FIRST COMMENTARY

The firm of theory: its definition and existence

The central theory of economics remains the neoclassical theory. It is essentially a theory that imparts much understanding about the price system as an allocator of resources, but it also contains a simple model of the business firm. Why is there a firm in the theory of price? What defines the firm in this theory? This, the first commentary of the present volume, is the proper place to settle these two epistemological issues. How I do settle them is at considerable variance from the manner that has become popular in the newly emerging theory of the firm. The answers to these questions lead quite naturally to the second commentary, which deals with theories of the internal organization of firms. Modern treatments of the firm do not cleanly separate the organization of the firm from the question of its existence. In the writing of this volume, I have become convinced that mixing these two issues is a source of confusion. Just as a theory of the existence of the atom need have no close relationship to a theory of the inner workings of its electron system, so an explanation for the existence of the firm may have little to do with an explanation of the firm’s inner organization.

The most important example of comingling these two issues is found in R. H. Coase’s 1937 classic article, “The Nature of the Firm.” Coase sets his task explicitly at the beginning of part II of his article. “Our task is to attempt to discover why a firm emerges at all in a specialized exchange economy.” His reply, which has become extremely influential, is that the market, or price guidance, is not free. Transaction cost, if high enough, justifies the substitution of managed coordination for price-guided coordination. But then Coase goes on to apply transaction cost analysis to the inner organization of the firm, inquiring about the degree to which the firm is vertically integrated and the reliance it places on long-term employment contracts.

The interest that economists have shown in the existence of firms, by and large, has not expressed itself in an examination of facts pertaining to firms. Instead, there has been a search for a logical theory by which to justify the existence of firms in a price system. This purely theoretical slant is a result of the fact, alluded to by Coase, that the firm as a theoretical construct fits uncomfortably into the important body of price theory we call neoclassical economics. I now believe that the fit is not a poor one if the role of the firm in the theory is correctly perceived. The uncomfortable fit results from the
2 The economics of the business firm

improper way in which we have viewed the firm in this theory. I shall come back to this important issue later in this commentary; there I will give an alternative perception of the firm.

Contemporary literature perceives the identifying element in the firm to be its reliance on managed coordination. So does the older literature, and I begin by contrasting the two most influential early views about managed coordination and the existence of firms, those of F. H. Knight and R. H. Coase. The existence of the firm seemed not to be recognized as a problem in need of addressing by the literature surrounding neoclassical theory. That literature concerned itself with a related existence problem, the existence of profit in a supposedly highly competitive economy. In such an economy, the value of the marginal productivity of a factor determines the price it receives in the market. The competitive market tolerates no factor payment in the absence of a productive contribution. Profit is a payment to the owners of the firm, but the productive contribution made by owners, unlike the sweat of laborers, is unclear, and its rationalization is complicated by the fact that profit is in the nature of a residual, not a specified or easily calculated sum. The older literature found a productive function for ownership in the risk borne by a firm’s owners, and the existence of profit was attributed to the necessity for rewarding risk taking if there was to be risk taking. This rationalization of profit, which we may associate with the factual divergence of the real economy from the full information assumption of the perfect competition model, is a major target of the critical attack made by F. H. Knight in his influential book Risk, Uncertainty, and Profit (1921). It is as a consequence of this attack that the need to explain the existence of the firm is first recognized.

Knight’s explanation

The primary task assumed by Knight in his book is not to explain the existence of the firm or even its organization, but to explain the existence of profit. Knight shows that, even if the full information assumption of perfect competition is dropped so that risk can exist, risk offers no good explanation of profit because it is associated with known probability distributions of event outcomes. An insurance company can calculate an actuarially fair premium for the risk of fire and, upon charging this premium to business firms for insurance against the event of fire, it converts this risk to just another cost of doing business. Entry, exit, and competition generally must then treat risk as any other cost of production. As such, risk cannot give rise to profit. Business prospects that exceed all costs, including this risk cost, cannot exist in competitive equilibrium. Sufficient entry is attracted to eliminate the possibility of profit.

Knight then goes on to argue that imperfect information must reflect something more than risk before it can give rise to profit (or loss). Events must be
The firm in theory: its definition and existence

unpredictable, so that no conversion of uncertainty into a known cost of production is possible. Knight creates a new classification of imperfect knowledge to fit this need, and he labels it uncertainty. This is distinguished from risk by associating it with possible outcomes about which so little is known that no calculation of an actuarial type is possible. Uncertain events can give rise to profit or loss because entry into or exit from markets cannot rationally construct an anticipation of these events.

In acknowledging the presence of uncertainty, Knight fails to face the problem of how to redefine competitive equilibrium or of how to demonstrate the existence of equilibrium when competition and uncertainty are commingled. Uncertainty is a strange source of profit if Knight’s objective is to resuscitate the role of profit in guiding resources in a price-directed economy. Being unpredictable, it cannot rationally influence resource allocation decisions. Yet what makes profit an important variable in neoclassical theory, and in Smith’s earlier work on the *Wealth of Nations*, is the guidance it gives to resource flows. Moreover – and now at last we come to the point regarding the existence of the firm – uncertainty as the sole source of profit undermines an explanation of the existence of firms if firms are thought to exist in order to seek profit. Knight, rejecting risk as a source of profit and substituting an unpredictable source of profit for it, needs a new theory of why firms exist. This, at least, is my interpretation of why Knight strays from the central topic of profit to deal with the question of the firm’s existence. He does not himself acknowledge this, but he does deviate from the chain of reasoning that moves his analysis from risk to uncertainty to profit in order to offer a rationale for the firm’s existence. The rationale, significantly, depends not on profit but on risk, or, more accurately, on risk redistribution.¹

Knight sees in the firm advantages of redistributing risk between owner-manager and employees. The profit and loss consequences of fluctuations in business outcomes are absorbed by the owner-manager, who contracts to pay relatively stable wages to employees. Employees are thus insulated at least partly from fluctuations in the business outcomes. This reallocation of risk is efficient in Knight’s view, for he sees it as putting a greater proportion of risk on the party that he believes is less averse to bearing it, the owner-manager. Presumably, risk is not handled as well without firms.

In return for offering stable wages to employees, the owner-manager requires them to allow their activities within the firm to be supervised. To accept this supervision, employees in turn require a contractually fixed wage that attenuates risk. Hence, Knight explains the existence of the firm, the wage system,

¹. Knight never really reconciles the uncertainty source of profit and the sensibility of having resources guided by profit considerations. On this and other aspects of Knight’s uncertainty theory, see Demsetz (1988b).
4 The economics of the business firm

and recourse to managed direction of employees, all as concomitant to the
desire to allocate risk efficiently. This improvement in the distribution of risk
among the cooperating parties may be identified here as Knight’s productivity
explanation of the firm’s existence. An economic explanation for the existence
of firms must insist on a convincing description of conditions, hopefully fre-
quently encountered, that make the firm a productive arrangement, and I shall
couch my discussions of other theories of the firm’s existence in terms of their
productivity explanations.2

Knight’s search for an explanation of the existence of firms is, in some ways,
echoed by the different explanation put forward later by R. H. Coase. Just as
Knight’s explanation of superior risk allocation can be interpreted as an ex-
planation based on cost reduction, so can Coase’s. However, where Knight
relies on reductions in risk cost through risk reallocation, Coase relies on
reductions in coordination cost through the substitution of managed coordina-
tion for price-directed coordination.

Coase’s explanation

In putting forward his now famous explanation of the existence of the firm,
Coase looks to transaction cost (although risk may be one factor among many
that influences transaction cost). Managed allocation of resources within the
firm becomes viable if the cost of coordinating resources through market
arrangements (i.e., transaction cost) exceeds the cost of managing them within
the firm. The reduction in coordination cost achieved through managed co-
ordination is Coase’s productivity explanation of the existence of the firm.

To argue against Knight’s view, Coase observes that risk is often accepted
without the counterbalancing shift in control that Knight claims to be neces-
sary. He points to the fact that when firms buy goods from other firms, they
usually pay a relatively riskless contracted amount to suppliers, much as the
Knightian firm pays contracted wages to employees. Yet the firms doing the
purchasing do not ordinarily insist on, or acquire, supervisory control over
suppliers.3

Coase’s counterexample is not so telling a criticism as it first appears. Knight
can be interpreted to say that a difference in risk-bearing ability between

2. The reader must remember that our topic is the existence of the firm, not its organization. As
to existence and the use of the management/wage system, Knight relies mainly on risk re-
distribution. As to other aspects of the firm’s organization, Knight offers insightful considera-
tions of moral hazard.
3. Coase might have added, but did not, that stabilizing employee wages would bring about an
increase in the variability of employment. The wage system, if this increase in variability were
great enough, would increase the risk of income fluctuation borne by wage earners rather than
by owner-managers. Knight’s theory requires that this not be so.
The firm in theory: its definition and existence

cooperating parties is required to justify the formation of a firm. This difference is taken by Knight as obvious when it comes to owner-managers and employees, but, he might argue, not when it comes to two firms. No improvement in risk allocation is achieved by adopting contractual arrangements between firms like those that exist between a firm and its employees if the owners of these firms do not have systematically different risk preferences. Hence, Knight might argue, two firms cannot on grounds of superior risk allocation justify a merger into a single firm.

There is, however, a more serious problem with Knight’s theory. The owner-manager can be compensated through market transactions for bearing a greater proportion of the risk inherent in the firm’s operations. The smaller the share of risk borne by employees, the lower is the wage that satisfactorily compensates them for their work. There is no clear need for the owner-manager to receive his reward for risk-bearing through the supervision of employees. Similarly, the price that governs exchange between firms and their suppliers, in Coase’s counterexample, should reflect how risk is shared, and Coase, in discussing this counterexample, makes passing reference to the market’s ability to set prices that compensate for risk bearing. Since wage or price adapts to accommodate to the actual distribution of risk, risk redistribution cannot, in and of itself, provide an acceptable rationale for the firm or for the supervision of employees. The price system, unaided by institutions such as the firm, can cope with the risk distribution problem. It is somewhat surprising that Knight should have overlooked this, for his discussion of the problem of basing profit on risk clearly recognizes the existence of a market-determined price for risk bearing.

But Knight might have a reply even to this point. Although he writes about risk redistribution as already described, he might claim that he has been misinterpreted. The fluctuations in outcomes that the firm experiences are really due to uncertainty, not to risk. The predictability that is required to gauge the wage concessions appropriate to compensate owner-entrepreneurs for their agreement to pay relatively fixed wages is lacking if the fluctuations in the outcome of business activity reflect uncertainty. This cannot be, however. The value of these concessions is measured by the value of the control over employee actions that has shifted to the owner-entrepreneur in return for the payment of fixed wages.

Knight’s rationalization of the existence and organization of the firm makes more sense if one simply views the management of employees as a method for reducing risk, not merely redistributing it. Reduction of risk is obviously in the interest of owner-managers, given that they have taken on a large share of the risk inherent in business operations. Reduction of risk may be what Knight really had in mind, but, if so, he did not make this claim with any clarity, and a literal interpretation of his theory justifies the existence of firms in terms of risk redistribution.
6 The economics of the business firm

The profession is indebted to Coase for pressing it to pay attention to transaction cost and for showing how this cost affects institutional arrangements such as the firm. Coase’s theory of the firm’s existence does not suffer from some of the same weaknesses as Knight’s. But coordination cost, like risk cost, offers only a possible avenue by which the existence of a firm might be justified for reasons of productivity enhancement. Potentially there are many paths to cost reductions, and each may have a bearing on whether the firm or the market is the chosen instrument of economic organization. Coase’s reliance on reduction in coordination cost may be sufficient to explain the existence of firms within an otherwise price-directed economy, but it may not be necessary. More important is the question of empirical significance. The empirical significance of Coase’s explanation and of other explanations is a more unsettled issue than is supposed by the emerging modern theory of the firm.

One important alternative source of the firm’s productivity is implicit in neoclassical theory, but to understand it fully we must understand more clearly just what we mean by a firm. It is a peculiarity of Coase’s and Knight’s discussions, and of much of the new literature on the firm, that precise definitions of the firm are lacking. “Managed coordination” is used to represent the firm. These discussants treat the firm as if the managed coordination of resources is unique to the firm, but managed coordination is not so clean a cutting edge as their discussions suppose. Management takes place in households, which, I presume, are not firms. Furthermore, market transactions, which certainly are not regarded as firms in this literature, almost always require at least a modicum of management. A customer instructs a clerk to select one item from the shelf and not another; at purchase, he instructs the clerk to use a particular credit card and to put the item in a bag. The clerk responds as if he is being managed by his department head rather than by the customer. An investor instructs her stockbroker to check on the correctness of information that has come to her attention, to enter orders in certain amounts at certain times, and to call back with verifications of transactions. Do we have managed coordination within a firm or managed coordination across a market? Moreover, if market transactions involve conscious management, so conscious management involves market transactions. The firm is itself a nexus of contracts. So, if transaction cost rises, there is a substitution in favor of managed coordination only if contracts used to form the firm do not rise in cost as much as do other transactions.

A reexamination of the firm in neoclassical theory

The firm as conceptualized by neoclassical theory is not a mere appendage to the theory; it is an important aid to the theory’s search for understanding of the price system. To understand why this is so, one must recognize the nature of
The firm in theory: its definition and existence

The coordination problem that neoclassical theory tackles. The problem is to see how the price system works, and the task of the price system is to cope with the interdependencies that exist in a modern economy. The theory confronts this problem by constructing a hypothetical economy in which people depend on others. The construction depends on two characteristics of economic activity: extreme decentralization and extreme interdependency. Extreme decentralization deprives all firms and households of influence over price. So they do not set price; the system does. This aspect of neoclassical theory is well understood. The need for interdependency is not.

Self-sufficiency, the opposite of interdependency, is production for one’s own consumption. An economy populated by self-sufficient persons poses no social coordination problem with respect to resource use (although it does with respect to the distribution of wealth). The social dimension of the production problem contemplated by neoclassical theory comes from reliance on others, especially on strangers. It therefore serves this theory to populate its hypothetical economy only with persons who are not self-sufficient, especially with persons who are dependent on strangers. The importance of interdependency to an understanding of the price system becomes obvious in a comparison of the following two scenarios:

1. Robinson Crusoe, stranded alone on his island, must decide how much water to consume from his inventory of fresh water and how much time and effort to allocate to securing new supplies of water.

2. Hundreds of thousands of individuals in a populous society secure water from thousands of different persons who own and operate wells.

The first scenario involves decisions but poses no social coordination problem. Plans are conceived and executed by a single person. If Crusoe makes no calculating errors, his plans lead to appropriate actions and predictable outcomes. The chain from decision to action to outcome is direct, open, and conscious, and, I may add, it is managed coordination. The fact that Crusoe manages his own time and effort, rather than someone else’s, does not deprive his decisions, plans, and activities of the “managed” label, especially if there exist potential intertemporal inconsistency problems in Crusoe’s behavior. Management does not depend on the existence of complex organization. But self-sufficient management lacks the interesting social dimension that motivates neoclassical theory. The potential for conflict and inconsistency between Crusoe’s use of resources and the use of resources by others does not exist.

The second scenario poses a significant social coordination problem. Water is produced from many sites by persons whose activities are not coordinated in any obvious way, and this water is made available to many other persons who
8 The economics of the business firm

do not consciously plan or control water production. Is there any reason to believe that sensible quantities of water are produced and consumed? This type of question is a central concern to neoclassical theory. To face this question squarely, a hypothetical economy is modeled so as to be so interdependent that no one is self-sufficient in the sense that Crusoe is.4

This is accomplished with the aid of two “black boxes”: the household and the firm. The household sells its services to others and buys goods from others. It does not self-employ resources to produce goods for its own members; it offers its resources to firms. Firms buy or rent these resources, and they produce goods that are not for consumption by their owners and employees as such, but are for exclusive sale to households. The role of prices in accommodating this high degree of interdependency is of interest, not the manner in which households and firms manage their internal affairs. The contribution made by the household and the firm in this theory is to make the price system deal with extreme interdependency and decentralization. “In-the-household” production and “on-the-job” consumption are ruled out. This is implicit in the circular flow diagrams that once were a popular expository device of basic economics texts, but the theoretical significance of the circular flow was poorly understood. The texts mistakenly emphasized the circularity of the flow rather than the interdependency and decentralization exhibited in the transactions it pictures.

The firm as a specialized production unit

For lack of a better word, I use specialized to describe a production unit that produces for outsiders; this usage is different from specialized in the sense of doing only one thing, although the narrowing of tasks ordinarily accompanies specialization in the sense of doing for others. The firm in this theory is not just a black box, it is a specialized black box. It may do more than one thing within its boundaries, but whatever it does is directed toward use by outsiders, not insiders. No attention is paid to the complexities of managing production (or, as a counterpart, to the problems of dispute resolution within households). The role of this concept of the firm is to separate production from consumption, so that Crusoe-type self-sufficiency is absent. The coordination system for linking production and consumption that remains when self-sufficiency is absent is comprised of but two components: impersonally determined market prices and personally defined tastes. The two react to each other as described in neoclassical theory. The perfectly competitive firm is thus an important building block for a scenario that puts the price system to a test in which it is the only coordination mechanism for joining production to consumption. The firm in

4. Persons remain self-sufficient to the extent that they make their own decisions relative to the prices they face.
The firm in theory: its definition and existence

This theory need not be an organization at all. A single owner-manager-employee is all that is needed, and indeed neoclassical theory pays no attention to the organization problems that abound within real firms. Organization unnecessarily complicates things, when all that is needed from the firm is that it separate production from consumption; production must be exclusively for consumption by outsiders. What is needed is a concept of the firm in which production is exclusively for sale to those who are formally outside the firm. This requirement defines the firm (for neoclassical theory), but it has little to do with the management of some by others. The firm in neoclassical theory is no more or less than a specialized unit of production, but it can be a one-person unit.

The transaction and specialization theories compared

This concept of the firm differs from that widely used in contemporary literature on the firm and in the older works of Knight and Coase. In that literature, firms and markets are viewed as substitutes because the emphasis is on alternative ways of coordinating the use of resources. Here the emphasis is on specialized production, not on the internal organization of the firm. From this perspective markets and firms are not substitutes; if there is a relationship between the two, it is a complementary relationship. Markets do not produce goods for others, because they do not produce. Hence, in the specialization theory of the firm, markets cannot substitute for firms, and neither can the price system. Markets are conceptualized arrangements for matching bids and asks, for exchanging entitlements, and for revelation of the prices that accomplish these exchanges, but the activities that accomplish these things are housed in firms. These firms may be brokers, members of organized exchanges, contract lawyers, financial news gatherers and promulgators, etc. All are specialized producers of exchange services who may or may not rely heavily on managed coordination in the production of these services.

In a world in which everyone possesses perfect information about prices and technologies, price coordination is a pseudonym for self-management, so that what might be considered as substitutes are self-management and management by others. Prices do not coordinate; they supply information. Each owner of resources, knowing all there is to know in a perfect information world, can self-manage his or her resources, placing them in their highest value uses. As information becomes costly and less perfect, self-management gives way to management by others because this is an efficient way to take advantage of specialized knowledge. The relevance of specialized knowledge to the internal organization of the firm is a topic to which we return in the next commentary. The specialization theory gives an entirely different perspective on the existence of firms from that given by the transaction cost theory. In transaction
10  The economics of the business firm

cost theory, if transaction cost becomes lower (relative to management costs),
activities previously undertaken within firms are now accomplished in markets.
In the limit, if transaction cost goes to zero, firms cease to exist and all activities
are accomplished by markets; this is interpreted as the substitution of markets
for firms. And if transaction costs are prohibitively high, the market is aban-
doned in favor of within-the-firm allocation of resources. In the specialization
theory, a reduction in transaction cost causes firms to subdivide, with each firm
tending to focus production on a smaller part of the goods spectrum. This
subdivision is what the transaction cost theory calls the substitution of the
market for the firm, but the specialization theory calls it only an increase in the
number of firms. While it is true that the interface between these new firms is
not managed (if these firms are price takers), whereas before the subdivision the
interface was managed, and this make it appear as a reduction in reliance on
managed coordination, it is also true that the larger number of firms require
more central management units. Hence, one cannot claim that there is less
reliance on managed coordination if transaction cost is reduced. And if trans-
action cost goes to zero, firms do not disappear from the face of the earth. They
cannot, in the specialization theory, because goods are still produced for sale
to others. Firms simply achieve a maximum feasible division of activities or a
maximum feasible number of separate management units.5

The specialization theory, analyzing a reduction in transaction cost, views
the interposition of a market between activities previously carried on within one
firm, not as the substitution of market for firm, but as the substitution of two
firms for one. Assuming that specialization is productive, this substitution can
increase both the total output of the economy and the fraction of this total that
is produced in firms (as compared to the fraction produced in self-sufficient
fashion). The degree to which persons rely on others (strangers) is increased as
a result of a reduction in transaction cost. Reliance on markets also increases,
but not by eliminating firms. The number of firms and the degree to which they
sell to outsiders increase, as does reliance on markets. The importance of firms
and of markets in the economy correlates positively.

From the perspective of the specialization theory, an increase in transaction
cost reverses this process; the number of firms decreases. This change is also
called for by transaction cost theory, but the interpretation is again very dif-
f erent. Transaction theory interprets this as a substitution of firms for markets
and therefore as an increase in the importance of managed coordination. Spe-
cialization theory treats it simply as a reduction in the number of firms and in
the number of independent management units, but not as a substitution of

5. Just how many firms might exist if transaction cost were zero depends on considerations of
economies of scale and scope, but each firm would tend to specialize more than if transaction
cost were positive; “specialization” here means that production is focused on a smaller number
of goods.