7–2

NUMBER OF MAN-YEARS HIRED	TOTAL WAGE BILL AT 6 PER MAN-YEAR	TOTAL PRODUCT	SURPLUS OVER WAGE BILL
1	6	5	−1
2	12	13	1
3	18	23	5
4	24	38	14
5	30	50	20
6	36	60	24
7	42	68	26
8	48	75	27
9	54	81	27
10	60	86	26
11	66	89	23
12	72	91	19

When we are speaking of "applying" laborers to a plot of land, we can equally well speak of "applying" a plot of land to the laborers. When the marginal product of men declines in Table 7–1, we can say it is because there are more men per acre, or fewer acres per man—only the proportions are important. The law of diminishing returns is completely symmetrical, and it is a matter of indifference which input we hold fixed and which we vary.

The symmetry can be illustrated by deducing the marginal product of land from Table 7–1, on the assumption that 10 acres of land were in the plot. We may proceed along these lines: As an approximation (to be discussed in Chapter 8), if eight units of labor on 10 acres yield 75 units of product, then 9 units of labor on $\frac{9}{8} \times 10$ $(= 11.25)$ acres will yield $\frac{9}{8} \times 75$ $(= 84.375)$ units of

product. (That is, proportional increases of all inputs lead to proportional increases of output.) The table tells us that 9 men on 10 acres yield 81 units of product. We may now calculate the marginal product of land by comparing the outputs with 10 and 11.25 acres, holding labor at 9 units:

$$\frac{84.375 - 81}{11.25 - 10} = \frac{3.375}{1.25} = 2.7 \text{ per acre.}$$

When we move from a ratio of labor to land of $\frac{8}{10}$ to one of $\frac{9}{10}$, we found that the marginal product was 6 per man. Now, as we reverse the movement and go from a ratio of labor to land of $\frac{9}{10}$ to one of $\frac{9}{11.25} = \frac{8}{10}$, we find that the marginal product of land is 2.7 per acre.

We give both marginal product curves (with $L = 10$) and the total product curve in one diagram (Figure 7-1, based on Table 7–1). As we move to the right, the ratio of labor to the land rises; as we move to the left, the ratio of land to labor rises. The diagram is divided into three stages, which correspond to three possible stages of returns:

1. In the first stage the marginal product of the land is negative.
2. In the second stage the marginal products of both factors are positive and diminish as the factor increases.
3. In the third stage the marginal product of the labor is negative.

The first and third stages are thus completely symmetrical.[2]

The entrepreneur will seek to be in the second stage, where neither input is being used in so large a quantity as to reduce the level of output. Even if labor is free, he will go only to the end of the second stage, and even if land is free he will stop at the beginning of the second stage. This latter condition was approached in colonial days, when land was almost free. The colonists were properly lavish in their use of land relative to labor, despite the frequent complaints of European visitors who were accustomed to the more intensive utilization of more expensive land and transferred their notions of appropriate technique to inappropriate relative prices of labor and land.

[2] These precise relationships between average and marginal products hold only if a given proportional change in all inputs leads to an equal proportional change in output; see mathematical note 9 in Appendix B.

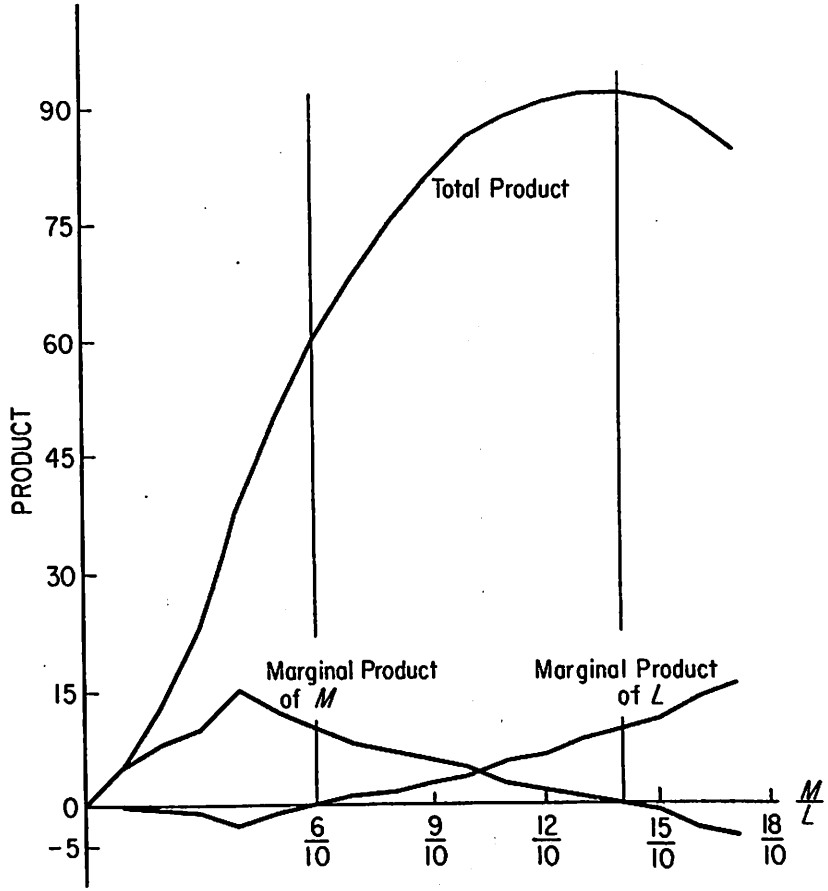


Figure 7-1

In our examples we have assumed that if the ratio of labor to land is sufficiently small, no product will be obtained. This is not impossible: one man-hour applied to an entire 160 acre farm will yield nothing but a brisk stroll. Nor is it necessary. Suppose we apply a variable amount of fertilizer to given quantities of land and labor. If no fertilizer is used, some product will nevertheless be obtained, so the total product curve begins some distance above the origin. An example is given in Table 7–3.

Table 7–3

POUNDS OF FERTILIZER	BUSHELS OF WHEAT PRODUCED	MARGINAL PRODUCT*
0	18.3	—
43	28.6	10.3
86	37.1	8.5
129	39.0	1.9
172	39.5	0.5

* Per 43 pounds of fertilizer.
SOURCE: The example is taken from F. L. Patton, *Diminishing Returns in Agriculture* (New York: Columbia University, 1926), p. 34.

We have so far assumed also that there is an initial stage of increasing marginal returns to labor and this is also possible but unnecessary. Marginal product may begin to diminish with the first units of the variable service; this is also illustrated in Table 7–3, although the size of the increments of fertilizer is so large that we cannot be sure that an initial stage of increasing marginal returns has not been overlooked.

The converse is also possible: the initial stage of increasing marginal product may be so broad that the demand for the required product is obtained before the second stage is reached. But if the productive service being held constant is divisible it would be unnecessary even in this case to employ it with a negative marginal product. Suppose we need only 13 units of product, given the production schedule of Table 7–1. Using again the approximation that proportional changes in all inputs yield proportional changes in output, we may proceed as follows: 6 units of the variable service with 10 units of the constant service yield 60 units of product, so $\frac{13}{60} \times 6$ (= 1.3) units of the variable service with $\frac{13}{60} \times 10$ (= 2.17) units of the constant service will yield $\frac{13}{60} \times 60$ (= 13) units of product. Hence by throwing away (10 − 2.17) = 7.83 units of the constant service we can save (2 − 1.3) = 0.7 unit of the variable service, still obtaining 13 units of product. If the fixed service is divisible, the entrepreneur will not operate in a region of increasing marginal returns to the variable service (and of negative marginal returns to the constant service).[8]

[8] It would be imprecise to say that by this device we have converted increasing returns into constant marginal returns to the variable service, for we are not holding the quantity of land in use constant.

The phrase "diminishing returns" has become part of ordinary language, so people now say that they stopped reading a book because they reached the point of diminishing returns. It is hopeless to fight against popular usage, but one should at least notice that almost always this usage is nonsensical unless reference is being made to diminishing *total*, not marginal, returns. One should indeed stop reading a book (even this one) if he is losing ground, unless it is ground that is a positive nuisance, but commonly the person means that the additional (marginal) pleasure or instruction is not sufficient to justify the time for further reading. I recommend the following language, especially with elderly aunts: I stopped reading the book because its marginal utility per minute had fallen below the marginal utility of alternative uses of my time, including sleep. This language is not only correct but has the interesting effect of always shifting the conversation to sleep.

The Role of Adaptability

The law of diminishing returns requires that we hold constant the quantity of one (or more) productive factors as we vary the quantity of the factor we are studying. In its most literal sense, this constancy implies that the quantity and form of the constant productive factors be unchanged: if we vary the number of men building a house, we nevertheless hold the number and type of tools constant. This is perfectly possible, and will of course usually yield fairly sharply diminishing returns, because if the tools appropriately equip n men, a larger number will have to resort to more primitive methods of work or tool-sharing.

There is another sense in which a factor may be held constant: its economic quantity (or value) can be held constant. We can hold the house-building tools at $2,000, say, but vary their form so that they are most appropriate to whatever quantity of labor we employ. With fewer men, we use fewer and more elaborate tools; with more men, we use more, but less elaborate, tools. Or conversely, if we are examining the marginal productivity for tools, we can hire fewer but abler workmen (with the same aggregate payroll) with fewer tools, and more but less able workmen with many tools.

This broader sense of "constancy" is obviously more appropriate when we are studying the behavior of an entrepreneur who seeks to maximize the output from given resources, if he can in fact

change the form of the constant factors. And normally he can make this change if given time: sooner or later the particular factors need to be replaced and they can then be replaced by more appropriate "constant" factors.

If the fixed productive factor need not be changed in form when the quantity of the variable productive factors is changed, the fixed factor is called adaptable. Adaptability is complete when the form of the constant factors is such that, whatever the quantity of the variable factors, the maximum output (with the known technologies) is achieved.

The difference between the products obtainable with partial and complete adaptability is illustrated in Figure 7-2. The extreme case of zero adaptability, it may be noted, would arise with fixed proportions—where the constant factor was literally incapable of being used with more or less than a critical quantity of the variable factor—and in this case the total product "curve" would simply be point P_0.

We shall later argue that the productive service which we arbitrarily hold constant in order to exhibit diminishing marginal returns is often actually fixed for the entrepreneur in the short run. Then he cannot make any magical transformation of the constant productive factor—it requires time to wear out such factors (if they are durable) or to rebuild them. Since the firm will nevertheless usually have a fluctuating output even in the short run, the entrepreneur will seek to have a flexible productive system—one which operates with tolerable efficiency over a considerable range of outputs. This flexibility can usually be achieved (at a cost): for example, it is possible to design an oil refinery so it can vary substantially the proportions in which gasoline, fuel oil, and other products are obtained from given crude oil. In terms of our diagram, the flexible plant will have a lower output at X_0 because, if versatility is expensive, a larger quantity of the constant factor is needed, but the marginal product will not fall so rapidly when the variable productive service is increased.

The Proof of the Law

The law of diminishing returns is, as we have said, an empirical generalization, not a deduction from the laws of matter. An empirical law (as we learned from the law of demand, p. 24) cannot

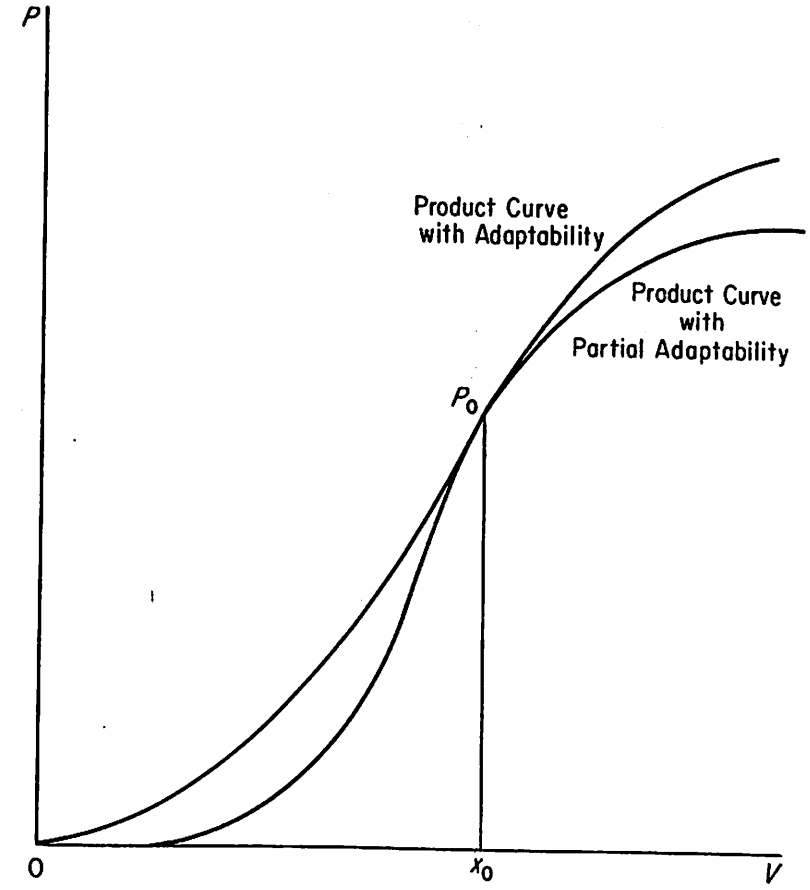


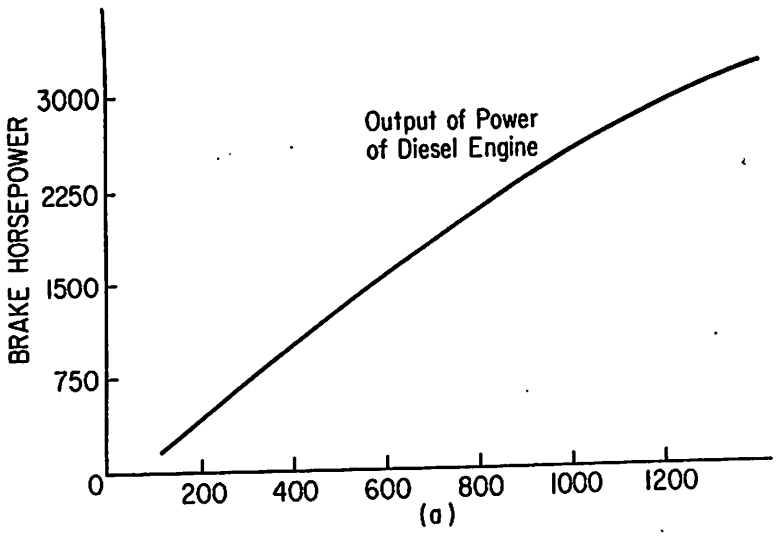
Figure 7-2

be proved by producing instances of its operation. This is not to say that such direct empirical evidence is irrelevant: in particular the law was immediately accepted by economists when it was first proposed simply because it seemed so clearly operative in agriculture. We could now produce a vast number of illustrations and in fact do give two samples in Figure 7-3 and Table 7-3.[4] A method

[4] Figure 7-3 is based upon "Trials of the T.S.M.V. Polyphemus," *The Institution of Mechanical Engineers, Proceedings,* 121 (1931), 183 ff. The equation of the total product curve is

$$Y = -128.5 + 2.740X + 0.0005110X^2 - 0.0000005579X^3,$$

where Y is brake horsepower and X is fuel input (pounds per hour).



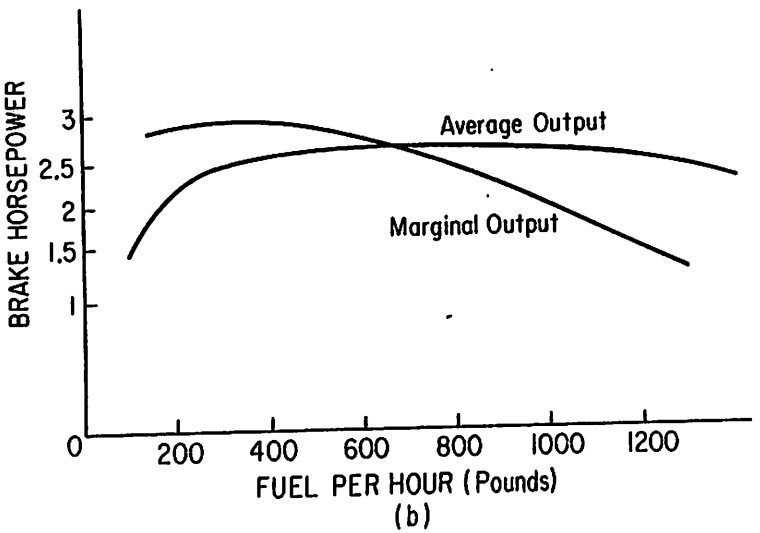


Figure 7-3

of testing the law that is especially relevant to economic analysis will be provided later in the chapter.

A large number of attempts have been made to prove the law by deriving it from self-evident facts. Perhaps the most famous proof assumes the opposite of diminishing marginal returns, and deduces that all the wheat in the world could be grown in one flower pot. It proceeds like this: suppose we have increasing marginal products on a 10 acre farm:

VARIABLE SERVICE	TOTAL PRODUCT
0	0
1	5
2	15
3	30
4	50

(The total product is then $\frac{5}{2}V + \frac{5}{2}V^2$ where V is variable service.) We proceed:

If 2 units of V on 10 acres yields 15,
1 unit of V on 5 acres yields 7.5.

Again:

If 4 units of V on 10 acres yields 50,
2 units of V on 5 acres yields 25,
1 unit of V on 2.5 acres yields 12.5.

It is evident that by decreasing the quantity of land we are increasing the total product from given quantities of the variable service. Using our equation for total product,

1000 units of V on 10 acres yields 2,502,500,
1 unit of V on $\frac{1}{1000}$ acres yields 2,502.5.

Since $\frac{1}{1000}$ of an acre is still a very big flower pot, let us do this once more:

1,000,000 units of V on 10 acres yields 2,502,002,500,000,
1 unit of V on 10^{-5} acres yields 2,500,002.5.

It is clear that we are doing very well by reducing the quantity of land: 10^{-5} acres is a plot of 5.2 sq. inches. The result should not surprise us unduly, however: if there are constant returns to scale (so reducing every input by K per cent reduces the product by K per cent) and if there are increasing marginal returns to one

factor, there must be negative marginal returns to the other, so reducing the latter naturally increases the total product (p. 127). But unfortunately for the proof, there is no basis for saying that there must be constant returns to scale, as we shall see in the next chapter. So the proof is inconclusive, as proofs concerning the real world have a habit of being.

SHORT-RUN COST CURVES

In order to isolate the marginal product of a productive service, we have held the quantities of the other factors constant. This procedure can be applied to any combination of factors: any one input (or group of inputs) could be varied, the remainder being held constant.[5]

In actual life, however, it is usually the case that the entrepreneur can vary the quantities of some inputs much more easily and quickly than the quantities of others. The proprietor of a factory can vary within a few days the number of employees he hires, the rate of supply of raw materials, the number of hours he operates the plant. It may require weeks, however, to hire specialized executives, or to obtain specialized machinery (which may have to be built to order), or to enlarge the factory building. The proprietor of a retail store can increase quickly the number of sales clerks and the supplies of goods, but it will take longer to enlarge the store. The proprietor of an electric generating company can expand quickly his use of fuel, but requires several years to obtain an additional generator.

This is loose language: when a proprietor says that he can quickly buy more steel sheet, but requires 7 months to obtain a new stamping machine, he is not being precise. At a sufficiently high price, one can buy a stamping machine from another company and have it installed in 24 hours; at a very high cost one can have a new machine built in a month by working around the clock. When we say that in the short run some inputs are freely variable, we mean that their quantity can be varied without affecting their price (for given quality). When we say that other inputs are not freely variable we mean that their quantities can be varied within the

[5] When more than one productive service is variable, they must be combined in the most efficient proportions, which may vary with the rate of their use. This problem is discussed below, p. 146.

given time unit—be it a week, a month, or a year—only at a considerable change in their price: if we try to sell the specialized machine, it has little value in other uses; if we try to buy more, price rises sharply for early delivery.

Fixity and variability are matters of degree. The plant's supply of electricity can be increased instantaneously (without a change in price); it may require five years to find a gifted designer. In order to simplify the formal theory, economists define "the" short run as a period within which some inputs are variable, others fixed. Clearly there are many short runs, and the number of freely variable productive services increases as the period of time is lengthened.

The rate at which a firm expands its use of "fixed" factors depends not only on the cost of rapid change but also upon how long output is expected to run at a high or low rate. Suppose a firm has a "plant" (fixed factors) appropriate to a rate of production of 100 units of output per week. (The determination of the right amount of plant is taken up in the next chapter.) If now 130 units is the desired output (due to a rise in price), the firm will immediately begin to increase its plant if this new rate of output is expected to last for years. But if it is a short term fluctuation, which will probably be followed by an output rate of 70, it will be supplied only by varying the use of variable factors (and probably by inventory changes). In general no variation in plant size will be made if the fluctuation in output is expected to be temporary. In addition, even a permanent change in output, if it comes unexpectedly, will for a time be handled primarily through changes in the "variable" productive services.

The cost curve appropriate to these temporary changes in output is the short run marginal cost. We may prove the primacy of marginal cost from first principles. The cost of any action (such as increasing output 10 units) is the alternative use of the resources required to achieve this action. The short run fluctuations of output by definition involve no change in the "plant" (fixed factors), so there is no foregone alternative to the more intensive use of the plant.[6] The only foregone alternative is the amount spent on additional units of variable services.

[6] If the plant will wear out faster at higher rates of output, the extra cost is chargeable to the increased output. This cost (called user cost) is usually minor.

The definition of marginal cost is

$$MC = \frac{\text{Increase in Total Cost}}{\text{Increase in Output}}.$$

Since the increase in total cost is equal to the increase in the number of units of variable services times their price (which is constant to the firm under competition), we may rewrite this definition as

$$MC = \frac{\text{Increase in Quantity of Variable Services}}{\text{Increase in Output}} \times \text{Price of Variable Services}$$

$$= \frac{\text{Price of Variable Services}}{\text{Marginal Product of Variable Services}}.$$

Hence marginal cost varies inversely to marginal product, and the law of diminishing marginal product is equivalent (under competition) to the law of increasing marginal cost.

For reasons which do not bear close scrutiny, it is conventional to define a considerable variety of short run cost curves for the competitive firm. They may be illustrated with the arithmetic in Table 7–4, which is based upon the production schedule in Table 7–1, plus the assumption that units of the variable service cost $5 and units of the constant service $4. The definitions of the various costs are:

1. Total fixed cost = quantity of the fixed productive service times its price.
2. Total variable cost = quantity of the variable productive service times its price.
3. Total cost = total fixed cost plus total variable cost.
4. Marginal cost = increase in total cost divided by the increase in output.
5. Average fixed cost = total fixed cost divided by output.
6. Average variable cost = total variable cost divided by output.
7. Average cost = average fixed cost plus average variable cost = total cost divided by output.

The last four curves are illustrated in Figure 7-4.

We have said that only the marginal cost curve is relevant to short run changes in output: we can go a step farther and say that only that portion of the marginal cost curve *above* average

Table 7–4

UNITS OF VARIABLE SERVICE	UNITS OF FIXED SERVICE	TOTAL PRODUCT (= OUTPUT)	TOTAL VARIABLE COST	TOTAL FIXED COST	TOTAL COST	AVERAGE VARIABLE COST	AVERAGE FIXED COST	AVERAGE COST	MARGINAL COST
0	10	0	0	$40	$40	—	∞	∞	—
1	10	5	5	40	45	$1.00	$8.00	$9.00	$1.00
2	10	13	10	40	50	0.77	3.08	3.85	0.62
3	10	23	15	40	55	0.65	1.74	2.39	0.50
4	10	38	20	40	60	0.53	1.05	1.58	0.33
5	10	50	25	40	65	0.50	0.80	1.30	0.42
6	10	60	30	40	70	0.50	0.67	1.17	0.50
7	10	68	35	40	75	0.51	0.59	1.10	0.62
8	10	75	40	40	80	0.53	0.53	1.07	0.71
9	10	81	45	40	85	0.56	0.49	1.05	0.83
10	10	86	50	40	90	0.58	0.47	1.05	1.00
11	10	89	55	40	95	0.62	0.45	1.07	1.67
12	10	91	60	40	100	0.66	0.44	1.10	2.50
13	10	92	65	40	105	0.71	0.43	1.14	5.00
14	10	92	70	40	110	0.76	0.43	1.20	∞

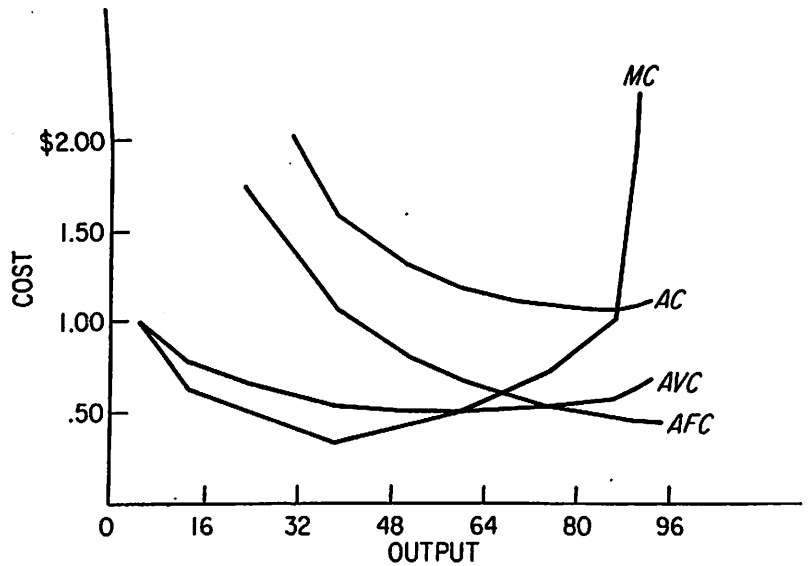


Figure 7-4

variable cost is relevant. To show this, we must first show that a competitive firm will operate where marginal cost equals price. It operates at this output because profits are then maximized. The demand curve of a competitive firm is a horizontal line: its output is too small to affect the market price. Hence, when the firm increases output by one unit, it increases

1. Receipts by the price of the unit.
2. Costs by the marginal cost of the unit.

Hence profits will rise after a unit increase in output if price exceeds marginal cost; and profits will rise after a *decrease* of a unit in output if price is less than marginal cost.[7]

[7] The rule may be derived algebraically. When output rises by Δq, profits rise by

$$p\Delta q - [C(q + \Delta q) - C(q)],$$

where $C(q)$ is the cost of producing q. If profits are at a maximum, they will not either increase or decrease with a small change in output, so this expression must equal zero. Rewriting it,

$$p = \frac{C(q + \Delta q) - C(q)}{\Delta q},$$

and the expression on the right is of course marginal cost.

We illustrate this rule in Figure 7-5. When the price is P_1, if the firm expands its output from X_0 to A, it will add RST more to costs than to receipts; if it contracts output to B, it will reduce receipts by RMN more than it reduces costs. (Recall that the area under a marginal curve between two points is the change in the total between these points.)

But if price falls below P_0, the firm faces a different choice. When price is P_2, if the firm operates at X_1 (where marginal cost equals

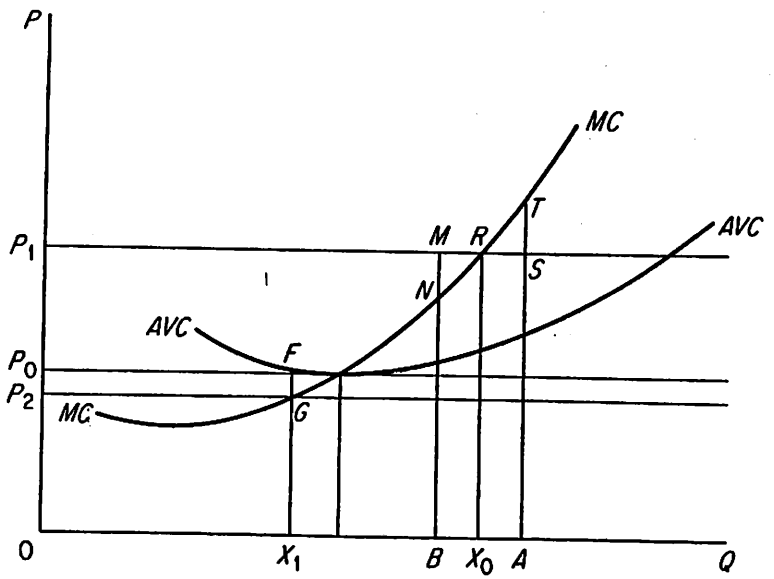


Figure 7-5

price) it will have total variable costs of $0X_1$ times X_1F, which exceed the receipts ($0X_1$ times P_2). By closing down the plant temporarily (recall that short run curves are appropriate only to temporary fluctuations), it will save money. Hence the firm will not operate below a price of P_0.

We define the supply curve of a competitive firm as the amounts it will supply at various prices. This supply curve is (in the short run) the firm's marginal cost curve above minimum average variable cost.

The Suspicious Character of Average Costs

Four cost curves were presented in Figure 7-4: average fixed, average variable, average and marginal costs. Average fixed cost is wholly uninteresting: it is the cost (per unit of time) of the "fixed" factors divided by output. It is always a rectangular hyperbola, and it is always useless. Average variable cost, we found, had one use: to determine the minimum effective point on the marginal cost curve; otherwise it too is dispensable.

Average cost is rather more popular in economics, and deserves fuller—but not necessarily kinder—treatment. The problem it poses is simply this: it cannot be trusted to stay put. Suppose a firm is making very handsome profits or losses on the usual average cost calculations: price is well above average cost or well below it, where average cost of course includes interest at the going rate on investment.[8] Suppose further that the profits or losses will persist for a considerable time. We claim that there will be a tendency for average costs to rise or fall to where they equal price.

To understand this shiftiness of average costs, let us ask why this competitive firm makes an unusually large or small rate of return on its investment for a considerable period of time. The answer must be that it has superior resources (including possibly management) so its costs are comparatively low, or inferior resources, so its costs are comparatively high. But then these superior resources are really worth more, and the inferior resources less, than the values at which they are carried on the books. If the resources are owned by the firm (say, a piece of land), there may be no tendency to write up the value of a superior resource to its true value, because accountants are conservative. On the other hand, the accountants will not object strongly to writing down the value of the inferior resource.

Whether the resources are revalued or not, another factor leads to movements of average costs. If the firm is sold, its price will be determined by its expected earnings. If these earnings are high, the firm will sell for more than book value, and if earnings are low it will sell for less than book value. If the buyer values the

[8] The "of course" should not lead the student to believe that it will be included in usual accounting procedures; accountants have been unwilling to include interest on investment (other than interest on debt) in cost.

enterprise at its cost to him, then by definition it will earn the going rate of return—average cost will move to equality with price.

If a firm used no specialized resources, the valuation of inputs would be much simpler, for then by definition the alternative product of a resource would be its cost to the firm (and industry). Once specialized resources enter, however, there is no valid basis for fixing their value other than discounting their future earnings—and average cost begins to follow price.

Revaluations of assets will not affect marginal costs because the revaluations do not depend upon the firm's output. Suppose there is a rise in the industry's output because of an increase in demand, so a given superior resource (say, a piece of land) should be cultivated more intensively for maximum profit. If the plot is cultivated more intensively, it will have a larger marginal product (by the law of diminishing returns; see MP_L in Figure 7-1), and should be revalued upward. But even if the owner of the plot mistakenly failed to use it more intensively, its value would rise—for the value of an asset is determined by what others would pay for it. Hence the asset becomes more valuable whether or not its owner varies output.

The actual amount of asset revaluations is unfortunately almost completely unknown.[9] The effects of restraints imposed by accountants and tax laws are in the direction of preserving historical costs (costs as historically made and recorded). Historical costs, if rigorously adhered to, eliminate certain methods of capitalizing gains and losses, but introduce other departures from the alternative cost concept appropriate to maximum profit behavior.[10]

The Proof of Rising Short-Run Marginal Costs

We have pointed out that marginal cost varies inversely with the marginal product of the variable factor, so the law of diminishing returns implies that the short-run marginal cost curve has a

[9] The most extensive study is Solomon Fabricant's *Capital Consumption and Adjustment* (National Bureau of Economic Research, 1938), esp. Ch. 12. Of 272 corporations reporting during 1925–34, 66 made capital write-ups, 140 capital write-downs—the period was obviously dominated by the Great Depression.

[10] Many of the problems encountered in analyzing historical costs are dealt with in the literature on national income and wealth.

positive slope. It would appear that this ends the matter of proof, but it does not.

A series of statistical studies have found that short-run marginal cost is approximately constant until "capacity" is approached. Capacity in turn is usually defined as the output at which marginal costs become very inelastic.[11] The typical marginal cost curve, according to this literature, is that illustrated in Figure 7-6. Clearly

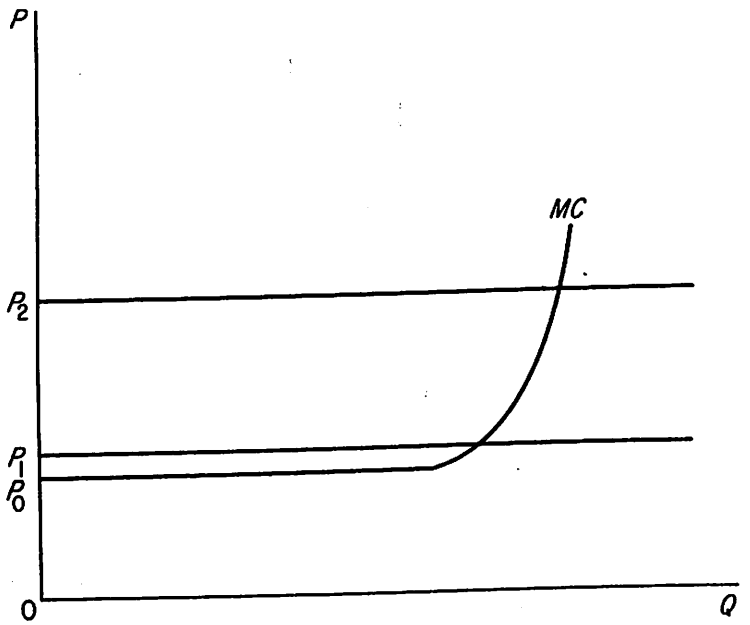


Figure 7-6

this literature denies the short-run validity of the law of diminishing returns.[12]

Rather than delve into the statistical studies which yield horizontal short-run marginal costs,[13] it is possible to test the validity

[11] We shall quarrel with this definition later.

[12] Numerous examples and references are given in J. Johnston, *Statistical Cost Analysis* (New York: McGraw-Hill, 1960).

[13] The studies have been criticized as having linear biases in the statistical procedures, and defended against this charge, with no clear victory for either side.

of this cost curve indirectly. If marginal costs are essentially constant up to the output at which they rise rapidly, under competition a firm's output (set where marginal cost equals price) will be nearly constant at all prices above this constant marginal cost, and zero at lower prices. Thus in Figure 7-6, the output of the firm varies little at prices between P_1 and P_2, but falls to zero at prices under P_0. Where marginal costs display this behavior, then, short-run variations in the output of the industry will come about almost exclusively through variations in the *number* of plants in operation. But if marginal costs rise steadily with output, much of the industry's fluctuation in output will come from fluctuations in the rate of output of each plant, and little from fluctuations in the number of operating plants.[14]

In this form, the hypothesis that short-run marginal costs are constant can be tested against readily observable facts. As an example, consider the American cotton spinning industry. The output of the industry may be measured by spindlehours, the "plant" by active spindles, and the output per plant by "hours per active spindle". Then the percentage change in output (spindle hours) from one quarter year to the next will be approximately equal to the sum of the percentage changes in active spindles and in hours per spindle.[15] This calculation has been made by quarters from August 1945 through June 1959, separately for the southern states (where the industry has grown slightly) and for the New England states

[14] The argument, it may be noted, can be extended also to noncompetitive firms which operate more than one plant within a market area. Unless the plants have equal constant marginal costs, a monopolist will minimize costs by operating the lower marginal cost plant at "capacity" and making all adaptations to changing output in the plant with higher marginal costs.

[15] The output of an industry is $Q = Nq$, where N is the number of plants operating, q the output per plant. By definition,

$$\Delta Q = N\Delta q + q\Delta N$$

and

$$\frac{\Delta Q}{Q} = \frac{\Delta q}{q} + \frac{\Delta N}{N}.$$

Hence the relative change in output is equal to the sum of the relative changes in N and q—the magnitudes used in our test. For large changes in output, a term $\Delta q \Delta N/Q$ should be added, and (as is customary with such formal partitions) divided arbitrarily between N and q. Here the cross product term is neglected.

(where the industry has been declining very substantially). We may tabulate the average of the 55 quarterly changes:[16]

Section	PER CENT OF CHANGE IN SPINDLE HOURS DUE TO	
	Change in Active Spindles	*Change in Hours per Spindle*
Southern states	9.2	90.5
New England	21.8	76.5

The conclusion is clear: even in the declining branch of the industry the overwhelming part of changes in output is achieved through variations in the rate of operation of plants (here, hours per spindle), not by variations in number of active plants (here, active spindles).[17] In this industry short-run marginal costs are rising: I suspect that in most industries they do so.

RECOMMENDED READINGS

Friedman, M., *Price Theory*, Chicago: Aldine, 1962, Chs. 5, 6.

Stigler, G. J., "Production and Distribution in the Short Run," *Journal of Political Economy*, 47 (1939), 305–27. Reprinted in *Readings in Income Distribution*.

Viner, J., "Cost Curves and Supply Curves," *Zeitschrift für Nationalökonomie*, 3 (1932), 23–46. Reprinted in *Readings in Price Theory*.

PROBLEMS

1. A producer with two plants wishes to produce a given output at the lowest possible cost. Under what conditions will he close down one of the plants?

[16] The source of the data is: U. S. Bureau of the Census, "Cotton Production and Distribution," Bulletins 186, 189, 193, and 196 (Washington, D. C.: U. S. Government Printing Office).

[17] In the declining branch of the industry, one would expect a larger role for plant reductions simply because the industry is declining. And when short-run output changes are divided into increases and decreases, we find that the role of plant changes is more important in the case of declines of output:

New England States

Direction of Change in Quarterly Output	PERCENT OF CHANGE IN SPINDLE-HOURS DUE TO	
	Change in Active Spindles	*Change in Hours per Spindle*
Increase	20.3	82.2
Decrease	23.9	68.6

2. You are given the following production function:

INPUT OF *A*	0	1	2	3	4	5	6	7
OUTPUT	100	101	103	105	106.8	108.4	109.9	111.3

(a) Draw the marginal and average products of *A*.

(b) Draw the marginal and average products of *B* (the other productive factor). Ten units of *B* underly the foregoing schedule. Use the constant returns to scale equation (p. 133).

3. An economy consisting of farms has the unusual production function for each farm:

NUMBER OF MEN	MARGINAL PRODUCT
1	20
2	15
3	19
4	14
5	18
6	13
etc.	etc.

If you had 10 farms and 40 employees how would you allocate them among farms? If wages are $40, construct the marginal cost schedule of output.

4. The law of diminishing returns was originally stated as an historical law, that is, it asserted that the marginal product of labor on land would decline as population grew. If true, what would have happened to (1) aggregate agricultural land values, and (2) prices of farm products relative to manufactures, over long periods?

chapter eight

Production: Returns to Scale

No such sweeping generalization as the law of diminishing returns has been found for the relationship of output to inputs when all inputs are varied. We are accordingly driven to consider alternative possibilities: when all inputs are increased in a given proportion, output may increase in a greater or lesser or equal proportion. The economist must then determine, when he is analyzing the automobile or shoe or radio repair industry, whether it has increasing, decreasing, or constant returns to scale, and we shall discuss later the methods of empirically determining economies of scale.

THE PROPER COMBINATION OF INPUTS

Let us begin by asking a basic question: if we wish to produce at a certain rate, in what proportion shall we use the various inputs? This question is not answered directly by the law of diminishing returns, for it told us only how many men were needed to produce a given product, given that they worked on 10 acres of land, or (since the law is reversible) how many acres were needed, given a labor force of 8 men. There are many different combinations of inputs that will yield the desired product, and obviously the cheapest combination will maximize the producer's profits.

The cheapest combination obviously depends upon the relative prices of the inputs and in fact the least cost combination is given by the rule: a dollar's worth of any input should add as much product as a dollar's worth of any other input. For if a dollar's worth of input A has a marginal product of (say) 5 units, and

that of B only 3 units, then we can

(a) Buy \$1 less of B, suffering a decline of product of 3 units,
(b) Buy \$0.60 more of A, obtaining 3/5 of the marginal product of a dollar's worth, or 3 units of product, and
(c) Pocket the \$0.40.

This rule may be stated as an equation of minimum cost:

$$\frac{\text{Marginal Product of } A}{\text{Price of } A} = \frac{\text{Marginal Product of } B}{\text{Price of } B} = \frac{\text{Marginal Product of } C}{\text{Price of } C},$$

for all inputs, no matter how many.

When the price of one input increases, this rule of minimum cost tells us that we must use less of this input (thus increasing its marginal product) and more of the other inputs (thus decreasing their marginal products).

This analysis has obvious analogies to the problem of the consumer dividing his income among commodities in order to maximize satisfaction. In fact the same apparatus of indifference curves can be used, with the obvious modification that now we shall call such curves isoquants (equal quantities), and define the isoquant (Figure 8-1) as those combinations of inputs which yield the same product. When we reduce the quantity of one input (A) by a small amount (ΔA), we reduce the product by ΔA times the marginal product of A ($= MP_a$). Thus if the marginal product of men is 6, when we reduce the quantity of labor by 0.25 (one-fourth of a day, say), we reduce the total product by $0.25 \times 6 = 1.5$. In order to offset this reduction, we must increase the other input (B) by such an amount (ΔB) as to produce this much, so

$$\Delta A \cdot MP_a + \Delta B \cdot MP_b = 0 \qquad (\Delta A < 0),$$

along an isoquant. Hence the slope of an isoquant is

$$\frac{\Delta B}{\Delta A} = -\frac{MP_a}{MP_b}.$$

Corresponding to the consumer's budget line, there will be an outlay line for the entrepreneur. With a given expenditure E_0, he

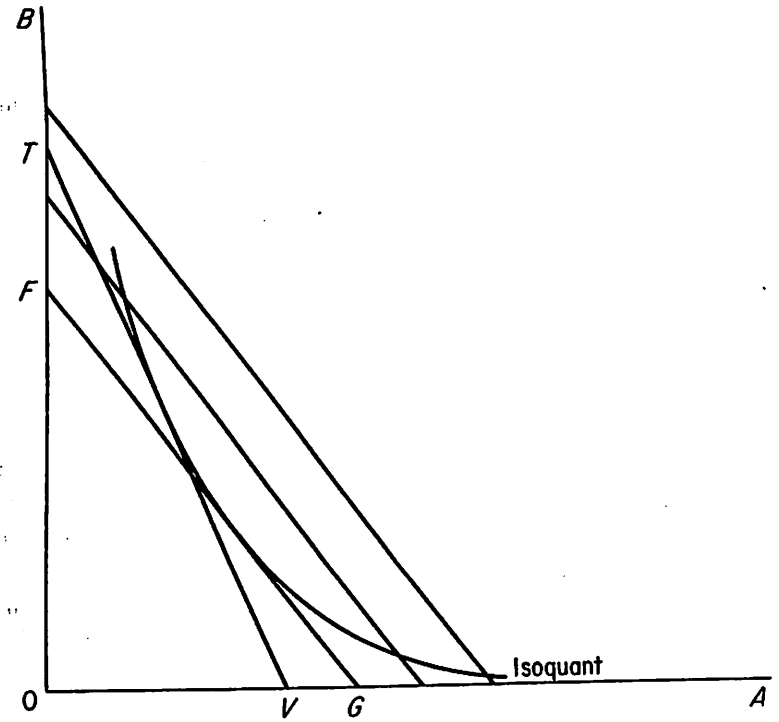


Figure 8-1

can buy all combinations of A and B such that

$$E_0 = AP_a + BP_b,$$

and there will be a different outlay line for every amount of expenditure. We draw three outlay lines in Figure 8-1 (temporarily ignoring TV), each higher one representing a larger total expenditure. The entrepreneur will choose outlay line FG because this is the lowest line which touches the isoquant, and in general the lowest outlay line to yield the desired product is tangent to the isoquant. But the slope of the outlay line is[1]

$$-\frac{P_a}{P_b},$$

[1] For a constant outlay,

$$\Delta A \cdot P_a + \Delta B \cdot P_b = 0,$$

$$\frac{\Delta B}{\Delta A} = -\frac{P_a}{P_b}.$$

and the tangency implies that

$$\frac{P_a}{P_b} = \frac{MP_a}{MP_b},$$

another form of our condition for minimum cost. The proposition that less will be used of an input if its price rises is illustrated by increasing the price of A, leading to the new outlay curve, TV, which is necessarily tangent to a convex isoquant to the left of the original equilibrium.[2]

The student will find many different uses of this technique, which is generally employed where one wishes to analyze three variables without recourse to solid geometry. (Here the three variables are two inputs and output; with consumer indifference curves they were two commodities and utility.)

CONSTANT RETURNS TO SCALE: THE SIMPLEST CASE

The simplest possibility with respect to economies of scale is that there are none: when output is increased in any proportion, exactly equal proportionate increases of all inputs are required. Then if the prices of productive factors are not affected by the firm's rate of output, as they will not be under competition, total costs vary proportionately with output.

Constant returns to scale are commended upon a very simple ground: if we do a thing once, we can do it twice. If we use A and B to produce P, why should not $2A$ and $2B$ produce $2P$? Perhaps they should, but it must be emphasized that there may be cheaper ways of producing $2P$. Where painting was done by hand, it may now be feasible to use a spray gun; where a man performed tasks X and Y, it may now be feasible to have him specialize in task X with a gain in efficiency. These are questions of fact, and we cannot state that in general they will be, or will not be, possible. If these examples suggest that *at most* we must double inputs to double output, that also is not true. For the tasks of coordinating a larger enterprise may increase so rapidly that large enterprises are inefficient. We shall examine these possibilities of increasing and decreasing returns shortly.

[2] See mathematical note 6 in Appendix B on the relationship of diminishing returns to convexity.

If there are constant returns to scale, obviously marginal costs will be constant for all outputs when the inputs are in proper proportion. For K per cent more output requires K per cent more of each input, and since the prices of productive factors are constant to a competitive firm, total costs also rise by K per cent. Hence average and marginal costs are constant.

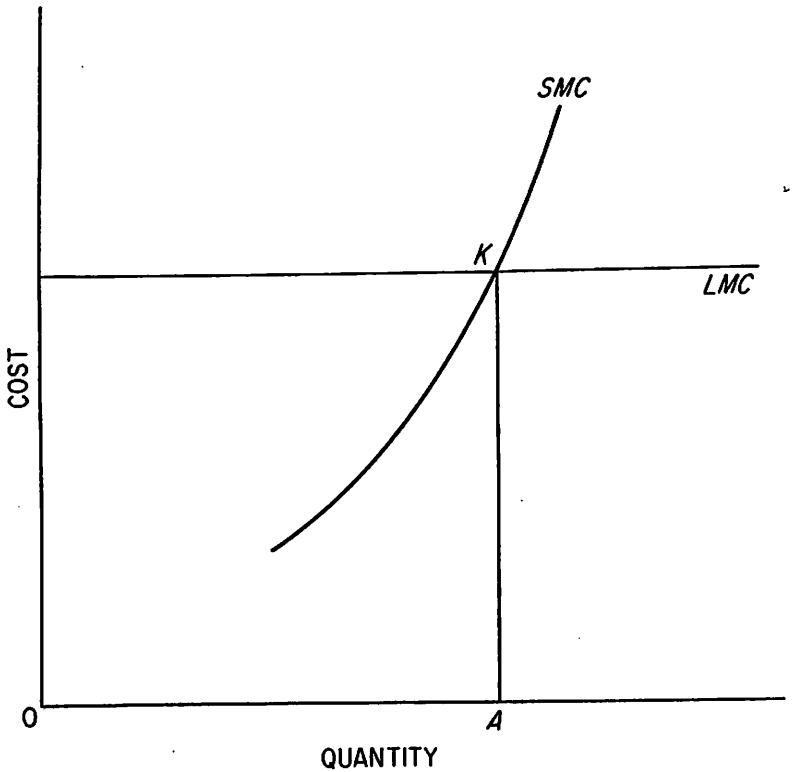


Figure 8-2

We have identified cost curves which reflect complete adjustment of all inputs with the "long run", and those which reflect changes in part of the inputs with the "short run" (p. 134). Hence "short-run" marginal costs will rise (because of diminishing returns) even though "long-run" marginal costs are constant. Consider Figure 8-2. At output A, a certain quantity of each input yields minimum marginal (and average) cost, AK. If we vary part of the inputs, we shall obtain the marginal cost curve, SMC.

It is evident that short-run marginal costs may be greater or smaller than long-run marginal costs, depending upon the rate of output. At the point where the two curves intersect, the marginal product of the plant ("fixed factors") divided by its price is equal to the marginal product of the variable factor divided by its price. At larger outputs the marginal product of the plant rises and the marginal product of the variable factor falls, so the long run minimum cost condition is not fulfilled; the opposite relation holds below the intersection.

Even though the long-run average and marginal cost curves of the firm are horizontal (and in fact identical) under constant returns to scale, no such relation need hold for the industry. As we shall see in Chapter 10, the industry cost curves are affected also by changes in the prices of inputs (which are constant to the individual firm). Nevertheless, even for the industry constant returns to scale is the overwhelmingly popular assumption in scientific work. The so-called Cobb-Douglas function is

$$P = aC^{\alpha}L^{1-\alpha},$$

where P is product, C is capital, and L is labor. This production function yields constant returns to scale,[3] and it has an almost monopolistic position in economic literature. Its popularity is not due to its demonstrated validity as a description of actual production functions, however. Rather, it is used because (1) it yields diminishing returns to each productive factor separately, (2) it is simple to handle, being linear in logarithmic form, (3) in many investigations the precise nature of returns to scale is not very interesting, and constant returns is a convenient simplification, and (4) of a remarkable property of constant returns to scale which we must now mention.

Euler's Theorem. Euler's theorem on homogeneous functions is the august name attached to this final property of constant returns. The theorem is a simple one: it says that if there is constant returns

[3] If we vary each input in a given proportion, say changing C to (λC) and L to (λL), we get

$$a(\lambda C)^{\alpha}(\lambda L)^{1-\alpha} = a\lambda^{\alpha+1-\alpha}C^{\alpha}L^{1-\alpha} = a\lambda C^{\alpha}L^{1-\alpha} = \lambda P,$$

so the product increases in the same proportion.

to scale, then the total product is equal to the sum of the marginal products of the various inputs, each multiplied by the quantity of its input.[4] Thus if the production function is

$$P = f(A, B, C, \ldots)$$

and there is constant returns to scale,

$$P = A \cdot MP_a + B \cdot MP_b + C \cdot MP_c + \cdots.$$

Since the theorem has been in the mathematical books for two hundred years, we can assume its truth, and here present only an example. Consider the simple production function

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

which Paul Douglas believed to be descriptive of American manufacturing; here P, C, and L are product, capital, and labor, all in index number form. If $L = C = 200$,

$$P = 200^{1/4}200^{3/4} = 200.$$

Increase labor to 201, and the product rises to

$$P = 200^{1/4}201^{3/4} = 200.749.$$

If now C is increased to 201, with L held at 200,

$$P = 201^{1/4}200^{3/4} = 200.249.$$

Hence the marginal product of L is $200.749 - 200 = 0.749$, and that of capital is $200.249 - 200 = 0.249$. The sum of marginal products times quantities of factors is

$$200 \times 0.749 + 200 \times 0.249 = 199.60,$$

which is approximately what Euler's theorem asserts. The small discrepancy in product arises because we use finite increases in the inputs: the theorem holds strictly only for infinitesimal changes.

Euler's theorem entered economics in order to solve the problem whether, if each productive factor is paid at the rate of its marginal

[4] The definition of a homogeneous function of degree k is that if

$$P = f(A, B, C, \ldots),$$
$$\lambda^k P = f(\lambda A, \lambda B, \lambda C, \ldots),$$

where λ is any positive number. When k is unity, the function is homogeneous of the first degree, and this is our definition of constant returns to scale.

productivity, the total product would be sufficient and only sufficient. It was received with considerable hostility: Edgeworth remarked that "Justice is a perfect cube, said the ancient sage; and rational conduct is a homogeneous function, adds the modern savant." The modern savant, Philip Wicksteed by name, abandoned the argument, but the simplicity and manageability of the homogeneous functions have overcome any scruples on realism and they are immensely popular among economists to this day.

Variable Returns to Scale

Phrases such as "economies of mass production" testify to the widely held belief that as an enterprise expands its scale of operations, it will be able to reduce average costs. Popular beliefs are seldom a safe guide in economics, and here they are especially suspect. Laymen observe that more electricity (or transistor radios or electric dishwashers) are made than formerly, and that prices have fallen (or, in a period of inflation, risen less than a comprehensive price index). These observations are correct, but the passage of time also allows technological advances to take place, so the effects of scale of operations and technological advance are not separated. Returns to scale (like diminishing returns) refer to the behavior of output relative to inputs when the "state of the arts" is given.

Increasing returns to scale arise when a doubling of output does not require a doubling of every input. The causes of increasing returns are:

1. There may be some unavoidable "excess capacity" of some inputs. A railroad has a tunnel which is essential for given traffic, but can handle twice as much traffic. The emphasis here is on "unavoidable." If the railroad has unused locomotives, in the long run they can be sold or worn out, and hence do *not* give rise to increasing returns.

2. Many inputs become cheaper when purchased on a larger scale. There are quantity discounts because of economies in larger transactions. Often equipment costs less per unit of capacity when larger sizes are ordered (see Table 8–1).[5]

[5] Containers have the property that their contents increase as the cube of dimensions, the surface (and material required) as the square.

Table 8–1

Prices of Ball-Bearing Induction Electric Motors, 1800 rpm (February 1950)

HORSEPOWER	PRICE	PRICE PER HORSEPOWER
1.0	$59	$59.00
1.5	69	46.00
2.0	80	40.00
3.0	89	29.67
5.0	106	21.20
7.5	139	18.53
10.0	176	17.60
25.0	327	13.08
50.0	559	11.18
100.0	1073	10.73
150.0	1633	10.89
200.0	2085	10.42
500.0	3207	6.41
1000.0	5819	5.82

3. More specialized processes (whether performed by men or machines) are often possible as the scale of operations increases: the man can become more expert on a smaller range of tasks; the machine can be special purpose.

4. The statistical laws of large numbers give rise to certain economies of scale. For example, the inventory of a firm need not increase in proportion to its sales, because there is greater stability in the aggregate behavior of a larger number of customers.[6]

If these forces are dominant, the long-run marginal cost curve of the firm will have a negative slope—there will be economies of scale. An illustrative long-run marginal cost and several short-run marginal cost curves are given in Figure 8-3: each short run curve represents a different amount of "fixed plant." The corresponding aver-

[6] See W. J. Baumol, "The Transactions Demand for Cash: An Inventory Theoretic Approach," *Quarterly Journal of Economics* (November 1952). A similar argument may be made with respect to risks of failure. See also the results on servicing of machines in W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed. (New York: Wiley, 1957), Vol. I, pp. 416–21.

age cost curves are also given in Figure 8-3. These average costs are exclusively alternative costs—the input prices are those necessary to keep the resources in this industry and exclude all "rents."

Decreasing returns to scale arises out of the difficulties of managing a large enterprise. The larger the enterprise, the more extensive and formal its administrative organization must be in order to pro-

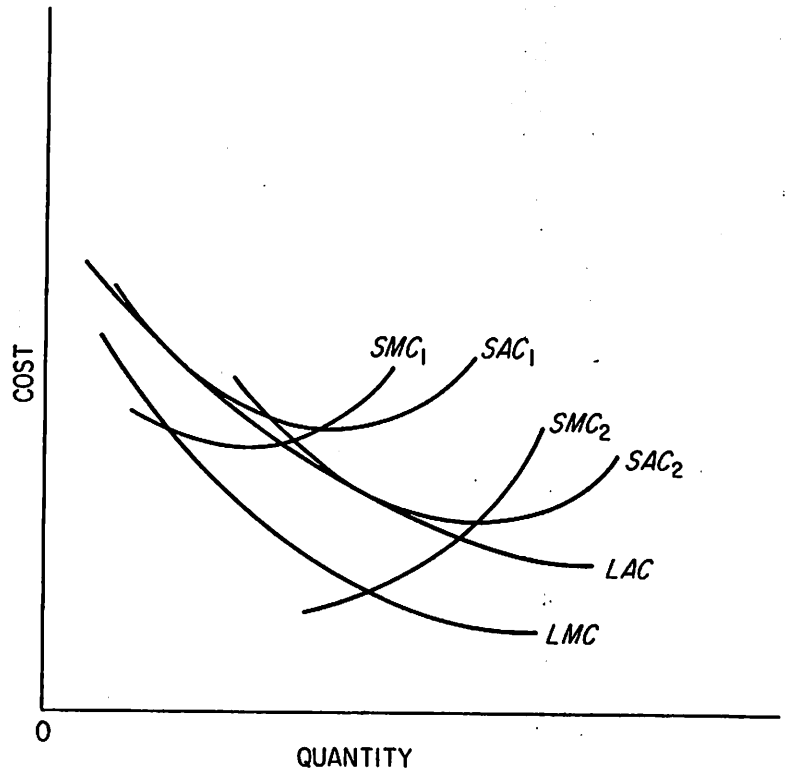


Figure 8-3

vide the information necessary for central decisions and the sanctions necessary to enforce these decisions. A large organization must be less flexible—policies cannot be changed frequently and still be carefully controlled.

The decentralization of a large organization might be considered a way in which to avoid the rigidity of size, and this has indeed become a fashionable practice at times. A fundamental contradic-

tion is encountered here, however: as the parts of a large enterprise are decentralized, the gains of economies of scale are simultaneously sacrificed. It would be possible to give each manager of a store complete autonomy, but then the organization which owned a thousand stores would become a mere investment trust: there could be no gains from quantity purchases or joint advertising.

This source of inefficiency of large size is given little weight in the popular literature: size is almost equated with efficiency. Yet anyone who watches a line of automobiles start forward as a traffic light changes will be impressed by how each additional driver starts a little later than his predecessor, so it takes considerable time for the motion to be communicated to the twentieth car, even when all the drivers can see the light change. This same slack is encountered in large organizations, so when frequent changes are called for, a large company is very inept. The industries making style goods (women's apparel and shoes, novelty toys, and so forth) are consistently dominated by smaller and more flexible companies. Again, those enterprises requiring very close coordination of skills of men are seldom large scale: no novel can be written by more than two persons (and of these at most one can be a woman), no orchestra can have 300 members and still be called symphonic. And in general intricate decisions cannot be made well by committees, which is the reason the greatest of industrial and political empires must have one head, whose familiarity with the details which underlie his decisions becomes vanishingly small.

Capacity. The notion of capacity is widely used, but seldom defined precisely. Yet it is an ambiguous concept even at best. In the normal case of variable proportions, the absolute maximum attainable output from a given set of fixed factors might be used—obviously a firm has no "capacity" limitation in the long run when all inputs can be increased. But the maximum attainable output is never known—it is, for example, the output of a farm or a factory when "no expense [or variable factor] is spared," and no one has been foolish enough to devote unlimited resources to this end.

Sometimes the technology of production seems to invite a fairly clean notion of capacity. For example, a blast furnace runs day and night, so it would appear to have a definite limit on output per month. Actually it does not: the charge can be varied; oxygen can be used, and the shut-down period can be shortened, so plants

have operated for considerable periods at more than 100 per cent of capacity. Yet the qualifications are minor, and in the short run "capacity" has a reasonable determinate meaning here. Such cases are uncommon.

It seems clear that economic capacity should be defined in a way that takes account of costs—no one cares about the output that could

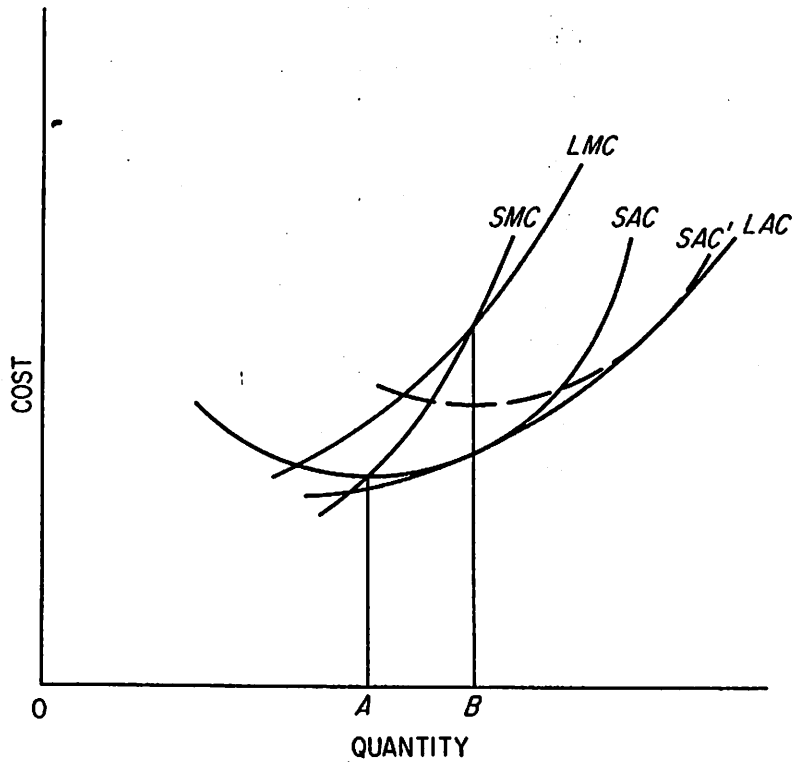


Figure 8-4

be obtained only at literally prohibitive costs. Two definitions have been proposed: capacity is (1) the output at which short-run average costs are at a minimum, and (2) the output at which short- and long-run marginal costs are equal. The definition of words is necessarily arbitrary, but there is a persuasive argument for the latter definition—it is more relevant to entrepreneurial decisions. We may illustrate its relevance by Figure 8-4.

On the minimum cost definition, capacity is OA; on the marginal cost definition it is OB. Suppose an entrepreneur with the plant represented by the short-run average cost curve wished to operate permanently at OB. On the minimum cost definition he is operating beyond capacity, and this suggests that, given time, he will build a larger plant. But he will not: this plant has the minimum average cost of any possible plant for output OB, and the larger plant denoted by SAC' would obviously have higher costs for the desired output. A definition which leads us to say a firm will willingly operate permanently beyond capacity seems undesirable. The definition of capacity in terms of the equality of short- and long-run marginal costs does not have this flaw: then it will always be true that in the long run a firm will expand if it wishes to maintain a rate of output which is beyond present plant capacity.

EMPIRICAL MEASURES OF ECONOMIES OF SCALE

When one looks at the size distribution of firms in a competitive (or, for that matter, noncompetitive) industry, he will always discover that a large variety of sizes exist at any one time. We may illustrate this variety with the corporate income tax data in Table 8–2. Assets are not an ideal measure of firm size, but they will do.

We observe that there is a considerable range of sizes of firms at any one time. This could be explained by the failure of some companies to reach the optimum size, due to errors of judgment or the time required to grow to the optimum size. But the range of sizes persists over a considerable period of time (a longer period, indeed, than our tables reveal). This persistence can only be explained by the fact that there is more than one optimum size.

The optimum size of a firm—we shall define "optimum" shortly—depends upon the resources which a firm uses. All firms in an industry do not have identical resources. Some have managers who are effective in running a small concern; others have managers who capably run a large concern. Some have large holdings of natural resources, others buy their raw materials. Some are located where labor is cheaper, others where electrical power is cheaper.

Table 8–2

Percentage Distribution of Assets by Company Size in Selected Manufacturing Industries (1954 and 1958)

Asset Size Class	*1954*	*1958*
A. Knitting mills		
Under $100,000	2.4%	3.4%
$100,000– 500,000	13.6	14.9
500,000– 1,000,000	12.0	13.1
1,000,000– 2,500,000	19.2	18.1
2,500,000– 5,000,000	14.8	16.3
5,000,000–10,000,000	13.7	12.5
10,000,000–25,000,000	12.8	19.4
Over 25,000,000	11.5	2.4
TOTAL	100.	100.1
B. Engines		
Under $500,000	.1	1.2
$500,000– 1,000,000	.6	.4
1,000,000– 2,500,000	2.4	1.6
2,500,000– 5,000,000	5.5	3.3
5,000,000– 10,000,000	6.1	0.
10,000,000– 25,000,000	32.5	20.2
25,000,000$ 50,000,000	35.1	31.1
50,000,000–100,000,000	16.6	42.3
TOTAL	99.9	100.1

SOURCE: *Statistics of Income*, 1954, 1958.

Such differences are compatible with all firms having equal long run marginal costs.[7]

If we observed the distribution of firms by size in an industry over a period of years and it did not change (random fluctuations aside), one could make several valid inferences. First, the firms of every size would on average be operating in a region of constant or rising long-run marginal costs—for if marginal costs were declin-

[7] The optimum size of firm is commonly defined as that which has minimum long-run average costs. As soon as we allow resources to differ, it is not possible to say that long-run average costs *excluding* rents will be equal for the different firms. The varying qualities and types of resources imply that some are specialized to the industry—that is, some resources will earn more in the industry than they could earn elsewhere. Average costs *including* rents can of course be equal.

ing to any size, these firms would expand and acquire a larger share of the industry's output. And second, the firms of various sizes would be equally efficient, because if any size were more efficient, this size would be more profitable and firms would tend either to move to this preferable size or to leave the industry.

In fact the basic definition of a firm of optimum size is that it can maintain itself indefinitely in competition with firms of other sizes. This test of optimality is all inclusive: it takes account of the ability of the firm, not merely to produce goods efficiently, but also to introduce new technology at the proper rate, cope with changes in consumer tastes, adapt to a changing geographical market in the product or resources, and so on. A test of comparative efficiency that is not all inclusive would not allow us to predict the survival of the most efficient size of firm.

Some sizes of firms decline as a share of the industry; for example, corporations with assets under $10,000,000 making engines had 14.7% of industry assets in 1954, only 6.5% in 1958 (Table 8–2). When the decline is large enough or persistent enough to overcome the possibility that it is due only to random fluctuations,[8] as is true in this case, we may conclude that these size classes are comparatively inefficient. On this interpretation, the large firms in the engine industry were more efficient than the smaller firms: there were economies of scale. In the knitting industry, on the contrary, there was a decline in the role of larger firms, so there were diseconomies of scale.

RECOMMENDED READINGS

Douglas, P. H., "Are There Laws of Production?" *American Economic Review* (March 1948).

Marshall, A., *Principles of Economics*, London: Macmillan, 1922, Bk. IV, Chs. 8–13; Bk. V, 3–5.

Robinson, E. A. G., *The Structure of Competitive Industry*, London: Nisbet, 1935.

Stigler, G. J., "The Economics of Scale," *Journal of Law and Economics*, 1 (1958).

[8] Random forces would be accidental events unrelated to the size of firm over a long period: floods or other catastrophes, an unusual number of deaths of entrepreneurs in a given period, unusual interruptions of supplies of materials, and so forth.

PROBLEMS

1. Prove that long-run and short-run marginal costs are equal where long- and short-run average cost curves are tangent.

2. Suppose a production process, contains three "machines": A, with a "capacity" of 20 units; B, with a capacity of 75 units; and C with a capacity of 210 units. Each machine has costs of $10 plus 10¢ per unit up to these limits of capacity, after which an additional machine must be employed. Calculate the average costs for outputs of 10, 20, and so on, up to several hundred units. Then determine minimum cost output. The problem of reconciling processes with different efficient sizes is called "balance of processes."

3. Using a Cobb-Douglas function, $P = C^{1/4}L^{3/4}$ calculate isoquants for $P = 100, 200, 300$. (For the first isoquant, since $P = 100$, $\log 100 = 2 = \frac{1}{4} \log C + \frac{3}{4} \log L$ and assign various values to C or L.) Draw some price lines tangent to these isoquants, $P_L = 1$ and $P_C = 2$. (Perhaps P_L is wage rate per hour and P_C rental cost of machinery per hour.) Calculate also the long-run average cost curve.

4. Statistical studies of costs of firms or plants of different size often commit the regression fallacy—which has already been encountered in the discussion of the consumption function. It yields economies of scale simply because of random fluctuation, even though there "really" is constant returns to scale. It may be illustrated as follows:

(a) Consider 10 firms, with average outputs of 100, 200, . . . , 1000, respectively.

(b) Each firm's costs in any one year are $5 per unit (variable costs) plus $5 times its average output. Thus the firm with an average output of 300 has costs of 300 × $5 = $1500 plus $5 times the output in the given year.

(c) Output in a given year consists of average output plus or minus a random fluctuation.

(d) The random fluctuation is obtained by flipping a coin, adding 10% of average output for each consecutive heads (if heads appear first) or subtracting 10% for each consecutive tails (if tails appear first). Terminate the flipping when the run of heads or tails ends.

Calculate the costs in a given year. Compare graphically with average costs when there are no random fluctuations in output.

chapter nine

Additional Topics in Production and Costs

The cost curves developed in the preceding chapter are those commonly used in economic analysis. Yet they deal with only a particular kind of production process, and there are many problems for which they require modification or extension. In this chapter we discuss three such extensions: multiple products; external economies; and finite production runs. Each is sufficiently important to deserve attention, and in the process more will be learned of the standard cost curves.

MULTIPLE PRODUCTS

Multiple products made their entrance into economic analysis in Great Britain, so the traditional example of multiple products has been the steer, which yielded a hide and beef. It is at least approximately true that these products are yielded in fixed proportions: a steer has only one hide. Hence if we attempt to construct a cost curve for (say) hides, we shall find that we cannot do so: we cannot vary the output of hides, holding the output of beef constant. The only possible cost function is that for a composite unit of (hides and beef), and given competition, it will be a matter of indifference to producers whether hides sell for \$20 and carcasses for \$1, or hides sell for \$1 and carcasses sell for \$20. Demand conditions will determine relative prices.

The case of multiple products produced in fixed proportions is, in fact, really not a case of multiple products so far as production is concerned. In a cost diagram, we may relabel the output axis $(A + B)$, and now employ the cost curves of the single product

firm. There is no difference between calling (beef and hide) one product and calling H_2O water.

As a general rule, however, the products of a firm can be produced in variable proportions. This is obviously true in many cases: a department store can sell more or less of any one product; a shoe factory can make more or less of one kind of shoe; a farmer (the nation's agricultural policy permitting) can grow more soybeans and less wheat. Variability is also possible in many more subtle cases: a petroleum refinery can vary the proportion of crude oil distilled into gasoline. Conversely, what looks to be independent productive activities—a firm produces steel and cement in different plants—may be related by some common element: for example, the cost of raising capital for the cement plant will probably be lower, the larger the steel plant.

When the proportions among the products are variable, it is possible to derive a separate marginal cost for each product. Consider the hypothetical data for a petroleum refinery in Table 9–1. We

Table 9–1

OUTPUT OF FUEL OIL	OUTPUT OF GASOLINE (GALLONS) *100*	*110*	*120*	*130*
100	\$2.45	\$3.55	\$4.85	\$6.35
110	3.90	4.80	5.90	7.20
120	5.45	6.15	7.05	8.15
130	7.10	7.60	8.30	9.20

define the marginal cost of gasoline as the increase in total cost divided by an increase in the output of gasoline, the quantity of fuel oil being held constant. For example, the marginal cost of 110 gallons of gasoline, when the output of fuel oil is 120 gallons, is

$$\frac{\$6.15 - \$5.45}{10} = \$0.07 \text{ per gallon.}$$

There will be a marginal cost curve for gasoline, or in general for any one product, corresponding to each possible output of the other product or products. This poses no real problem in the theory: we can simply write (in the competitive case)

$$MC_G(G, F) = P_G,$$

that is, that at equilibrium the price of gasoline will equal its marginal cost, which depends upon the quantities of gasoline (G) and fuel oil (F) produced, and similarly for fuel oil:

$$MC_F(G, F) = P_F.$$

The two equations can then be solved simultaneously.

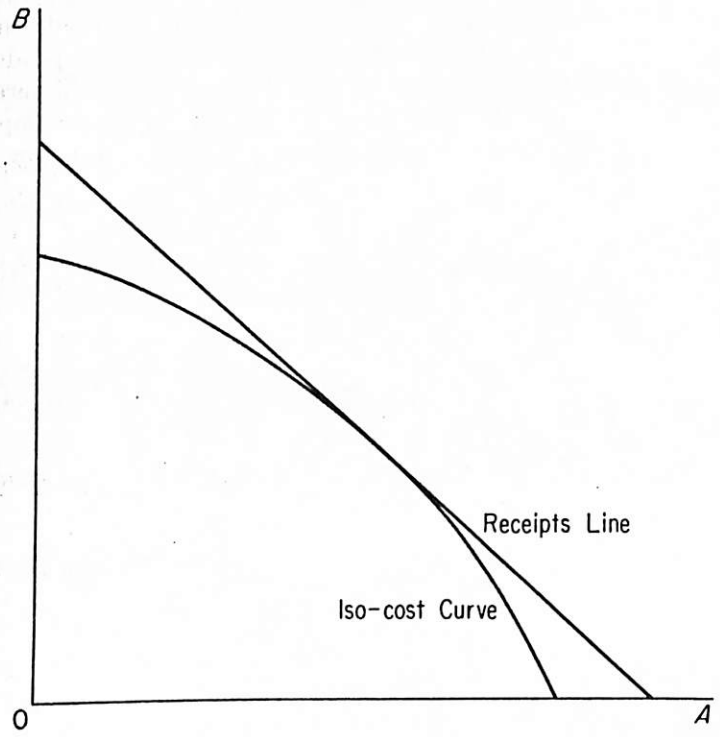


Figure 9-1

An equivalent geometrical procedure is to construct indifference curves (called isocost curves), which represent the quantities of the products which can be produced at a given total outlay. We display one such isocost curve in Figure 9-1. It is concave to the origin because as one continues to substitute one product for the other in the production process, smaller amounts of the other are obtained for given decreases of the one—marginal costs of each

product are rising.[1] Under competition the receipts from the sale of the products by a firm can be drawn in the diagram as a straight line: receipts are $Ap_a + Bp_b$, and prices are constant. The firm will operate where the isocost line touches (is tangent to) the highest possible receipts line, and this is equivalent to equating marginal cost and price.[2]

There is no corresponding possibility of calculating the average cost of one of several products. It is worth noticing that even though impossible, it is done every day. The costs which are common to several products—a machine or raw material used in producing both, an executive who manages the production of both—are often divided among the products in proportion to their separable variable costs, or in proportion to their sales. Such an allocation must be arbitrary, for there is no one basis of allocation that is more persuasive than others. Indeed *any* allocation of common costs to one product is irrational if it affects the amount of the product produced, for the firm should produce the product if its price is at least equal to its minimum marginal cost.

EXTERNAL ECONOMIES

An external economy is a source of reduction in cost which is beyond control of the firm. One firm in a competitive industry has no influence upon the prices of inputs, so if their prices fall as the industry expands, this is an external economy. Conversely, if input prices rise as the industry expands, the rise in cost of a firm represents an external diseconomy. The external factors may work upon coefficients of production as well as on input prices: for example the growth of traffic congestion in a community may force a firm to use more trucks to deliver a given quantity of goods.

[1] The argument of mathematical note 6 in Appendix B is applicable with changes of language.

[2] The slope of the price line is

$$\frac{\Delta B}{\Delta A} = -\frac{p_a}{p_b}.$$

An isocost curve is given by $\Delta A \cdot MC_a + \Delta B \cdot MC_b = 0$, or

$$\frac{\Delta B}{\Delta A} = -\frac{MC_a}{MC_b}.$$

Cost Curves for Industry-wide Output Changes

The cost curves of a firm presented in Chapters 7 and 8 were constructed on the assumption that the firm has no influence upon the prices of the factors of production it uses.[3] Under competition this is of course (by definition) the proper assumption. But when all the firms in a competitive industry simultaneously increase or decrease output, their aggregate effect is often to change the prices of inputs. Since we are normally interested much more in the behavior of the industry than of the firm, it is desirable to have cost curves which take account of the impact of the industry's rate of output on input prices.

The direct method of dealing with this dependence of the costs of one firm on the rate of output of the industry is to draw a different cost curve for the firm for each possible price of productive services. For example, when the price of the product is $0A$ and the output of the firm $0T$, the price of raw materials may be \$1 a pound, and the firm's marginal cost curve M_1 (Figure 9-2). When the price of the product is $0B$ and the output of the firm $0R$, the price of the raw material may be \$2 and the marginal cost curve of the firm M_2. Let us join points like T_1 and T_2 (and the innumerable other points we could find for other prices of the raw material) and label the curve M. Then M_1 and M_2 are the type of cost curves derived in preceding chapters, and M is the type of cost curve which we wish to employ in many areas. The distinction between the two types of marginal cost curves is clear:

> M_1 (or M_2) is the marginal cost curve when the prices of productive services are constant.
>
> M is the marginal cost curve when all firms in the industry are varying their rate of operation so marginal cost equals price.

Let us call the latter type of curve marginal cost for industry-wide changes. We argued that marginal cost curves of type M_1 have a positive slope under competition in both short and long run. If

[3] Implicitly it was also assumed that variations in the industry's output did not affect the coefficients of production. Exactly the same technique which will be presented to include the effects of changes in input prices on the cost curves will also take account of changes in production coefficients.

this is true, marginal cost curves for industry-wide changes will also have positive slopes unless, when the industry expands, the prices of productive services fall, in which case these curves may (not must) have a negative slope (we discuss this case later).

It should be kept in mind that curves of type M_1, which might be called marginal costs for single-firm changes, are the only type

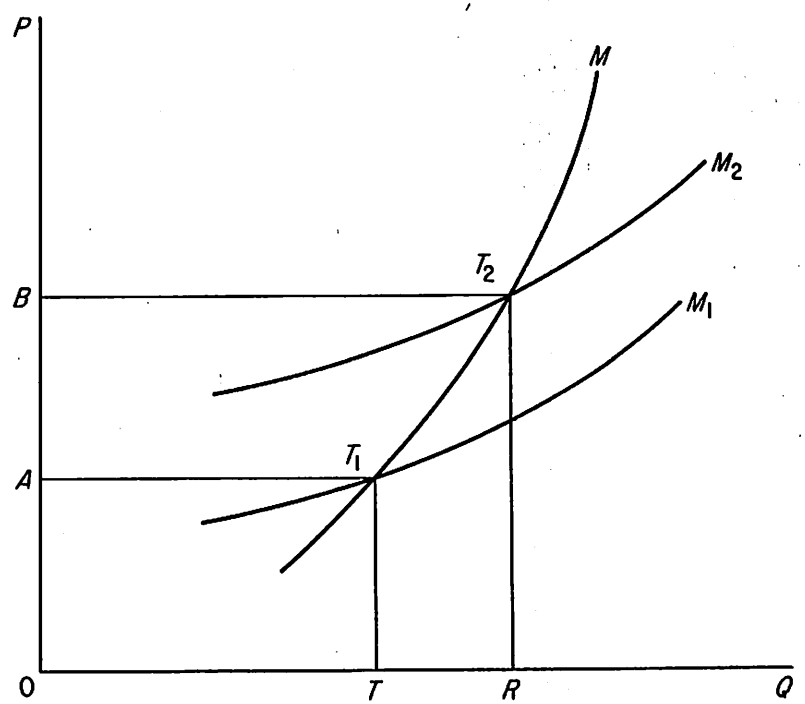


Figure 9-2

that the entrepreneur can individually move along: he cannot control the rate of output of the industry and thus the prices of productive services. The type M curves display the combined effects of the entrepreneur's selection of minimum-cost combinations of inputs (portrayed by M_1) and the repercussions on the firm of profit-maximizing behavior of other firms in the industry, over which the entrepreneur has no control. In this sense the type M curves are shorthand methods of describing the whole array of possible marginal cost curves of the firm (corresponding to all possible prices of pro-

ductive services), for they pick out the points (like T_1 and T_2) which are relevant to industry-wide changes.

The Functions of the Firm

The number of processes to which a raw material is subjected in its transformation into a finished consumer commodity is indeterminably large. We may, for example, distinguish the making of flour and the baking of bread, or we may distinguish the greasing of pans, the kneading of dough, or the lighting of ovens. The question arises: how are these functions divided up among firms? What determines whether retailing will be undertaken by manufacturers, or ore mining by steel companies, or credit extension by doctors?

A part of the answer lies in the technology employed. If letters are prepared on a typewriter, it would be extremely inconvenient to subcontract out the typing of the vowels. If an ingot must be reheated to be rolled, it is obviously more economical for the firm that cast the ingot to roll it while it is still hot.

But technology is usually not peremptory: there is often wide scope for variety in the ways productive processes are performed. The publisher of a book need not (and seldom does) print it; the printer seldom binds the book. Then a famous theorem of Adam Smith comes to our rescue: the division of labor is limited by the extent of the market.[4] Smith pointed out that small villages could not support highly specialized occupations, but that large cities could:

> In the lone houses and very small villages which are scattered about in so desert a country as the Highlands of Scotland, every farmer must be butcher, baker and brewer for his own family. In such situations we can scarce expect to find even a smith, a carpenter, or a mason, within less than twenty miles of another of the same trade. The scattered families that live at eight or ten miles distance from the nearest of them, must learn to perform themselves a great number of little pieces of work, for which, in more populous countries, they would call in the assistance of those workmen. Country workmen are almost every where obliged to apply themselves to all the different branches of industry that have so much affinity to one another as to be employed about the same sort of materials. A country carpenter deals in every sort of work that is made of wood: a country smith in every sort of work that is made of iron. The former is not only a carpenter, but a joiner, a cabinet maker, and even a carver in wood, as well as a wheelwright, a ploughwright, a cart and waggon maker. The employments of the latter are still more various. It is impossible there should be such a trade as even that of a nailer in the remote and inland parts of the Highlands of Scotland. Such a workman at the rate of a thousand nails a day, and three hundred working days in the year, will make three hundred thousand nails in the year. But in such a situation it would be impossible to dispose of one thousand, that is, of one day's work in the year.

[4] *The Wealth of Nations* (New York: Modern Library ed., 1937), pp. 17–21. I earnestly recommend that all of this book except p. 720 be read.

The gains from specialization operate in the same manner in a modern industrial society. As an industry grows, more and more activities are performed on a sufficient scale to permit firms to specialize in their full time performances: the making, and repairing, of machinery, the designing of plants, the testing of products, the recruiting of labor, the packaging of products, the collection of information on supplies, markets, and prices, the holding of trade fairs, research on technical problems, and so forth.

We may illustrate this development geometrically. Suppose the firm engages in three processes: processing raw materials (Y_1), assembling the product (Y_2), and selling the product (Y_3). For simplicity, assume that the cost of each function is independent of the rate of the other processes, and that the output of each process is proportional to the output of the final product.[5] The average cost of each function is shown separately, and the combined costs are the average cost of output for the firm (Figure 9-3). As we have drawn the figure, process Y_1 is subject to increasing returns, process Y_2 is subject to decreasing returns, and process Y_3 is subject first to increasing and then to decreasing returns. This situation may be perfectly stable in spite of the fact that the firm is performing function Y_1 at less than the most efficient rate and Y_2 at more than the most efficient rate.[6]

As the industry's output grows, the firms will seek to delegate decreasing and increasing cost functions to independent (auxiliary)

[5] This second assumption allows us to measure all processes along one axis; it has no effect on the argument.

[6] If the firm is a monopoly, it cannot specialize in process Y_1 and sell to other firms. It would be cheaper to buy Y_2 from several other firms than undertake it subject to decreasing returns, but if the costs of the other processes would be higher if Y_2 were not performed (contrary to the simplifying assumption in the text), Y_2 cannot be delegated.

industries. For example, when one component is made on a small scale it may be unprofitable to employ specialized machines and labor; when the industry grows, the individual firms will cease making this component on a small scale and a new firm will specialize in its production on a large scale. Thus, when the firm buys

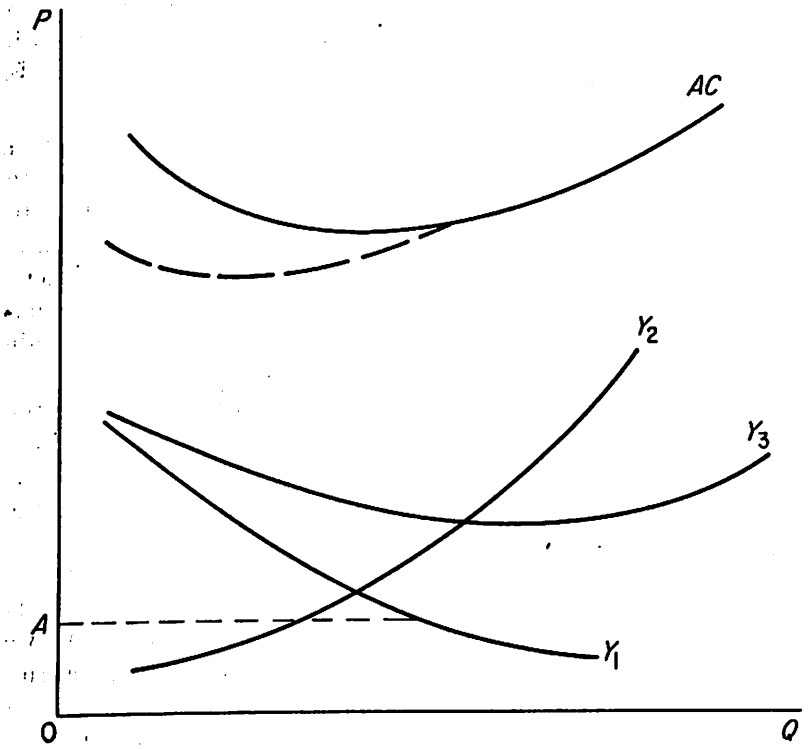


Figure 9-3

Y_1 at price $0A$, its average costs fall to the broken curves shown in Figure 9-3. Conversely, the firms will make only a part of the processes (Y_2) subject to increasing cost, and buy the remainder from independent firms.

A related explanation of the division of functions among firms is that those activities will be undertaken by a firm which are cheaper to administer internally than to purchase in the market. The transactions between firms are not free: there are costs attached to searching for prices, closing contracts, collecting payments, and so on.[7] Of course the coordination of activities within the firm is also not free: men and machines must be assigned tasks in an efficient manner and supervised to ensure that the efficient plan is followed. When a firm supplies only a part of its needs for some process (curve Y_2 in Figure 9-3) the rising costs of internal coordination are in fact the basic explanation for partial recourse to purchase. The cheaper market transactions become (due to improved knowledge of prices and greater security of contracts) the greater will be the comparative role of market coordination—firms will become more specialized.

Some external economies depend less on the growth of the industry than on that of the entire industrial system. As the economy grows, it becomes possible to establish a much more complete transportation system, a complex of types of banks and other financial institutions catering to specialized needs, an educational system that can train highly specialized personnel, and so on. These external economies are perhaps the decisive reason that the law of diminishing returns does not hold for an entire economy; it is highly probable that the American economy would be less productive if it were smaller.

FINITE PRODUCTION RUNS

The traditional laws of production are oriented to the problem of infinitely continued production: the farm will grow wheat this year, next year, and so on indefinitely. Many production decisions, however, involve a given volume or period of production. For example, the firm is to print 10,000 copies of a book, or produce 300 planes of a certain type; or, in the event of a fixed period, it is to supply (at a fixed annual rate) some item for 2 or 5 years.

The traditional theory does not directly cope with production for a finite run. For this theory is based upon continuous, unending flows of productive services, and under this condition it is a matter of minor detail whether the productive resources which yield the flows are durable or perishable: in either case they will be replaced when necessary. If the farm is to produce for only 10 years, how-

[7] See R. Coase, "The Nature of the Firm," *Economica* (1937); reprinted in Stigler and Boulding (eds.), *Readings in Price Theory*.

ever, and then be abandoned, it is clearly more efficient to use up the natural fertility of the soil than to maintain it. If only 5 units of a product are to be made, less specialized or less durable machinery will be used than if 500 units are to be made.

In the case of finite production runs, a theory of costs of great interest has been devised by Armen Alchian.[8] His analysis rests on the variation of total output (volume = V), the rate of production per period (q), and number of periods over which the item will be produced (m); in the simplest case these variables are connected by the equation, $V = mq$. Alchian has proposed a series of propositions concerning the behavior of total cost of the volume to be produced, of which the following are the most important:

1. The average and marginal cost per unit of total volume decreases as the total volume increases, holding the rate of production per unit of time constant.

Let the cost of a given total volume be the sum of discounted future expenditures. Then the proposition may be illustrated by the printing of a given book: once the plates have been made, additional copies (a given number per period) can be struck off at a relatively constant additional cost. The total cost (ignoring interest) will be approximately

$$\text{Composition Costs} + \text{Number of Copies} \times \text{Printing Costs per Copy}$$

so the average cost will be

$$\frac{\text{Composition Costs}}{\text{Number of Copies}} + \text{Printing Costs per Copy},$$

which decreases as the number of copies printed increases. There are usually some producers' goods which partake of the nature of stamping dies. In addition there are economies from "learning": as the length of the production run is extended (as it must be if V increases but q is held constant)—a variety of economies are uncovered by experience.

2. The marginal cost of output rises with the rate of output if volume is held constant.

[8] "Costs and Outputs," in *The Allocation of Economic Resources* (Palo Alto, Calif.: Stanford University Press, 1959); see also J. Hirshleifer, "The Firm's Cost Function: A Successful Reconstruction?" *Journal of Business* (July 1962).

If total volume is held constant, an increase in the rate of output per period implies a shortening of the number of production periods, so the proposition asserts that it is cheaper to produce a given number of units in (say) 4 years than in 3 years. This proposition is essentially an assertion of diminishing returns.

These marginal cost curves are illustrated in Figure 9-4. The marginal cost of volume is of special interest—it is the theoretical

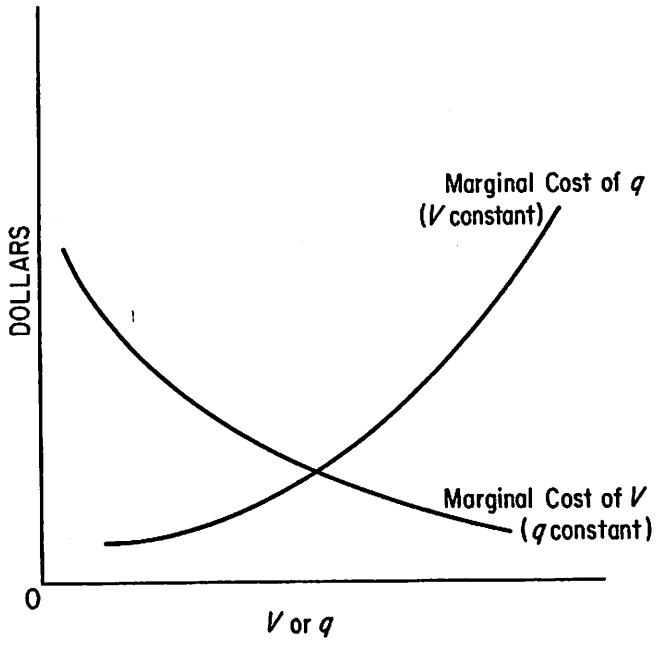


Figure 9-4

explanation for the almost universally observed phenomenon of quantity discounts. Whether we look at aggregate volume—the cost of a book of which 1,000 or 100,000 copies are printed—or at the size of an individual transaction—it costs less per copy to sell five copies than one—we find strong confirmation of the effects of volume on cost.

The relationship of marginal costs to aggregate volume has special relevance to the introduction of new commodities. These new commodities fall in price more rapidly through time than do the

prices of established goods, and the more rapid fall is due to the large increase in volume. Once the production of the commodity has achieved a substantial scale, these economies are exhausted and the traditional cost curves of infinite production runs become appropriate.

RECOMMENDED READINGS

Alchian, A., "Costs and Outputs," in *The Allocation of Economic Resources*, Palo Alto, Calif.: Stanford University Press, 1959, also A. Alchian and W. Allen, *University Economics*, Belmont, Cal.: Wadsworth Publishing Co., 1964, Ch. 21.

Coase, R., "The Nature of the Firm," *Economica* (1937); reprinted in Stigler and Boulding (eds.), *Readings in Price Theory*.

Stigler, G. J., "The Division of Labor is Limited by the Extent of the Market," *Journal of Political Economy*, 59 (June 1951), 185–93.

Young, A., "Increasing Returns and Economic Progress," *Economic Journal*, 38 (1928), 527–42.

PROBLEMS

1. An industry produces A and B in fixed proportions ($1A$ with $3B$). Average cost is constant at \$5 for $1A + 3B$. The demand functions are:

$$p_a = 48 - \frac{q_a}{10},$$

$$p_b = 60 - \frac{q_b}{3}.$$

Determine outputs and prices in long run equilibrium (assuming competition). Compare the effects of a tax of \$3 on A and \$1 on B.

2. Let total costs of producing A and B be

$$C = 10 + \frac{A}{2} + \frac{A^2}{10} + \frac{B}{5} + \frac{B^2}{25} + \frac{AB}{10}.$$

What is the marginal cost of 10 units of B when $A = 20$?

3. Construct the marginal cost curve for industry-wide changes from the production function in Table 7–1, the costs in Table 7–4, and the information that the price of the variable service is related to the purchases of the industry by the equation, $p_v = \$3 + Q_v/500$ and there are 100 firms. The marginal costs in Table 7–4 are then valid when the price is \$5, the purchases of Q_v are 1000, and the output of the industry is 8600.

4. (Due to A. C. Harberger.) Product X is produced by two factors of production, A and B. These factors must be used in fixed proportions, according to the recipe: $1A + 1B$ produces $1X$. The industry is competitive. Factor A has no use outside the industry, while factor B is so widely used outside the industry that the price of a unit of B is not influenced by variations in output in the X industry. The price of B is \$1. There are 1000 units of factor A, all of which are available at any price above \$0.50, none of which are available at a price below \$0.50. The demand curve for product X is $XP_x = \$2500$.

(a) What will be the equilibrium price and quantity of X?

(b) What will be the equilibrium price of factor A? of factor B?

(c) Suppose an excise tax of 20 per cent of the price to the consumer is imposed. What will be the price of X paid by the consumer? What will be the price received by the producer? How much X will be produced? What will be the price of factor A? of factor B?

(d) Suppose that a monopolist takes over industry X, and that he is assured that no entry will take place and no government will interfere with his operations, so long as he charges a single price for all the units of X he in fact delivers. What will be the price set by this monopolist? What will be the output of commodity X? What will be the price of factor A? of factor B?

chapter ten

The General Theory of Competitive Prices

Everyone knows that prices are set by supply and demand. A much smaller group, but one including careful readers of the preceding pages, knows what factors govern supply and demand. Our task is to gather these pieces of analysis and fit them into a general picture of the workings of competitive markets.

THE GENERAL PRINCIPLE

A competitive market must fulfill certain conditions if it is to be in equilibrium:

1. Each firm must be operating at the output which it deems most appropriate to the conditions of cost and demand.
2. The total quantity all firms wish to sell at the market price must equal the total quantity all buyers wish to purchase.

When these conditions are fulfilled, the price will be an equilibrium price—that is, it will have no tendency to change until supply or demand conditions change.

The first condition—an appropriate output of each firm—is in turn fulfilled when two conditions are met:

1. Each firm is in the industry which yields it largest profits.
2. Each firm is operating at the output where marginal cost equals price, which is the output which maximizes profits in this industry.

Quite clearly we are judging the "appropriateness" of an entrepreneur's decisions by whether they maximize his profits.

The extent to which the entrepreneurial behavior can be explained by efforts to maximize profits is a celebrated debating ground for economists.[1] We shall nevertheless use this assumption without extensive defense, and on two grounds. First, and most important, it yields a vast number of testable conclusions, and by and large these conclusions agree with observation. Second, no other well-defined goals have yet been developed and given empirical support.

These conditions of competitive equilibrium are readily translated into a diagram (Figure 10-1). For the firm the demand curve

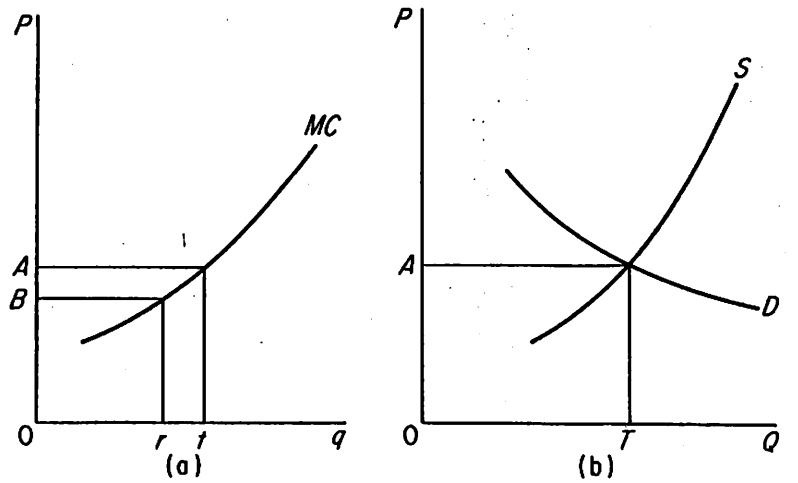


Figure 10-1

is a horizontal line, by our definition of competition that the firm be sufficiently small relative to the industry so variations in its output have a negligible influence on price. We may pause to notice that if our demand curve refers to this month (we shall soon look closely at the time dimensions), then the demand curve of the firm will be independent of next month's demand. Even if an unusually

[1] Even business men do not like this formulation. In one field study, when they were asked whether they maximized profits, they indignantly rejected the suggestion and pointed out that they were sincerely religious, public-spirited, and so on—as if these traits were inconsistent with profit-maximizing. But when the question was reformulated as: would a higher or lower price of the product yield larger profits?, the answer was, usually, no.

high price this month will lead to a reduction in industry demand next month, the individual firm cannot influence next month's price (say, by selling more cheaply now). Hence the demand curve of a competitive firm is independent of future conditions. Later we shall see that this is not true under monopoly.

The firm will operate at the output where its marginal cost curve intersects the demand curve. If we are examining (as we usually shall) forces that impinge on all firms in the industry, the marginal cost curve should be that which incorporates the effects of external economies—what we call the marginal cost for industry-wide changes (see pp. 166f.)[2] The firm operates at output $0t$ if price is $0A$, and at $0r$ if price is $0B$. The marginal cost curve thus traces out the firm's supply curve.

If we sum horizontally the marginal cost curves of the firms, we trace out the supply schedule of the industry (curve S). If there are 100 identical firms, then $0T = 100\ 0t$, and similarly for other outputs. The industry demand curve, D, is of course a conventional negative-sloping curve. The intersection of S and D establishes the equilibrium price.

This becomingly simple apparatus contains the essence of the theory of competitive prices. We can, and shall, clutter up the exposition in taking account of time periods, and of the entry and exit of firms from an industry, but the essence of the analysis will not change.

Two Normative Properties

Competitive prices are widely admired: by customers, for they connote the absence of monopoly power; by lawyers, since the antitrust laws are designed to achieve competition; and by economists. The economic advantages of a competitive price are two.

First, the division of output among firms is efficient in the sense that with no other division would the same output be so cheap to produce. Consider two firms which were not in competitive equilibrium (Figure 10-2). Firm 1 is operating at output $0b$, firm 2

[2] If a force were to impinge on only this one firm (say a tax on only this firm, or only this firm introducing a technological improvement) we should of course use the marginal cost curve for single firm changes in output.

at output $0d$. Clearly competitive equilibrium is lacking because the firms are not selling at the same price. If we reduced the output of firm 1 by ab, its costs would fall by $abmn$. If we increased the output of firm 2 by cd $(= ab)$, its costs would rise by $cdsr$. Clearly the costs of firm 1 would fall by more than those of firm 2 rose, so total costs of the two firms would decline for the given total

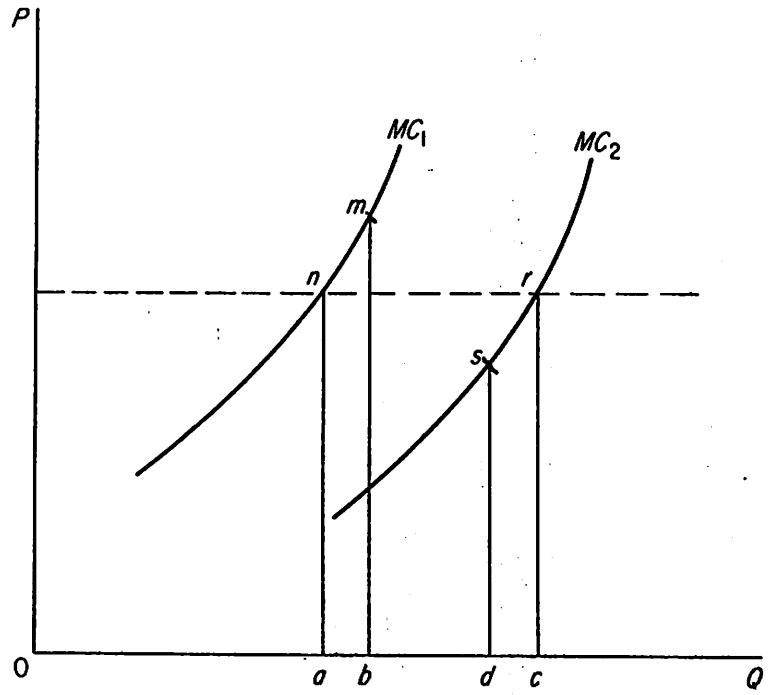


Figure 10-2

output. In competitive equilibrium marginal costs of all firms are equal, and thus no reduction in total costs would be possible by reshuffling output among firms.

Second, the output of the industry is "correct." The price is such that marginal cost equals price. The price is, for each consumer, the measure of the importance of an increment of the commodity—a demand price of $2 is implicitly a statement by each consumer that

a marginal unit of this commodity yields $2 of utility.[3] The marginal cost is the value (= alternative cost) of the resources necessary to produce a marginal unit. If price exceeded marginal cost (as it will be shown to do under monopoly), then consumers would gain by expanding output: a product worth about $2 is obtained by sacrificing the smaller alternative product (= marginal cost).[4] The gain from expanding output would persist until price had fallen to marginal cost.

These felicitous properties of competition are the basis for using competition as an ideal. But it is a limited ideal, quite aside from a qualification for decreasing cost industries to be discussed shortly. The ideal takes the distribution of income for granted, and if this distribution is unsatisfactory to a person, he may accept as ideal only that competitive equilibrium which rules with a satisfactory distribution of income. The ideal also takes consumers' desires for granted, and if a person disapproves of consumers' choices (and of their right to make their own choices), the competitive solution is again objectionable.

In fact almost everyone will make both of these criticisms of competition on occasion. No one believes that a destitute family should starve (income distribution) or that a consumer should be allowed to feed poison to his family (consumer sovereignty). Yet in a society where there is tolerable acquiescence in the existing income distribution, and consumers are believed to have a right to much freedom of choice, these normative properties are of great importance.

THE LONG AND THE SHORT RUN

Marginal cost is defined as the increment in total cost divided by the increment in output with which it is associated. Hence we shall have as many marginal costs for a given increment of output as there are relevant ways of producing this increment. If the firm operates its plant overtime its marginal costs will be governed by the additional wages, materials, power, and so forth. If the firm expands its plant, marginal costs will also include interest on the additional investment and appropriate depreciation charges.[5] If a new plant is constructed, marginal cost may include the salary of a new superintendent, etc.

The firm will normally handle short run fluctuations in output by varying its rate of operation of the existing plant (and by holding inventories). Investments in durable assets will be made on the basis of more persistent changes in output. We call the short run the period within which the firm does not make important changes in its more durable factors ("plant"), and the long run the period within which the size (and existence) of plants is freely variable. Clearly the short run is of no interest if a firm can quickly increase and decrease all inputs, and it is basically an empirical judgment that in general there will be important resources which cannot be worn out or built in (say) a year. The long run may also be longer for contractions than for expansions, or vice-versa.

The short-run marginal cost curve of a firm will rise more rapidly than the long-run marginal cost, because the law of diminishing returns will hold more strongly, the more inputs are held constant. Both curves (for single firm changes) must rise with output in the effective region if competition is to exist—if marginal cost fell with output but selling price did not (and it does not under competition), profits would increase indefinitely with increases in output and the firm would acquire a significant control over price. But marginal cost curves for industry-wide changes, which incorporate effects of external economies, may either rise or fall with output.

The Firm and the Industry

The industry's long-run supply curve, like its short-run curve, is the sum of the marginal cost curves of the firms in the industry. Its slope will therefore be governed by two factors:

[3] Recall that at equilibrium,

$$\frac{P_a}{P_b} = \frac{MU_a}{MU_b},$$

and if we call all commodities other than A money income (B), so $P_b = 1$ (the price of a dollar is 1 dollar),

$$P_a = \frac{MU_a}{MU_{\text{income}}}.$$

[4] We say "about $2" because as output expands, the demand price falls, and with continuous demand curves even a one-unit increase in output leads to a small fall in price—perhaps from $2.00 to $1.99999.

[5] If the additional plant were to be used for only one year (even though it might last 10 years with care), the appropriate depreciation rate is 100 per cent. If the additional output is to be produced indefinitely, only a fraction (1/10 by the now unpopular straight-line depreciation formula; $10/(1 + \cdots + 10) = 10/55$ by the sum-of-digit formula) should be charged off the first year.

1. The slope of the long-run marginal cost curve of each firm (for industry-wide changes).
2. The price at which firms enter or leave the industry.

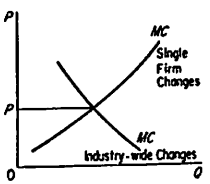
We have nothing to add on the first score: the firm will operate somewhere on its long-run marginal cost curve.[6]

The price above which firms will enter the industry, or below which they will leave, can be different for every firm (existing or potential) in the economy. It will take a higher price of aluminum pots and pans to attract a firm from cotton textiles than a firm from aluminum toys because the former firm's familiarity with the basic technology is less. It may take a lower price of trucks to attract a firm from agricultural implements than one from the hand tool industries because the large capital requirements are easier for the former firm to meet. It will take a higher price to attract a bachelor than a married couple into the corner grocery industry, because the latter has a captive labor supply.

The number and versatility of existing firms is so large, relative to the number in any one industry, that one would generally expect the number of entrants to increase rapidly as the price of the industry's product rose. Only if the industry employed specialized resources (say, a special kind of land) or if (what is ruled out under competition) there are barriers to entry would one generally expect numbers of entrants to be unresponsive to price in the long run.

The empirical evidence suggests that in fact a large part of the increases in output of a growing industry come from the existing firms.[7] Our geometry tells us that the existing firms will produce this additional output only if the long-run marginal costs of existing firms do not rise with output. This line of analysis therefore suggests that the long-run marginal costs in most industries (for single-firm changes) are relatively flat.

[6] One minor point may be noted. If the marginal cost curve for industry-wide changes falls with output, the firm will still operate where this marginal cost equals price. The individual firm never has a choice of where to operate on the curve for industry-wide changes, but the curve for single firm output changes leads to this output. The accompanying graph illustrates the point.



[7] For manufacturing industries some evidence is given in my *Capital and Rates of Return in Manufacturing Industries* (New York: National Bureau of Economic Research, 1963), pp. 31–34.

The Quicksilver Character of Competitive Industries

A large amount of effort is devoted to assisting or burdening competitive industries. The assistance may be a protective tariff, a subsidy, or some free governmental service. The burden may be a tax, a minimum wage, or a compulsory industrial safety device. Usually it is believed that the firms in the industry will reap the gain or bear the burden of the measure, at least in part. This belief is usually correct, but only temporarily.

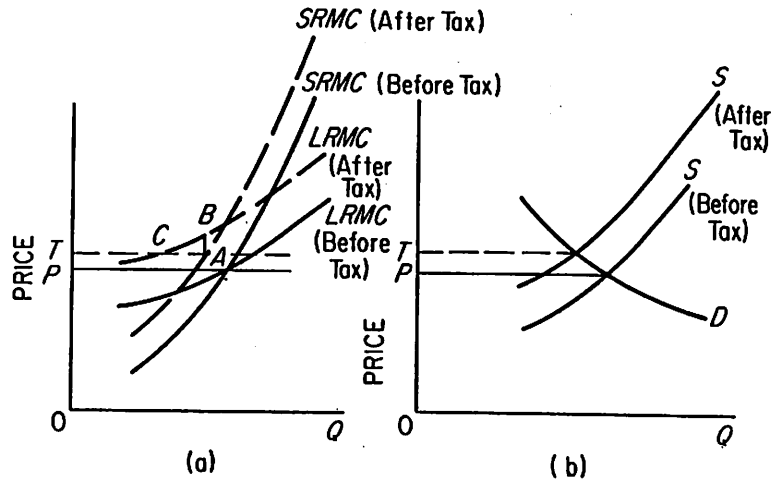


Figure 10-3

Consider a firm with the long- and short-run costs displayed in panel A of Figure 10-3, and selling its product at price $0P$. If a tax is now imposed on each firm, its costs may rise as indicated. The price will rise by a smaller amount than the tax if the demand is not completely inelastic (see panel B). Marginal losses of AB per unit of output will be incurred by the firms. With the passage of time resources will leave the industry because they can earn an amount elsewhere equal to their long-run marginal cost to this industry. Eventually the short-run marginal cost curves (and with them, their sum, the industry's short-run supply curve) will shift to the left enough to raise price to long-run marginal cost. The

contraction of output of a plant will be to some output larger than TC, because price will rise above $0T$ as the number of firms declines. The firms will again be earning a competitive rate of return. The analysis of a subsidy is completely symmetrical.

Only short-run gains or losses, therefore, can be given to the firms in a competitive industry. These gains may of course be large: if durable assets without alternative uses have on average a remaining life of 6 years, a firm may gain 3 or 4 years' return if the policy prevents the contraction of the industry,[8] or if it takes 3 years to build a new plant, extra gains may persist this long.

Even these temporary gains or losses will not be incurred, however, if the developments are fully anticipated. If the tax is anticipated, investment will have fallen appropriately by the time it is imposed. Similarly, if a tariff is expected, the industry's investment will have risen to where only a competitive rate of return is obtained when the tariff is imposed.

There is one group who may reap permanent gains or losses from policies designed to help or burden an industry: the owners of specialized resources. They will not have alternative uses for their resources, so their returns will vary directly with industry output. Thus the permanent beneficiaries of a subsidy on zinc will be the owners of zinc mines; the permanent losers from rent ceilings will be landowners.

Is the Output of Decreasing Cost Industries Optimal?

We have said that a competitive industry has an optimal output—when marginal cost equals price, resources are satisfying marginal demands in this industry as important as these same resources could satisfy elsewhere. Decreasing cost industries, however, pose a special problem.

Consider the long run cost curves for single-firm changes, LMC_1 and LAC_1 in Figure 10-4. Let us begin with a price of \$10, and an output of the firm of 1,000 units per week. Average costs are \$9.60, and the "profit" of $1{,}000 \times \$0.40 = \$400$ is the payment to the entrepreneur for his scarce services.[9] Total costs of production are $1{,}000 \times 9.60 = \$9{,}600$.

[8] The duration of the short-run gains will depend upon how much the industry would have to contract, as well as how fast it would contract, in the absence of the favoring legislation.

If now demand increases and price rises, the firm's output will rise to (say) 1,400 units. Since this is a decreasing cost industry, some inputs fall in price and the cost curves for single firm changes

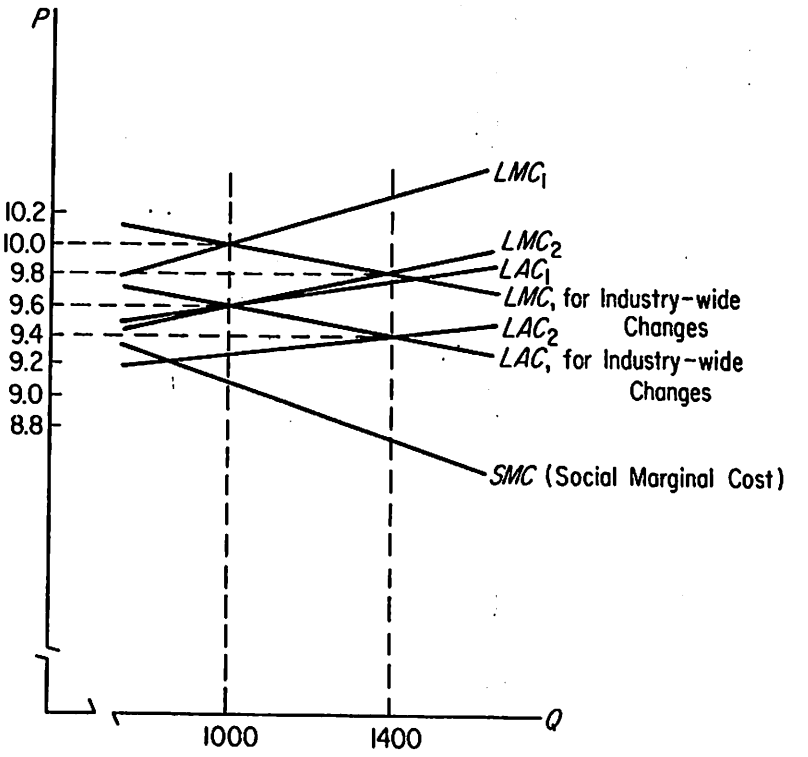


Figure 10-4

shift downward to LAC_2 and LMC_2. The price in the new equilibrium, we assume, is \$9.80 and the average costs \$9.40. (The long-run marginal cost for industry-wide changes is also presented as the locus of intersections of the various marginal costs with de-

[9] If this type of entrepreneurial service were not scarce, there would be sufficiently many firms in the industry, each operating at 1,000, that marginal and average cost would be equal.

mand.) Total costs are now $1{,}400 \times 9.40 = \$13{,}160$. Hence the marginal cost of output (for industry-wide changes) is

$$\frac{\$13{,}160 - \$9{,}600}{1400 - 1000} = \$8.90.$$

From the social viewpoint, it would be desirable for the industry to expand because price ($9.80) is in excess of marginal cost. No one firm will find this expansion feasible because when it expands output alone, it receives only $1/n$ of the reduction in input prices that results from the rise in output—the remainder goes to the other $(n - 1)$ firms.

Decreasing cost industries therefore operate at too small an output. The extent of the departure from a socially optimal output will depend upon the rate of fall of input prices; or more generally, on the extent of the external economies.

It might be, and in fact has been, argued that by a symmetrical argument increasing cost industries will be too large. It is true that when the firm buys more of an input subject to rising supply price, it will ignore the resulting rise in its price because this rise will be borne by the other firms. The arithmetic is indeed strictly parallel: let the supply of the input be

QUANTITY	PRICE	TOTAL COST	MARGINAL COST
100,000	$10	$1,000,000	
110,000	10.50	1,155,000	$\frac{\$155{,}000}{10{,}000} = \$15.50$

The firm will consider $10.50 to be the marginal cost of the input, since its purchases do not affect its price.

But the conclusion is false: increasing cost industries are not too large. The alternative product of the input must be $10.50, when 110,000 units are purchased by this industry, or the input could not be obtained at this price. The extra $5 is a rent accruing to the suppliers of the input who had previously received only $10.[10] No product is foregone as a result of this price increase—it is a transfer payment, ultimately from consumers of the product to owners of the input. The difference between the decreasing and increasing cost industries is this: the price increases of inputs do not represent foregone products, whereas the price decreases of inputs represent economies in their production on a larger scale.

[10] Their receipts rise by $100{,}000 \times \$0.50 = \$50{,}000$ and $\$50{,}000/10{,}000 = \$5$.

An Exercise in Analysis

The apparatus of competitive price theory is the staff of life for the economist: he uses it much more often than any other part of his knowledge, and it is the basis upon which most of his fancier knowledge is erected. A thorough command of the apparatus comes only from using it frequently, but we must be content here with a partial analysis of a general problem, the effects of protection of agriculture in an industrial society.

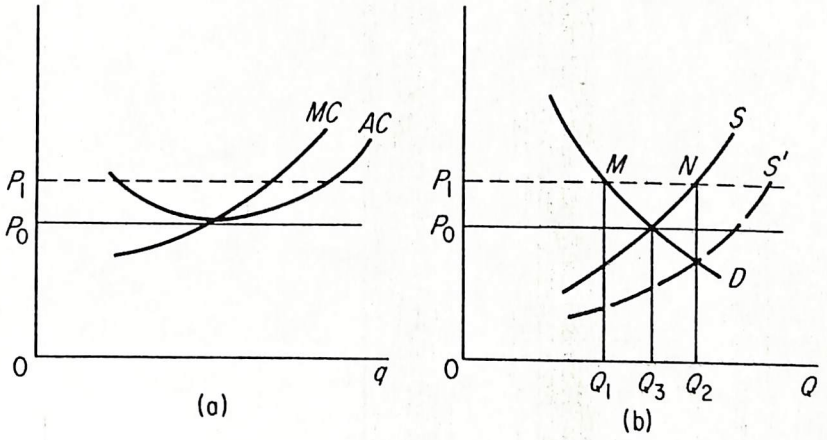


Figure 10-5

Agricultural industries, both in the United States and elsewhere, are often given assistance by price support programs. A governmental agency (the Commodity Credit Corporation is our leading instrument) will lend at designated prices against the product on what are called nonrecourse loans (loans which permit no assessment on the farmer if the agency fails to recover the full amount of the loan). The program is presumably initiated when the industry is earning less than the rate of return in other industries. Hence the initial position for a firm and the industry are something like the situation portrayed in Figure 10-5, A and B, with price p_0. The support price is set at p_1, and it obviously serves to increase output and diminish purchases, and to increase consumer expendi-

tures if demand is inelastic. In fact the increase in producers' receipts will be the sum of

Increase in consumer expenditures, $Q_1p_1 - Q_3p_0$,
Governmental loans, $(Q_2 - Q_1)p_1$.

In each period of time (say, crop year) the governmental stocks will rise by $(Q_2 - Q_1)$, assuming there is no entry of new firms or expansion of existing firms,[11] and that costs of production do not change for a farm. If technological progress lowers costs and shifts the industry supply curve to S', of course the governmental

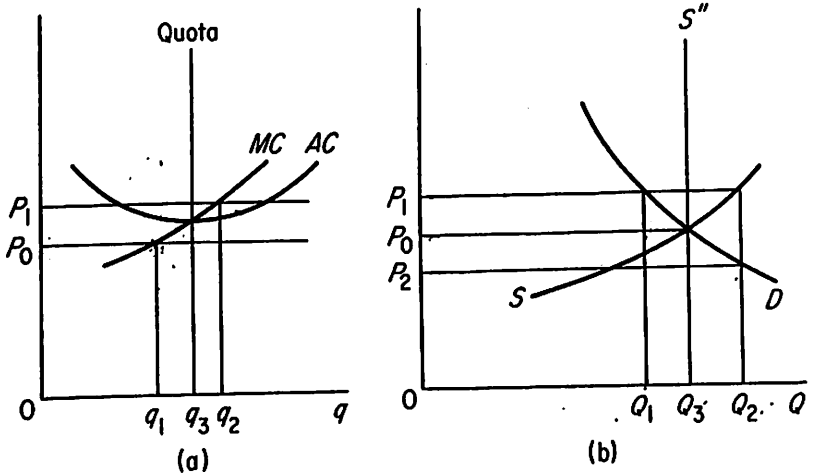


Figure 10-6

stocks increase more rapidly. Eventually there will be complaints at the growth of the governmental stocks (whether for reasons of expense or the outrage of some primitive ethical code), and production controls will be imposed. The controls may be direct output quotas for individual farms, or more commonly—because of the short-run fluctuation of yields due to weather changes—quotas on the acreage devoted to the product.[12] The situation is now illustrated by Figure 10-6, where Q_3 is the sum of the quotas of all

[11] The latter assumption is especially unrealistic, but is made to simplify the discussion.

[12] Since one input is being fixed, but others are free, the farmer will substitute other inputs (fertilizer, better seed), so output will not fall in proportion to the reduction in acreage.

farms. The annual increase in governmental stocks now decreases to $(Q_3 - Q_1)$. There is no saving to consumers, but governmental expenditures fall by $(Q_2 - Q_3)p_1$.

Let us accept without question the desirability of giving the producers the increase in income here achieved. This income increase for a typical farmer is $q_3(p_1 - p_0)$ *minus* the additional costs of growing the larger quantity $(q_3 - q_1)$, which is the area bounded by p_0, q_3, and MC in Figure 10-6A. The objections to giving this increased income in this manner are

1. Producers are using an unnecessarily large amount of resources to produce the output:

(a) Marginal costs will inevitably vary among firms—violating the optimum property discussed earlier.

(b). If an input is controlled, the substitution of other inputs will lead to the violation of another optimum condition: that inputs be used in such proportions that their marginal products are proportional to their social costs. Here too much fertilizer, and not enough land, will be used.

2. A portion of the output is unnecessary, and is measured by governmental purchases. Storage costs should be added.

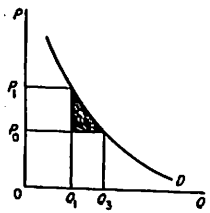
3. The price is above marginal cost (even accepting the combination of inputs used) so consumers would gain by an expansion of purchases.[13]

The first two components represent resources wasted; the third represents the consumer loss due to an inappropriate composition of output.

The same increase in income could be given to the farmers by other devices:

1. Output quotas could be made sufficiently small to raise prices and reduce costs the desired amounts. Then component 2 of waste would be eliminated; the other components of waste would rise.

[13] They would gain roughly the amount indicated by the shaded area in the accompanying figure. The increase in output $(Q_3 - Q_1)$ would require resources which would produce roughly $p_0(Q_3 - Q_1)$ worth of product elsewhere, which must be foregone. Notice that this is a measure of the utility gain to consumers; the money gain (if demand is inelastic) represents only a transfer from farmers.



Query: would the price still be p_1?[14]

2. The government guarantees each producer a price p_1 but the market could be allowed to become free, so the price would fall to p_2 (Figure 10-6B). This scheme would eliminate the first two components of waste, but retain the third (with price below marginal cost).

Query: how much should the guaranteed price be to keep farmers' incomes constant?[15]

3. Prices and output could be freed, and a direct subsidy paid. Then all components of waste would be eliminated.

Query: is the subsidy now larger or smaller than in case 2?[16]

We should notice that this third policy, and in fact all policies, raise other economic (to say nothing of political) questions. Each policy implies a different income distribution, immediately for farmers and consumers, ultimately for everyone through the implicit taxation necessary to finance the policies. The quota systems will benefit landowners who possess quotas, but not tenant farmers. The direct subsidy system (policy 3) and the quota systems must face explicitly the problem of dividing the benefits among farmers; the subsidized price system (policy 2) need not. All systems except the subsidy system will yield larger benefits to farmers as technology improves (and cost curves fall), which may be a factor in the opposition of farm groups to the direct subsidy plan.

RECOMMENDED READINGS

Knight, F. H., "Cost of Production and Price Over Long and Short Periods," *Journal of Political Economy*, 29 (1921), 304–35; reprinted in *The Ethics of Competition* (New York: Harper & Brothers, 1935).

[14] To keep the questions tolerably manageable, assume that the cost curves stay put, that is, there is no substitution of other inputs for land. Assume also that we are interested in "profits"; if some of the farmer's wage and interest income must be separated out of the cost curves, the geometry becomes complex. Then a farmer's receipts fall by $p_1(q_0 - q_1)$ at price p_1, and costs fall only by the area bounded by p_0, q_1, and MC in Figure 10-6. Hence price must rise above p_1 to maintain his profits.

[15] The rise in income from expanding output to q_2 would exceed the rise in costs (since MC is less than p_1), so the price would fall below p_1 if profits were maintained.

[16] The subsidy is smaller. The costs of the extra produce $(q_2 - q_3)$ which could be sold only at a price less than marginal cost, need not be incurred.

Marshall, A., *Principles of Economics*, London: Macmillan, 1922, Bk. V, Chs. 1–5.

Wicksteed, P. H., *The Commonsense of Political Economy*, London: George Rutledge & Sons, 1934, Vol. II, Bk. 3.

PROBLEMS

1. A general problem in pricing. (This is a summary of a problem constructed by the late Henry Simons, in *Economics 201: Materials and Problems for Class Discussion*, University of Chicago, n.d.)

An industry consisting of 1,000 firms produces a standardized product. Each firm owns and operates one plant, and no other size of plant can be built. The variable costs of each firm are identical and are given in the adjoining table; the fixed costs of each firm are $100.

OUTPUT	TOTAL VARIABLE COST	OUTPUT	TOTAL VARIABLE COST
1	$10	13	$101
2	19	14	113
3	27	15	126
4	34	16	140
5	40	17	155
6	45	18	171
7	50	19	188
8	56	20	206
9	63	21	225
10	71	22	245
11	80	23	266
12	90	24	288

The industry demand curve is $pq = \$255{,}000$. Calculate the marginal and average costs of a firm, and the demand schedule of the industry for prices from $10 to $20. (See p. 238 for the cost equation.)

PART I

(a) Draw the supply curve—that is, the sum of the marginal cost curves—and demand curve of the industry on the same graph (Figure 1). Read off the equilibrium price and quantity. Prove that the answer is correct by comparing quantities supplied and demanded at (1) a price $1 higher, (2) a price $1 lower.

(b) Draw the cost and demand curves of the individual firm on the same graph (Figure 2). Accompany these graphs with detailed textual explanation of their construction.

PART II

Congress now unexpectedly imposes a tax of \$4 per unit on the manufacture of this commodity. The tax becomes effective immediately and remains in effect indefinitely. Assume (1) no changes in the economic system other than those attributable to the tax; and (2) none of the changes due to the tax has any effect on the prices of productive services used by this industry.

(a) Draw the new supply curve and the demand curve of the industry (Figure 3). Read off the new equilibrium price.

(b) Draw the new cost curves and demand curve of the individual firm (Figure 4). Explain the details of the construction of these graphs.

(c) Why can the price not remain as low as \$15?

(d) Why can the price not rise to and remain at \$19?

(e) Precisely what would happen if the price remained for a time at \$16?

(f) At precisely what level would the price become temporarily stable? What does it mean to say this is an equilibrium level?

(g) Suppose the short-run equilibrium price to be \$17. How would you answer the query: "I don't see why every firm should produce 15 units per day when the price is \$17. It would make just as much if it produced only 14, for the 15 unit adds just as much to expenses as it adds to revenues." Precisely what would happen if some firms produced 14 units per day and others 15 units?

(h) Would short-run equilibrium be reached at a higher or lower price (and with larger or smaller output) if the elasticity of demand were lower (less than unity)? If it were higher (greater than unity)?

(i) What would happen if demand had an elasticity of zero? An elasticity of infinity?

PART III

As Figure 4 will reveal, the new minimum average cost is \$19. The short-run equilibrium price is \$17; hence this industry becomes unattractive as an investment, relative to other industries. As plants are worn out, therefore, they will not be replaced; plants will be junked sooner; and even maintenance will be reduced. To simplify the problem, we assume: (1) each plant has a life of 1000 weeks; (2) the plants in the industry are staggered so that, at the time the tax was imposed, there is one plant 1 week old, one plant 2 weeks old, and so on; and (3) at the time the tax was imposed, 20 plants were so near completion that it was impossible to divert them to other uses. These are completed at one-week intervals. Hence, for 20 weeks the price will stay at \$17, and then rise gradually as entrepreneurs fail to replace worn-out plants.

(a) What will the situation be at the end of the twenty-fifth week? (Answer in terms of "greater than" or "less than.")

(b) When 120 weeks have passed (900 plants left) will the price be above or below \$18?

(c) How many weeks must pass (how many plants must be scrapped) before the price rises to \$18?

(d) Will the output per plant increase or decrease as the number of plants declines?

(e) When 220 weeks have passed (800 plants left), will the price be above or below \$19?

(f) How many plants must be scrapped before the price rises precisely to \$19?

(g) What would the price be if the number of plants declined to 750? What would be the output per plant? What would happen to the number of plants?

(h) What happens to the short-run supply curve of the industry as the number of plants diminishes. Draw, on the same graph (Figure 5), the supply curve when there are 1000 firms and 800 firms. Compute elasticities of supply for these two curves at a given price.

(i) How could the process of adjustment, and the final equilibrium, be different (1) if the elasticity of demand were greater than unity; and (2) if the elasticity of demand were less than unity? (The significant points are price, output per plant immediately after the tax is imposed, and number of plants and total output at the new long-run equilibrium.)

2. The same problem with multiple products. Assume that the cost schedule in the foregoing table is for outputs of commodity X, and that for every unit of X, one unit of Y is necessarily produced. The demand curve for X is $pq = \$170{,}000$, and the demand curve for Y is

$$p = \$22 - \frac{q}{1000}.$$

PART I

(a) Verify that the industry is in equilibrium. The marginal costs of X and Y cannot be calculated separately (p. 162), so the supply curve of the industry refers to the equal quantities of X and Y forthcoming at any price. Hence draw the demand curves for X and Y and add them *vertically* to get the price per unit of X plus Y.

(b) Then, for the individual firm, draw the demand curves for X and Y and their sum against the costs, to find profits.

PART II

A permanent decrease in the demand for X now takes place unexpectedly. The new demand curve is pq = \$100,000.

(a) Find the new prices of X and Y and the loss per firm.

(b) What would be the effect on short-run prices of a more elastic demand for X? For Y?

PART III

Make the same assumptions about plant life and the rate of entry and exit of firms as in Problem 1.

(a) What will the prices of X and Y be when there are only 900 firms in the industry? What will losses per firm be?

(b) What is the number of firms consistent with the price of X plus the price Y equal to \$15? Is this the long-run equilibrium?

(c) If a technical change now permitted the proportions between X and Y to be variable within considerable limits, would you expect the price of X to rise relative to that of Y?

chapter eleven

The Theory of Monopoly

Let us now make an abrupt transition from the industry of many firms to that of one firm. This firm may owe its sheltered existence to a patent, the fact that it is much more efficient than any small rival, or to other circumstances which we shall discuss in the next chapter.

MONOPOLY PRICE

A monopolist is no less desirous of profits than a competitive firm, and is in a somewhat better position to achieve them. The monopolist will by definition face the industry demand curve, and take conscious account of the influence of his output on price. When he increases his output, the resulting fall in price will be borne by himself alone—not, as under competition, almost exclusively by rivals. Marginal revenue is therefore less than price, and is in fact given by the equation,

$$\text{marginal revenue} = p\left(1 + \frac{1}{\eta}\right),$$

where η is the elasticity of demand. It follows immediately that since no monopolist will willingly operate where marginal revenue is negative, he will never willingly operate where demand is inelastic.

Maximum profits are obtained when an increment of output adds as much to revenue as to cost, that is, at the output where marginal revenue equals marginal cost. We illustrate this principle in Figure 11-1, where output will be $0M$ and price MT.

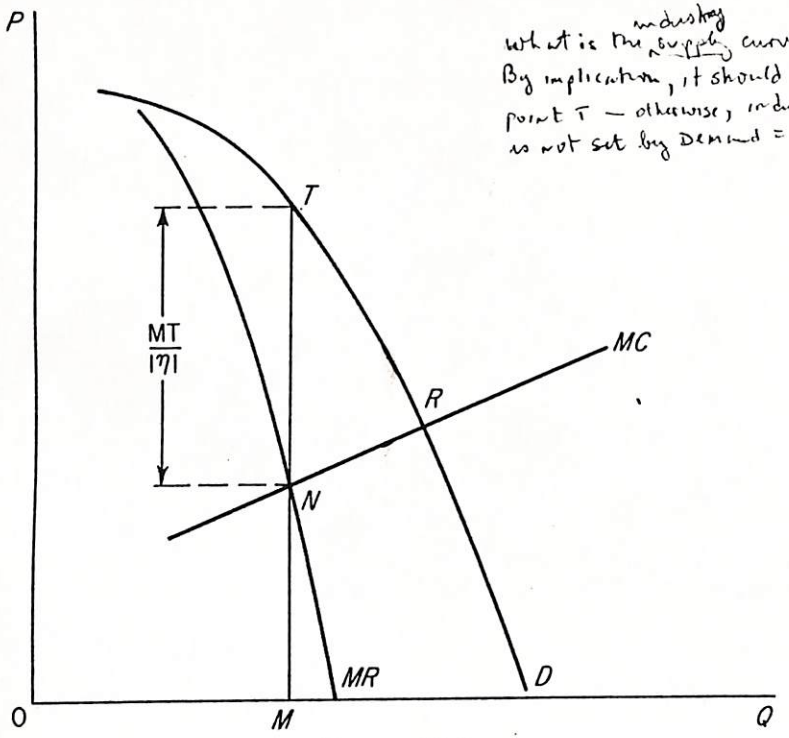


Figure 11-1

Monopoly and National Income

The Earl of Lauderdale criticized those writers who said that a nation's wealth was the sum of the wealth of its citizens:

> The common sense of mankind would revolt at a proposal for augmenting the wealth of a nation, by creating a scarcity of any commodity generally useful and necessary to man. For example, let us suppose a country possessing abundance of the necessaries and conveniencies of life, and universally accommodated with the purest streams of water—what opinion would be entertained of the understanding of a man, who, as the means of increasing the wealth of such a country, should propose to create a scarcity of water . . . ? It is certain, however, that such a projector would, by this means, succeed in increasing the mass of individual riches.[1]

[1] *An Inquiry into the Nature and Origin of Public Wealth* (Edinburgh: A. Constable, 1804), pp. 43–44.

Forming a monopoly of water and selling it, however, would lead to a reduction in national income, the noble Earl to the contrary.

The reply is superficially easy: the income of the monopolist would rise, but the (real) income of others who must now pay for water would fall. Yet this sounds like a simple transfer of command over the community's output, which would leave aggregate income unchanged. The reduction of real income would occur because wants previously satisfied no longer were satisfied, with no corresponding increase in output elsewhere (in fact a reduction, if resources are necessary to bottle and guard the water). If we constructed a price index to deflate money incomes, it would compare the cost of the bundle of goods produced before monopoly with its cost afterward, and the rise in this price index would imply a fall in real income.

The cost and demand curves need not be the same for a product if it is monopolized as they would be if a competitive industry produced it; in fact they will probably differ (more on this shortly). But if the cost and demand conditions were the same, we could measure the misallocation of resources which results with monopoly from Figure 11-1. At the margin, resources necessary to produce a unit of the product have a marginal cost, and hence an alternative product, of MN. In this industry however, they produce a product which consumers value at MT. Hence if output were expanded one unit, the product added here would exceed the product foregone elsewhere, and aggregate income would rise by NT. As additional units were produced, additional but declining gains would be achieved until marginal cost equalled price. The approximate triangle NTR measures the rise in income that would be achieved if output were to increase to the competitive level.

THE MONOPOLY DEMAND CURVE

The demand curve of a monopolist must have a negative slope. If a firm is the only producer of a commodity, and consumers display normal demand characteristics, the firm can sell more at a lower price. Only if there is at least one other producer of the identical commodity (oligopoly) will the monopolist's demand curve be horizontal over a significant range of outputs.

The slope of the demand curve will in general depend upon how good the substitutes for the monopolized good are, and how many

substitutes there are. The producer of any commodity is limited in his price-making power by the availability of other products which are close substitutes for it. Hence monopoly can arise (in the absence of collusion among producers) only if the product of the firm is substantially different from the products of all other firms—that is, if the cross-elasticity of demand for the output of this firm with respect to the price of each other firm is small. We should therefore say that the maker of any one brand of furniture is not a monopolist because, if he raises his prices, consumers will shift to other brands. Whether the maker of nylon is a monopolist depends upon the extent to which consumers will shift to silk or rayon if the price of nylon rises relative to the prices of silk and rayon. The telephone company is definitely a monopoly because telegrams, letters, bridge parties, and messengers are poor substitutes. If there are only a few producers of the good substitutes, we call the market structure oligopolistic.

This raises the question of when the substitutes are good or poor. Suppose there is only one grocery store at point A, but a road runs through A, and there are identical rivals on this road at B and C, and the cross-elasticity of demand for groceries at A with respect to prices at B or C is 0.05. Then we would say that A is a monopolist: he can raise his price 20 per cent and lose only 2 per cent of his customers to B and C (although he would lose customers also to other products).[2] Suppose now that 50 roads run through A, with two rivals like B and C on each road. Then there are 100 rivals, and with a 20 per cent rise in the price at A, sales at each of these other stores will rise 1 per cent—that is, the quantity demanded at A will vanish. Hence the power of a firm to set prices depends upon both the closeness of substitutes and the number of substitutes; many producers of poor substitutes may limit the firm as much as a few good substitutes.

Although there is no impropriety in calling a firm a monopoly if its demand curve has an elasticity of −100, there is also little purpose in doing so. The theory of monopoly will only tell us why this firm's price exceeds the competitive level by about 1 per cent $[= p(\frac{1}{100})]$, and this order of magnitude is not very interesting in

[2] If the various firms are of equal size, then η_{ap_b} may be taken as about equal to η_{bp_a}. Hence a 20 per cent rise in p_a will lead to roughly a 1 per cent rise in purchases at both B and C, and thus to a fall of only 2 per cent in purchases at A. See mathematical note 10 in Appendix B.

a world where the best measurements of marginal cost have more than a 1 per cent error. In general we shall wish to think of monopoly as involving demand curves which are not extremely elastic.

The monopolist's demand curve will depend upon the conventional determinants: the prices of substitutes and complements, incomes, and tastes. Incomes are beyond his control, but the prices of complements and substitutes are frequently capable of being influenced.

The entrance of the automobile companies into the finance business may illustrate the influencing of complementary prices. The purchase of an automobile depends upon the cost of credit as well as upon the price of the automobile, and in fact for buyers on credit the relationship is additive (down payments aside): the same increase in sales can be achieved by reducing the price of the car, or the cost of credit, by $10. If credit is supplied competitively, there is no profit in reducing its price further, but if it is supplied on monopolistic terms (by dealers), a reduction in price will benefit automobile producers. Of course it may be asked why a monopoly in financing automobile sales would not attract others besides the automobile manufacturers. The answer may be that entry is much easier for the manufacturers than for others, since they can compel the use of their credit facilities by their dealers as part of the franchise,[3] or the answer may be that the manufacturers simply were the first to be attracted by the gains. Indeed the main effects of the entry of the automobile finance companies would be (1) to redistribute profits between manufacturers and dealers, and (2) probably to lower credit costs to buyers of automobiles.[4] When the typical savings and loan association extends a mortgage, it writes the property insurance policy through an affiliated agency, which is a related instance of the exploitation of complementary demands.[5]

[3] Commercial banks did eventually enter into this line of finance.

[4] That the entry will not lead merely to a redistribution of monopoly profits from financing can be shown as follows. For a dealer the rate of return on selling cars will be at the competitive rate (assuming the automobile firm is not engaged also in philanthropy), but his rate of return on financing activities where *he* has monopoly power will be above the competitive level. Hence he will sacrifice auto sales to obtain more than a competitive rate of return from sales of finance, whereas the manufacturer will prefer an output mixture with more cars and less financing revenue.

[5] If the insurance agency business is competitive, the profits from this combination presumably come from the avoidance of selling costs.

Advertising

We could have discussed advertising earlier, for it will occur also under competition. Under competition, the main tasks of a seller are to inform potential buyers of his existence, his line of goods, and his prices. Since both sellers and buyers change over time (due to birth, death, migration), since people forget information once acquired, and since new products appear, the existence of sellers must be continually advertised. Price information poses heavy burdens: a store selling a thousand items would have to advertise perhaps 10,000 prices a year if it wished to remind people of its prices and notify them of changes.

This informational function of advertising must be emphasized because of a popular and erroneous belief that advertising consists chiefly of nonrational (emotional and repetitive) appeals. Even the seller of aluminum ingots or 2,000 horsepower engines advertises (and makes extensive use of solicitation through salesmen), although he is dealing only with more or less hard-headed businessmen.

What is true is that under competition the individual firm will not attempt to increase the desire for the product. Even if $1 of advertising would increase total sales of apples by $5, a single farmer would obtain only a tiny fraction of the industry's return, so only a cooperative advertising program would be feasible. A monopolist, on the other hand, would obtain the full returns from the advertising and hence undertake it.

Advertising, and selling activity generally, will be pursued like any other productive activities, until the expected returns and costs of various media are equated at the margin. It is commonly believed that advertising may first yield increasing, and then decreasing, returns—where we measure the marginal return of a dollar of advertising by the increase in receipts, holding output constant.

The return from a given advertisement will accrue gradually over time. Let us assume that the correct amount of advertising for a firm is $100,000 a year, and that it will reach 20 per cent of potential customers, who number 200,000. Moreover, assume that each year 5 per cent of the customers die or move away (and are replaced by births or immigrants), or forget the product once they have learned of it.

1. In the first year, $0.20 \times 200{,}000$, or 40,000 customers are informed.
2. In the second year,
   $0.95 \times 40{,}000$ old customers are still informed = 38,000
   New customers are $0.05 \times 200{,}000 = 10{,}000$
   Previously uninformed customers $= 0.95 \times 160{,}000 = 152{,}000$
   $0.20 \times 162{,}000$ uninformed customers = 32,400
   Total informed = 70,400
3. In the third year,
   $0.95 \times 70{,}400$ old customers are still informed = 66,880
   New customers are again 10,000
   Previously uninformed customers $= 0.95 \times (200{,}000 - 70{,}400)$, or 123,120
   (or, more simply, there are $200{,}000 - 66{,}880 = 133{,}120$ uninformed customers)
   $0.20 \times 133{,}120$ uninformed customers = 26,624
   Total informed = 93,504

This process can be continued, to yield the set of numbers of informed customers given in Table 11–1.[6] In eventual equilibrium, each year 10,000 new customers enter the market to replace those who leave, and 5 per cent of 166,667 informed customers (= 8,333) leave or forget the product. The number of uninformed customers is 41,667, made up of:

10,000 new customers, who replace
8,333 previously informed customers,
1,667 (= $0.05 \times 33{,}333$) previously uninformed customers,
31,667 (= $0.95 \times 33{,}333$) previously uninformed customers.

The accumulated advertising capital consists of the value of being known by 166,667 customers, and depreciates at the rate of 5 per cent a year. Since this depreciation is exactly offset by new advertising costing $100,000, the capital value of the advertising is $20 \times $100,000 or $2 million.

[6] See my "The Economics of Information," *Journal of Political Economy* (June 1961).

If customers turn over or forget quickly, of course the depreciation rate will be higher and the capital value will be smaller. But under these conditions larger amounts of advertising will be necessary to reach any given number of customers—so hotels catering to tourists will advertise more than apartment houses.

The effect of advertising on the elasticity of demand for the product is still a matter of conjecture. The primary purpose of advertising is of course to shift the demand curve to the right and upward. It is often said that the monopolist wishes a less elastic demand because he may then raise the price without a large reduction in

Table 11–1

Number of Customers Informed by a Given Rate of Advertising

YEAR	NUMBER
1	40,000
2	70,400
3	93,504
4	111,063
—	—
Eventually	166,667

sales. This presumably means that he prefers D_1 to D_2 (Figure 11-2); if so, it is false. Beyond output T (the monopolist may wish to operate beyond T to maximize profits) D_2 is a more profitable demand curve because a given quantity can be sold for a higher price. For the statement to be valid, one must restate it: the monopolist prefers a higher inelastic demand curve (D_3) to a lower elastic demand curve (D_2). This is indeed true, but it is also true if the words "elastic" and "inelastic" are interchanged or deleted.

Future Effects of Present Prices

We noted that a firm in a competitive industry must ignore the effects of the industry's price on future sales because it could not influence the price. A monopolist cannot ignore such future influences, and as a result the distinction between the long run and the short run loses much of its relevance in a regime of monopoly.

Suppose, to be concrete, that the demand curve of a monopolist this year is given by

$$q_t = 100 - p_t,$$

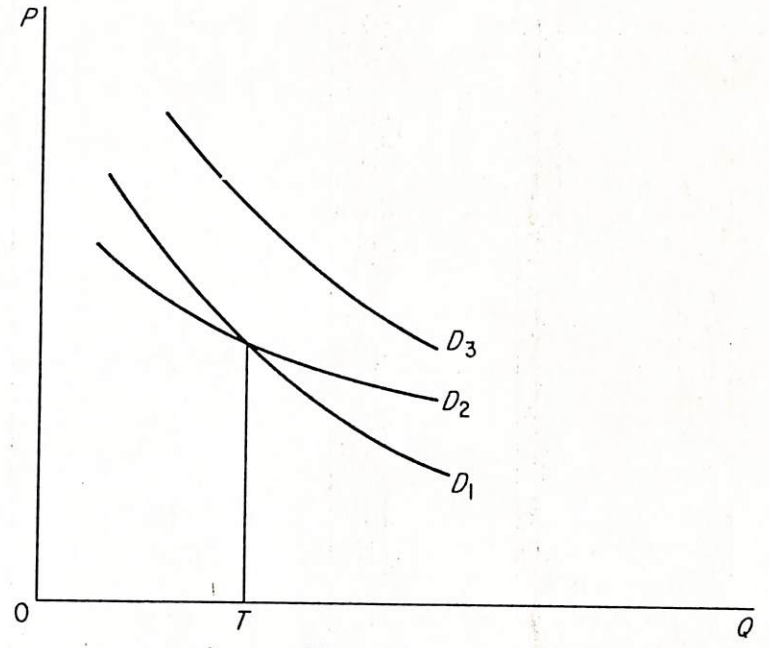


Figure 11-2

but that the demand curve next year is given by

$$q_{t+1} = 100 - p_{t+1} + \tfrac{1}{2}(30 - p_t).$$

This demand curve tells us that one more unit can be sold next year for every two dollars by which the price falls short of $30 this year. If consumers have delayed responses to prices, as we argued above (Chapter 3, p. 26), this sort of demand function is highly plausible.

The marginal revenue of output in the present year will then have two components:

1. The current marginal revenue, which may be calculated:

$$p_t = 100 - q_t,$$

$$\text{Revenue}_t = q_t(100 - q_t),$$

MR_t = Revenue from $(q_t + 1)$ units *minus* revenue from q_t units,

$$= (q_t + 1)(100 - [q_t + 1]) - q_t(100 - q_t)$$

$$= 99 - 2q_t.$$

2. The future marginal revenue from current output (we ignore discounting):

$$\text{Revenue}_{t+1} = q_{t+1}\{100 - q_{t+1} + \tfrac{1}{2}(30 - [100 - q_t])\}$$

MR_{t+1} = Revenue next period if $(q_t + 1)$ units sold now *minus* revenue next period if q_t units sold now,

$$= q_{t+1}\left(\frac{q_t + 1}{2}\right) - q_{t+1}\frac{q_t}{2},$$

ignoring terms which do not involve q_t, or

$$= \frac{q_{t+1}}{2}.$$

Hence the full marginal revenue from the sale of an additional unit this year is $99 - 2q_t + q_{t+1}/2$. Hence marginal revenue in the present period will be larger, the larger output is in the next period.

The same sort of phenomenon may arise on the cost side. Suppose, for example, a reduced output in the present period will lead to laying off men, and there is a substantial cost in rehiring. Then the full reduction in costs from a decline in current output will be less than the saving in wages by the amount of prospective rehiring costs.

These effects of the future will almost invariably be to increase the elasticity of current demand and cost curves. The rational monopolist must recognize the fact that people learn from experience, and that present acts therefore have future consequences. Yet this is often implicitly denied. Thus it is said that large buyers sometimes demand goods on unremunerative terms from small, competitive suppliers on threat of taking all their business elsewhere. As a single act this is possible, and quite possibly profitable, because it will pay the supplier to sell at a price above variable costs in the short run. But in the long run such suppliers will disappear if they do not earn a competitive rate of return. A monopolist (called a monopsonist in this buying role) who plays this game will therefore end up paying more than the competitive price, since suppliers would demand the equivalent of an insurance premium against such capricious behavior.

THE MONOPOLIST'S COST CURVES: MONOPSONY

The firm which is the only buyer of a productive service (a monopsonist) has the same power to control price in buying that a monopolist has in selling. The buyer will face a rising supply price (as a rule) and this supply price represents the average cost of

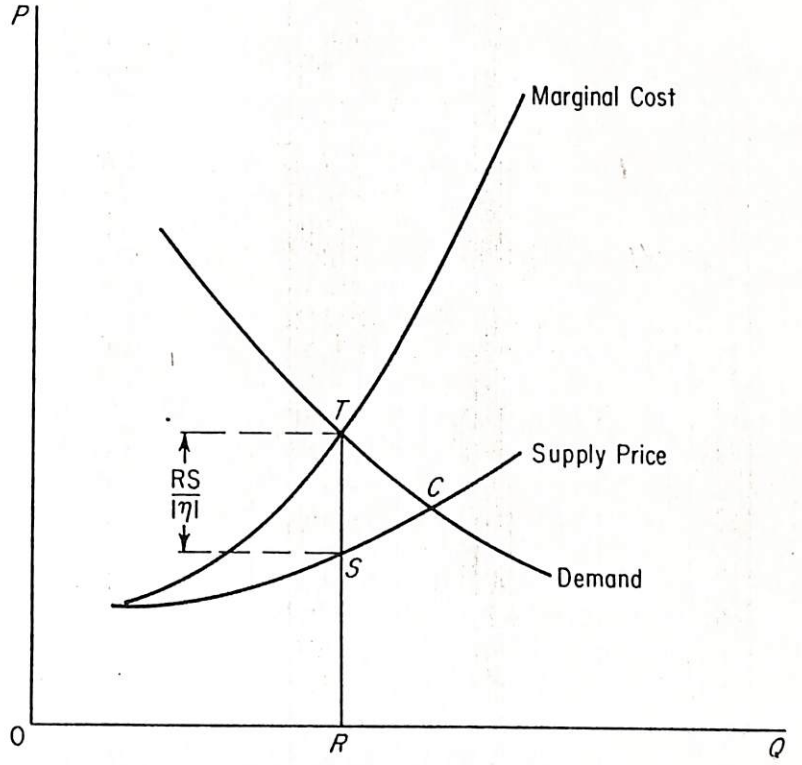


Figure 11-3

the productive service to him. The marginal cost will bear the usual relationship it has to an average, so $MC = p(1 + 1/\eta)$ where now η is the elasticity of supply. If we postulate also a demand curve by the monopsonist (it is analyzed in Chapter 14), he will buy that quantity which equates marginal cost and demand price.

We illustrate this monopsony situation in Figure 11-3. The quantity purchased will be $0R$, and the price paid to the suppliers RS.

The triangular shape, *STC*, will be a measure of the social loss arising because the resources are producing more valuable products here than in their alternative uses. The analogy to monopoly pricing is complete, and it is true in general that the formal analysis of monopoly power in buying is symmetrical with that of monopoly power in selling.

If a monopolist has any power on the buying side, he will be led to combine resources in different proportions from those which a competitive industry would use, and hence his cost curves will differ from those of a competitive industry.[7] He will in fact combine inputs A and B in such proportions that

$$\frac{\text{Marginal Product of } A}{\text{Marginal Cost of } A} = \frac{\text{Marginal Product of } B}{\text{Marginal Cost of } B}.$$

This condition for minimum cost has the same meaning that it had under competition: the marginal product divided by marginal cost is the additional product obtained by spending one more dollar on an input, and clearly if one input yields more per dollar at the margin than another, costs are not being minimized.

The monopsonist will substitute inputs whose prices rise slowly (whose supplies are elastic) for those whose prices rise more rapidly with quantity.[8] Therefore if his production function is the same as that which a competitive industry would have,[9] his average costs for given outputs would be less than those of the competitive industry. But as Figure 11-3 suggests, this "economy" is actually a waste from the economy's viewpoint.

Care must be taken, by both monopsonists and students, to know what supply curve they are dealing with. If a monopsonist buys from a competitive industry, in the short run the industry's supply curve (= sum of marginal costs) will have a positive slope because of diminishing returns. If a monopsonist should calculate a curve marginal to the firms' marginal costs, on average he will buy at

[7] Of course the comparison is with competitive cost curves for industry-wide changes—the only kind of cost curve a monopolist has.

[8] The marginal cost of a productive service to a monopsonist is $p(1 + [1/\eta_s])$, where p is the price of the service and η_s is its elasticity of supply. The monopsonist therefore uses relatively more of resources with more elastic supplies.

[9] In general it will differ because of economies or diseconomies of company size.

such prices as will impose losses on suppliers and in the long run enough firms will depart to force remunerative prices on him. Hence he has only short run monopsonistic power in this situation, and should use it only if he plans to contract his own scale. If the competitive industry's long-run supply curve rises because of rising input prices, however, he will take account of his indirect influence on input prices by calculating a marginal cost of the industry's product which is marginal to the industry's supply curve.

BILATERAL MONOPOLY

Bilateral monopoly arises when a monopolistic seller deals with a monopsonistic buyer. It would be pleasant to mention several important examples of this market structure, but its theory will serve to explain why it is seldom encountered (except in labor markets).

Suppose a monopolist has the marginal cost curve C (Figure 11-4). Then at fixed prices he would supply quantities indicated by this curve so it may be termed the average cost curve to the buyer, and then C' is the marginal cost of the commodity to the buyer. The monopsonist's marginal revenue product curve is R, and since he would purchase quantities on this curve for fixed prices, it is the average revenue curve to the seller, and R' is the marginal revenue curve to the seller. The monopolist would maximize profits by operating at output $0A$, and price AB, where his marginal cost (C) equals his marginal revenue (R'). The monopsonist would maximize profits by operating at output $0G$ and price GD, for here his marginal cost (C') equals his marginal revenue product (R). The objectives are inconsistent, so price under bilateral monopoly is said to be indeterminate.

Indeterminacy has a special meaning: the conditions of cost and demand are not sufficient to determine the price and quantity. Obviously if we look back at any year, there will have been a definite quantity and a definite price, but they will have been determined by factors outside the traditional theory: skill in negotiation; public opinion; coin flipping; a wise marriage. To say that a situation is indeterminate is a refined way of saying that it is not fully understood.

Joint profits of the two firms would be combined if they did not seek to exploit one another—that is, if they were content to exploit

is clearly discriminating. However, if it charges the same tuition for two classes whose costs per student differ by say $5, we should not call it discrimination because it would undoubtedly cost more than $5 to have separate fees for the two classes.

Conditions for Discrimination

The basic requirements for price discrimination are that there are two or more identifiable classes of buyers whose elasticities of demand for the product differ appreciably, and that they can be separated at a reasonable cost.

The demands of different buyers will be governed by the factors discussed in Chapter 3. Their elasticities may vary with

1. Income, as in the demand for medical care.
2. Availability of substitutes, as in the use of aluminum for cans facing great competition from tin plate and glass whereas aluminum in aircraft does not have good substitutes.
3. As a special case of substitutes, there may be rivals in one market (say, foreign) but not in the other (domestic).
4. Tastes, as when some buyers are eager to get early access to the commodity (a first run movie).

The form of discrimination is often more subtle than these examples might suggest. It has been common, for example, to lease rather than sell certain kinds of machinery, although the practice is declining somewhat due to antitrust convictions. When shoe machinery was leased, the basic charge was so many cents per pair of shoes processed—for example, 0.5¢ per pair for heel loading and attaching.[11] If use is not the chief cause of a machine's retirement, and it has more often been obsolescence, costs clearly are not twice as high for a machine which produces twice as many shoes, so discrimination is being practiced. The use of output as a basis for pricing is then a simple method of measuring the urgencies of desire of different manufacturers for the machine.

The tie-in sale may offer a still more indirect method of discriminating among customers. If the use of a machine is correlated with some other commodity—salt tablets for a dispensing machine, cards for a tabulating machine—the machine may be leased on a

[11] See Carl Kaysen, *United States* v. *United States Shoe Machinery Company* (Cambridge, Mass.: Harvard University Press, 1956), p. 322.

time basis and the user compelled to buy the related material from the lessor, who uses this material as a metering device to measure urgency of demand. For this explanation to hold, of course, the metering device must be sold at a non-competitive price.

Discriminatory Pricing

The monopolist will fail to maximize the receipts from the sale of a given quantity of his product unless the marginal revenue in each separable market is equal. For example, suppose he sells a given aggregate quantity in two markets at $10. If the demand elasticities are −2 and −3 the respective marginal revenues are $5 and $6.67, and the transfer of a unit from the former to the latter market will raise receipts by $1.67. In addition, the common marginal revenue must equal marginal cost.

The determination of prices may be illustrated graphically (Figure 11-5). Let the demand curves in two separable markets be D_1 and D_2, with corresponding marginal revenues MR_1 and MR_2. Then if the marginal revenue curves are added horizontally to get MR_t, we obtain the curve of aggregate quantities that can be sold at given marginal revenues. Output will be set where total marginal revenue equals marginal cost, or $0C$. This output will be sold in the two markets at prices P_1 and P_2, for at these prices marginal revenues are equal.

This analysis holds only if the markets are independent—that is, if the demand curve in one market does not depend upon the price set in the other market. This is seldom the case. Often there is some direct movement of consumers between markets: if first run movies get more expensive relative to second runs, some people will shift from the former to the latter. Often the movement is indirect. For example, if a railroad has no competition at point A but other transportation rivals at point B, we should expect demand for railroad transportation to be less elastic at the former point. Yet if the firms at A and B are in the same industry and selling in the same markets, in the long run the branch of the industry at A will decline if high rates are charged.

The theory of discrimination is only a special case of the theory of monopolies selling multiple products, and when the markets are not independent it is then necessary to treat the products sold in the various markets as fair substitutes for one another and employ

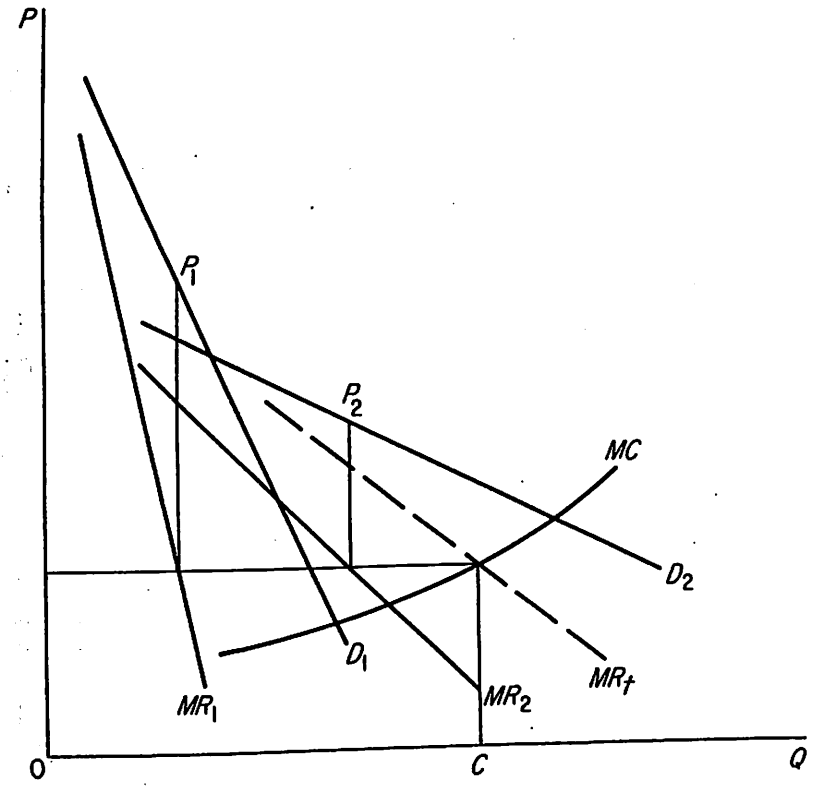


Figure 11-5

the theory of multiple products. That theory says simply that the monopolist will maximize profits if he equates the marginal revenue and marginal cost of each product. If the products are related in demand, however, one must calculate a "corrected" marginal revenue that takes account of the effect of the price of one product on the sales of others. For example, if product A has the demand schedule:

PRICE	QUANTITY	RECEIPTS
$10	100	$1,000
9	200	1,800

the crude marginal revenue is \$800/100 = \$8. But if this reduction in the price of A decreases the sales of a substitute product B,

also sold by the monopolist, from 500 to 400 units at a unit profit of \$3, then the net gain of receipts is only \$500 and the marginal revenue of A is only \$5.

Discrimination as a Condition for Existence

Although discriminatory prices are an inefficient method of allocating a commodity among individuals, they do yield a larger

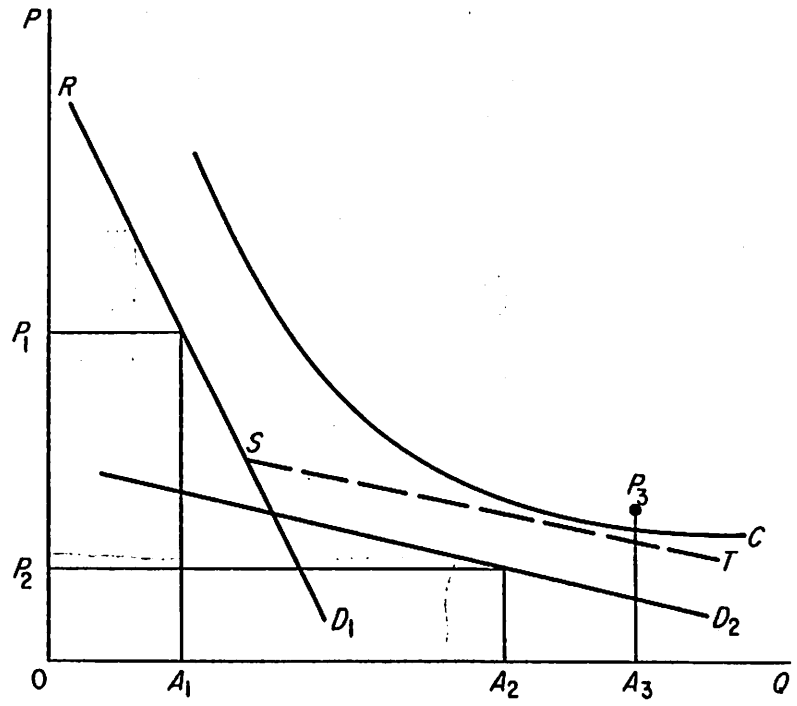


Figure 11-6

revenue than a single price system. Situations may therefore exist in which costs of production cannot be covered by receipts unless discrimination is practiced.

Consider, for example, a community with two classes of consumers, with the respective demand curves for a commodity, D_1 and D_2 (Figure 11-6). Adding these demand curves, the total demand curve is RST. The average cost of producing the commodity is C. Without discrimination, there is no output at which price is

so great as average cost. With discrimination, a quantity A_1 can be sold at price P_1, another quantity A_2 at price P_2, and the total quantity ($A_1 + A_2 = A_3$) sells for an average price of P_3, which exceeds its cost. This is, in a simplified form, the defense of price discrimination among commodities by railroads. In less extreme cases the output may be considerably larger (and also considerably smaller) with discrimination than without discrimination.[12]

Discrimination is then said to be defensible on the ground that each consumer must gain because he has his choice of buying the commodity or not, and hence he must gain if he buys it under discrimination. This is not necessarily true: the production of one commodity that is priced discriminatingly will often affect the prices of other commodities. If a railroad will haul coal for 1 cent per ton-mile and diamonds for $100 per ton-mile, the shipper of diamonds may be compelled to use the railroad because it has driven out of existence the former (competitive) stagecoach industry that hauled both commodities for 5 cents per ton-mile. Still, discrimination may be defensible on this ground.

The dilemma posed by an industry whose existence depends upon discrimination is this: if price exceeds marginal cost, there are marginal social gains from expanding output; but if total revenue falls short of total costs, the resources as a whole may satisfy more important demands elsewhere. Some economists accordingly propose a two-price system: a lump sum fee plus a price per unit equal to marginal cost. This method of pricing is in fact used when an initial installation charge plus a charge per unit is imposed. Another solution is to subsidize the loss resulting from a price equal to marginal cost from the public treasury—a solution especially appealing to the buyers of the product. Almost all genuine solutions involve much more than the reaching of optimum output: the distribution of income, the incentives to economic progress, and related economic and political questions are inevitably introduced.

RECOMMENDED READINGS

Henderson, A. M., "The Pricing of Public Utility Undertakings," *Manchester School*, 25 (1947), 223–50.

[12] There is no simple rule on the effect of discrimination on output; see J. Robinson, *The Economics of Imperfect Competition* (London: Macmillan, 1933), pp. 188–95.

Hicks, J. R., "The Theory of Monopoly," *Econometrica*, 3 (1935), 1–20. Reprinted in *Readings in Price Theory*.

Hotelling, Harold, "Stability in Competition," *Economic Journal*, 39 (1929), 41–57; reprinted in *Readings in Price Theory*.

PROBLEMS

1. If the marginal cost of a monopolist were, $MC = 60 - 3q (q < 21)$ and his demand curve were $p = 50 - q$, where would he operate? Deduce the condition for stable equilibrium.

2. Under discrimination the demand curve of a monopolist is made up of two parts:

$$p = 160 - 8q \text{ and } p = 80 - \frac{q}{2}.$$

Plot these demand curves, and the marginal cost curve, $MC = 4 + q$. Determine prices in the two markets and total profits; compare with price and profit with nondiscriminating monopoly.

3. Calculate the short-run marginal cost of a monopsonist, given the production function of Table 7–1 and the supply curve of the variable service: $p = \$6 - q/10$ (for $q < 50$).

4. A monopolist has a set of buyers, each of whom has the demand function,

$$p = 100 - q$$

and the monopolist has constant marginal costs = $10. He charges a fixed license fee which each buyer must pay in order to purchase the product, and also charges for each unit.

(a) What license fee will be set if there is no income effect upon the demand for the commodity? (Hint: the maximum fee is the consumer surplus.)

(b) What fee will be set if the quantity a consumer buys falls 1 unit (at any price) for each $10 of the fixed fee?

5. The marginal reduction in price from reading one more advertisement, or seeing one more dealer, is (on average) a diminishing function of the number examined.

(a) Will rich people pay higher or lower prices than poor people?

(b) Will people read more ads on kitchen stoves or on toasters?

(c) Will a store advertise each price? Each price change? If not, which?

Profits in Economic Theory
Michael Howard 1983
The MacMillan Press LTD, London

17
Effective Demand

Macro Effective Demand

Introduction

All forms of neoclassical theory deny the possibility of effective demand failures. In the next section we examine the reasons for this. The focus of attention is on Walrasian theory, which (as we have already seen in Part III) is the most refined product of neoclassical theorising. A denial of the possibility of effective demand failures does not imply, however, a denial of the phenomena conventionally identified as unemployment. This will be explained in the section following.

Obviously, any economics utilising a notion of effective demand must undermine neoclassical theory in some way, so we subsequently examine traditional Keynesian arguments on this issue. They prove to be rather weak. Nevertheless, there does exist stronger material from which effective demand theory can be formulated. This forms the topic of the later sections.

Walras's Law

Let us assume a market economy in which there are n commodities and re-examine the structure of Walrasian demands and supplies. The value of aggregate demand would be given by the expression

$$p_1D_1 + p_2D_2 + \ldots + p_nD_n = \sum_{i=1}^{n} p_iD_i \qquad (17.1)$$

where $p_i (i = 1, \ldots, n)$ is the price of commodity i and $D_i (i = 1, \ldots, n)$ is the sum of agents' demands for commodity i. The value of aggregate supply is defined analogously by the expression

$$p_1S_1 + p_2S_2 + \ldots + p_nS_n = \sum_{i=1}^{n} p_iS_i \qquad (17.2)$$

where $S_i (i = 1, \ldots, n)$ represents the sum of agents' supplies of commodity i.

The D_i and S_i therefore represent the market demands and supplies of agents who plan in accordance with the neoclassical assumptions. Consumers choose maximal consumptions subject to budget constraints and producers maximise profits subject to technological constraints. With these behavioural patterns it is easy to show that for any set of prices, not just an equilibrium set of prices, the magnitudes of (17.1) and (17.2) must be equal.

The value of producers' demands differs from the value of their supplies by an amount equal to profits. If consumers are *non-satiated*, so that they exhaust their budgets,[1] the value of their demands will equal the value of the assets they supply, including labour services, plus the value of profits which they receive from firms (it being assumed that consumers own firms). Consequently, the value of their demands differs from the value of their supplies by an amount exactly equal to that of producers. However, the differences are of opposite sign, so that when agents are taken all together, the value of aggregate demand is equal to the value of aggregate supply. Thus we have

$$\sum_{i=1}^{n} p_iD_i = \sum_{i=1}^{n} p_iS_i \qquad (17.3)$$

or

$$\sum_{i=1}^{n} p_iE_i = 0 \qquad (17.4)$$

where $E_i (i = 1, \ldots, n)$ is the excess demand for commodity i, defined by $D_i - S_i$.

The expression (17.3), or (17.4), is known as *Walras's law* (sometimes also called *Say's law*). As we have seen, this follows from three apparently weak assumptions: namely, that consumers maximise subject to budget constraints, that no consumer is satiated and that producers maximise profits subject to constraints of technology. Its implications are, however, not weak. It means that the structure of neoclassical theory precludes the possibility of there ever being an effective demand failure.

Unemployment in Neoclassical Theory

Both a Walrasian intertemporal equilibrium and a Walrasian temporary equilibrium involve all markets clearing. There may be an excess supply of particular commodities but they would have a price of zero (see p. 80). These commodities are most appropriately termed redundant rather than unemployed. There will be no unemployment in the sense of there being resources in excess supply at positive prices.

This is a non-controversial conclusion. However, its empirical implications are not clear cut. Neoclassical economists have never denied the possibility of unemployment as conventionally perceived. They have traced its cause to frictions and imperfections in the operation of markets[2], and today there are those, of whom Friedman is the most eminent, who maintain that appropriately specified concepts of Walrasian equilibrium can explain phenomena which are usually identified as unemployment (these economists are frequently referred to as 'monetarists' or 'new classicals' or the 'Chicago school'). In other words, their argument is that if the notion of equilibrium approximates sufficiently closely to the conditions of actual economies, then economic phenomena frequently conceived as unemployment can exist in equilibrium, and equilibrium theory can explain their determinants. In examining this argument, we shall concentrate upon the unemployment of labour but the ideas are easily generalised.

The determining structure of real market economies is conceived to be comprised primarily of tastes, technology, asset ownership and government policies. This structure

determines the phenomena observed in such economies. However, the determination is a stochastic one. Economic variables, like market prices, reflect the structure but not in a completely deterministic way. Instead, these variables show random disturbance and their values can be accurately forecast only 'on the average'. Thus it is only possible to know the probability with which a particular variable will take a specific value. It is not possible to predict with complete certainty.

Consequently, the empirically relevant concept of equilibrium is that of a rational expectations Walrasian temporary equilibrium (see pp. 112–15). As a temporary equilibrium, all currently operating markets clear, and as a rational expectations equilibrium, agents' price expectations are, 'on the average', correct. Thus in such an equilibrium the expected frequency distribution of future market-clearing prices held by agents is the distribution which will be actually encountered if the structure remains unchanged. Such an economy can experience fluctuations in real and monetary variables but all agents are adjusted to this. Consequently, the path of an economy in a sequence of rational expectations Walrasian equilibria would be approximated by a Walrasian intertemporal equilibrium.

A rational expectations Walrasian equilibrium is the appropriate conception of equilibrium because it is a terminal state. It represents a situation in which all agents are accommodated to the structure of the economy, in the sense that markets clear and the probabilities assigned to events by different agents are consistent and correct so there is no element of *systematic* error in expectations.

In this state there are no positively priced commodities in excess supply. However, there may be phenomena which convention or policy identifies as unemployment. For example, stochastic variability in commodity markets may be reflected in labour markets. Old workers will be retiring and new workers entering the labour force. Some existing workers will be relocating occupationally. Market imperfections, such as unions and minimum wage laws, can result in unduly low competitively determined wage rates. All of these processes may involve periods of temporary idleness, search for alternative employment and permanent abstention from work, all of

which, on the basis of conventional definitions, are classified as unemployment.

This unemployment is often called 'natural', using the term in the Wicksellian sense of referring to equilibrium. Thus, for example, Friedman writes:

> The 'natural rate of unemployment' . . . is the level which would be ground out by the Walrasian system of general equilibrium equations, provided there is embedded in them the actual structural characteristics of the labour and commodity markets, including market imperfections, stochastic variability of demands and supplies, the costs of gathering information about job vacancies and labour availabilities, the costs of mobility, and so on.[3]

The natural rate can change if the structure of the rational expectations Walrasian equilibrium changes.[4] Moreover, deviations from the natural rate can occur if agents do not adjust to the new structure instantaneously. Friedman is fond of locating the major cause of such changes in monetary shocks.[5] For example, imagine for simplicity that the initial equilibrium is a stationary state with a constant price level. If a monetary contraction takes place, according to Friedman's monetary theory, money wages and wage expectations will be required to take lower values in the new equilibrium.[6] If this is not immediately recognised by workers, they will interpret the money wage reductions they encounter as a reduction in real wages. This will occasion substitution into search activities, leisure, etc.,[7] and will be reflected in a rise in recorded unemployment.

'Unemployment' is higher because workers have mistaken a fall in absolute prices for a change in relative prices. They do so because they lack system-wide information on the basis of which new 'rationally expected' prices can be immediately determined.[8] They only have detailed knowledge about the sectors of the economy in which they operate and it is this tunnel vision which allows workers in general to confuse changes in money prices for changes in relative prices. It will take time for the market opportunities subjectively percieved by workers to coincide with the objective market situation,

as workers learn new forecasting rules which give results consistent with the new structure. Until they do so, expectations will be incorrect 'on the average' and they will misallocate their resources.

Unemployment above the natural rate is a disequilibrium phenomenon in the sense that it reflects that the economy is away from a terminal state, i.e. is not in rational expectations Walrasian equilibrium. However, the unemployment is an equilibrium phenomenon in the sense that agents are optimising, on the basis of the information they have,[9] and markets are always cleared. There is no 'effective demand failure' and Friedman's view is that expectations will automatically correct themselves quickly to reflect the new structure. This view is widely held by those who adhere to this form of neoclassical unemployment theory.[10]

Effective Demand Failures: The Traditional Arguments

Many theorists who have played a major role in the formalising of modern Walrasian theory have not been impressed with the application of this theory to explain unemployment. Their own view as to the status of Walrasian theory is to emphasise its counterfactural usefulness (see pp. 85–6),[11] and their suspicion of the theory outlined in the preceeding section derives in part from the fact that it treats the results of formal Walrasian theory in a most cavalier fashion. These results show that the existence of unique and stable equilibria can only be guaranteed on very stringent assumptions which are unlikely to be fulfilled in actual economies.[12] Disbelief that Walrasian theory is the appropriate path along which an understanding of unemployment should progress is also buttressed by historical experience. Some economists have viewed with incredulity the work of those who would seek to explain the heavy and persistent unemployment in the 1930s with models in which markets continually clear.[13]

It is true, nevertheless, that traditional Keynesian arguments used to account for effective demand failure are theoretically weak. Viewed in the light of Walrasian theory, they simply will not bear the weight placed upon them.

One such argument is associated with Robinson. It maintains that Keynes saw 'clearly that to recognise that the future is unknown brings down the whole structure of orthodox theory'.[14] It is true that Keynes (1937) perceived Knightian uncertainty to be pervasive in market economies. However, it is not true that this in itself is a decisive objection to Walrasian theory, as we have seen in Chapter 13.

More commonly expressed arguments focus upon the role of liquidity preference in maintaining interest rates 'too high' and on interest-inelastic investment. It is argued that, due to speculative expectations, money can become the preferred asset at rates of interest above the level required for full employment. Furthermore, even if it were possible for money rates of interest to fall to zero, investments may not be sufficiently interest-sensitive so as to ensure complete utilisation of resources. These arguments have become the standard fare of intermediate macroeconomic texts. In terms of the typical *ISLM* model they can be represented by Figures 17.1 and 17.2 respectively, where Y_F represents full-employment output. By themselves, these arguments have no force against Walrasian theory, and therefore no substance in accounting for effective demand failures. Both arguments relate to the functional form of particular aggregate demand relationships and this is not an issue which threatens the existence of Walrasian equilibria. The continuity of demand and supply relationships is

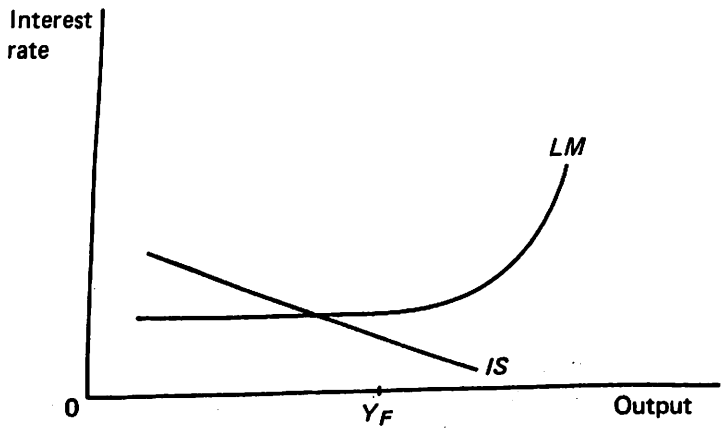


Figure 17.1

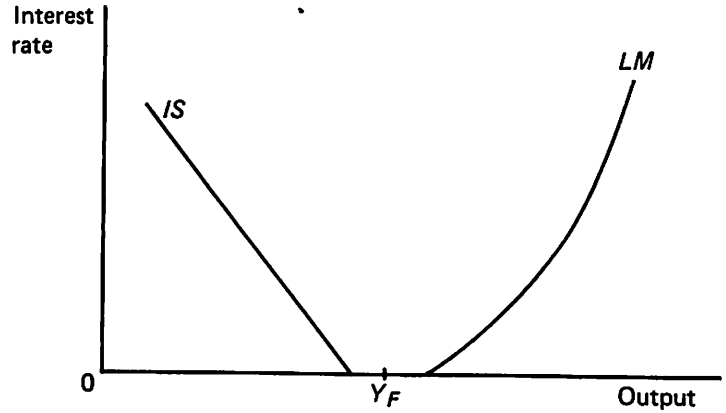


Figure 17.2

another matter (see pp. 80–1) but discontinuities play no role in the typical presentation of these arguments.

Undoubtedly, the most popular argument employed to account for effective demand failures concerns downwardly rigid money wage rates which are too high to ensure full employment. Nevertheless, by itself this argument is powerless to do so. Even if it is accepted that such rigidities characterise all labour markets, this is not sufficient to provide a rationale for effective demand failures. The reason lies in Walras's law, outlined above. One implication of this is that an excess supply of any commodity with a positive price will be balanced by an equivalent value of excess demand in other markets. (If in equation (17.4) it is assumed that some $E_i < 0$ with positive prices, then there must be other $E_i > 0$ since $\Sigma_i p_i E_i = 0$.) Consequently, there is no effective demand failure.

The problem with all these arguments is that they seek to question Walrasian results without questioning the Walrasian conceptualisation of demands and supplies upon which these results rest.

Quantity Constraints

In Walrasian theories of competitive economies, agents are assumed to formulate their demands and supplies in the belief

that they can trade in whatever quantities they deem desirable as long as they provide equivalents in exchange. Consequently, consumers are assumed to maximise utility in terms of available goods and are constrained only by a budget which is dependent upon their assets and prices, while firms are assumed to maximise profits subject only to a technological constraint. Taken together, the resulting demands and supplies yield Walras's law and the conclusion that effective demand failures are impossible.

The limitation of this Walrasian conceptualisation can be explained by considering a set of prices in which Walrasian demands and supplies are inconsistent. Obviously in this case not all demands and supplies can be realised simultaneously, and some of those agents on the long side of markets will be rationed if trades actually occur at these prices. (The *tâtonnement* mechanism, as we have seen on page 123, assumes trades will not take place. However, this is an obviously unreasonable characterisation of actual market behaviour.) In these circumstances it is not unreasonable to hypothesise that agents will develop expectations as to the probability of rationing in the future. If these probabilities are non-zero, this will affect other demands and supplies. For example, a consumer who is quantity-rationed in the sale of labour and expects this to continue at future dates will not necessarily change his or her willingness to supply labour from that specified by Walrasian theory, but is likely to reduce demand for currently available consumption goods. A firm which is quantity-rationed in the sale of its output and expects this to continue at future dates will not necessarily change its willingness to supply output from that indicated by Walrasian theory, but is likely to reduce demand for currently available labour. From this, there follows a number of important implications.

First, there are now two types of demand and supply. There are Walrasian demands and supplies which, following Clower (1965), are frequently called 'notional' demands and supplies. There are also quantity-constrained demands and supplies which depend, like notionals, upon prices, budgets and technology but also, unlike notionals, upon quantity constraints operative now or expected to be operative in the future. This means that a concept of effective demands and supplies becomes meaningful. Take the two examples of the previous paragraph. In the case of the consumer, his or her effective supply of labour would be defined as the notional supply, and his or her effective demand for any consumer good would be the quantity-constrained demand. In the case of the firm, its effective supply of output would be its notional supply and its effective demand for labour would be its quantity-constrained demand.

Second, effective demand failures are now possible because Walras's law does not extend to effective demands and supplies. It is possible for

$$\sum_{i}^{n} p_i E_i^e < 0$$

where E_i^e are excess effective demands. It is thus reasonable to imagine economies in states where some markets show excess supplies at positive prices and these are not balanced by excess demands on other markets. This means that there can be genuine unemployment. For instance, continuing with the above example, it is possible to envisage the following situation. Consumers' demand for goods is constrained by their inability to sell all the labour they supply, while firms do not employ more labour because the demand for goods is less than the amounts the firms are willing to supply. There is therefore an excess supply on both goods markets and labour markets simultaneously.

Third, economic agents' actions become dependent upon quantity variables in ways suggested by traditional Keynesian models. The level of aggregate consumption expenditure, for example, becomes dependent upon an income magnitude, which is determined by both quantities and prices. Moreover, a change in a quantity variable may alter others. The relaxation of a rationing constraint on labour sales will increase consumption expenditures, leading to a relaxation of quantity constraints on firms' sales, leading in turn to an increased demand for labour. In short, multiplier processes, which have no foundation in Walrasian theory, become possible. This also implies that the co-ordination of economic activities becomes a more complicated question to analyse because

these activities can be interelated in more complex ways than is specified in Walrasian theory. Certainly the stability results of the latter are of little relevance since they pertain only to notional demands and supplies.

Fourth, from this perspective Keynesian economics appears more general than does Walrasian economics. The latter is seen as a special case of the former because it examines the particular case in which effective demands and supplies coincide with notional demands and supplies. This is certainly in line with Keynes's own view of the status of neoclassical economics.

Effective Demand Failures and Equilibrium

The ideas outlined in the preceding section can be traced back to the work of Clower (1965) and Leijonhuvud (1968), who derived them from Keynes (1936). They have been extended into new concepts of equilibrium by many economists, some of whom previously worked within the confines of Walrasian theory. These new concepts of equilibrium can be placed into two broad categories.

First, there has been the formulation of temporary equilibrium models, involving fixed prices and in which agents' maximisations take account of perceived quantity constraints in current and future periods. Equilibrium is defined as a situation where agents' maximisations generate constrained trades, i.e. effective demands and supplies, which are consistent. Various types of equilibria are possible and some of them involve genuinely unemployed resources.[15] The weakness of these models lies in treating prices as being exogenously fixed. The rationale for doing so is the belief that in modern capitalist economies quantity adjustments initially dominate price adjustments as the response to any change in effective demands and supplies. Consequently, such models do not imply that prices never change. The economy is pictured as moving through a sequence of quantity-constrained temporary equilibria, in each of which prices are given but between which they may change. If and how they change depend upon the

determinants of prices. However, since the analysis concentrates upon a single period, prices are exogenously specified.

The second approach tries to overcome this weakness of fixed-price models. The essential idea is that agents who are rationed would willingly change prices if such changes were thought to yield a beneficial relaxation in the quantity constraints to which they are subject. Nevertheless, this willingness does not necessarily translate into price changes. Whether or not agents do change prices depends upon the 'conjectures' they hold as to how price changes will affect quantity constraints. For example, if an unemployed worker conjectures that a large reduction in his asking wage will only have a negligible impact upon the probability of gaining employment, this asking wage is unlikely to be reduced. The focus of attention is therefore upon what circumstances generate pessimistic conjectures, and the models, as so far developed, place emphasis upon incomplete information, imperfect competition and social conventions.[16] These can be such as to produce equilibria in which prices and quantity constraints are correctly forecast, so that economic processes terminate in states involving effective demand failures similar to those represented by fixed-price models but which continue over successive periods.

Conclusion

Concepts of effective demand and effective demand failures have been clearly established as theoretically viable. However, this does not imply that economic theorists will abandon neoclassical theory. The approach economists favour depends in part upon the pre-analytic vision they have of the phenomena which theory seeks to explain in a disciplined and orderly manner. We have touched on this elsewhere (see pp. 153–5) and it is also of relevance to theories of effective demand. An argument against the relevance of Keynesian models and favourable (at least in spirit) to neoclassical theory is in fact easily constructed from the way many neoclassical economists seem to perceive the nature of the market system. It could run as follows.

Market systems are essentially systems of voluntary trades. This means all parties to a set of trades must realise the maximal benefits possible, otherwise that set of trades will not be maintained. If any agent decides to form a new set of trades, he or she will need to signal other agents of this. The price system is one means of communication but it is not the only one. Market systems have evolved, and are still evolving, many systems of communication. In a situation Keynesians call effective demand failures, mutually beneficial trades obviously exist and it is equally obvious that it is in the interests of agents to locate them. Thus there can be a strong presumption that the Keynesian diagnosis is faulty and there can be an equally strong presumption in favour of an economics which formally incorporates, however inadequately, the adaptability and flexibility of market systems.[17]

Notes to Chapter 17

1. Non-satiation means that, no matter how large the consumption of any consumer, each consumer would prefer a larger consumption. This does not preclude any consumer from being satiated in the consumption of a particular commodity. It only implies that there is no consumer who is completely satisfied in the consumption of all commodities simultaneously.
2. See, for example, Dobb (1937), Schumpeter (1954), and Feinberg (1978).
3. Friedman (1969, p. 102).
4. Friedman (1969, p. 103).
5. See, for example, Friedman (1969, pp. 103–5; 1976, ch. 12).
6. See, for example, Friedman (1969).
7. See, for example, Lucas (1981, p. 48).
8. See, for example, Friedman (1976, ch. 12), Phelps (1970), and Lucas (1981).
9. Lucas (1981, pp. 4, 156, 242, 245).
10. See Phelps (1970) and Lucas (1981).
11. See also Hahn (1973; 1981), Arrow (1967; 1974), and Arrow and Hahn (1971).
12. See Hahn (1965; 1971; 1980a) and Tobin (1980, ch. 2).
13. See, for example, Rees (1970).
14. Robinson and Eatwell (1973, p. 48). See also Shackle (1967; 1972; 1974), Davidson (1972), and Coddington (1976).

15. Benassy (1975), Drezè (1975), Grandmont (1977), Malinvaud (1977; 1980), and Muellbauer and Portes (1978).
16. Negishi (1976; 1979), Hahn (1977; 1978; 1980a; 1980b), Akerloff (1979), and Buiter (1980).
17. These sentiments seem to be particularly pronounced in the work of Friedman. See Friedman (1962) and (1980).