Foundations and Trends® in
Technology, Information and Operations Management
Vol. 5, Nos. 3–4 (2011) 147–309

DOI: 10.1561/0200000029

The Economic Foundations of Supply Chain Contracting

By Harish Krishnan and Ralph A. Winter

Contents

Foundations and Trends® in
Technology, Information and Operations Management
Vol. 5, Nos. 3–4 (2011) 147–309

DOI: 10.1561/0200000029

The Economic Foundations of Supply Chain Contracting

Harish Krishnan[1] and Ralph A. Winter[2]

[1] *Sauder School of Business, University of British Columbia, 2053 Main Mall, Vancouver, British Columbia, V6T 1Z2, Canada, harish.krishnan@sauder.ubc.ca*

[2] *Sauder School of Business, University of British Columbia, 2053 Main Mall, Vancouver, British Columbia, V6T 1Z2, Canada, ralph.winter@sauder.ubc.ca*

Abstract

Why do supply chain contracts take the forms that they do? Which contracts should firms adopt to coordinate incentives along a supply chain? This monograph synthesizes the theory of contracts along supply chains. It integrates developments from two largely separate literatures, the management science literature on supply chain coordination and the economic literature on vertical control.

1

Introduction

1.1 Setting the Stage

A supply chain is the sequence of firms involved in the production of a product or service, from the procurement of raw materials through the production of intermediate inputs to the distribution of the product to ultimate buyers. Supply chain management involves all economic decisions along this chain, including product design, the choice of inputs at each stage, the choice of which suppliers to use, transportation of both inputs and final products, inventory decisions at each point of the chain, and ultimate pricing. No single firm controls all decisions along the supply chain. Many firms, each with their own management and shareholders, make decisions that must be coordinated.

Real-world contracts reflect the need for this coordination. We observe in supply chains an extraordinarily rich array of agreements. Contracts may specify complex nonlinear pricing schedules. Contractual restrictions may be imposed on input suppliers well beyond the obligation to supply at a specified price. Restrictions imposed on downstream firms include restraints on prices or territories, exclusivity requirements of various types, or minimum quantities. We observe inventory risk-sharing arrangements, e.g. buy-back contracts in which an input supplier agrees to purchase unsold inventory; revenue-sharing and other royalty agreements; loyalty contracts, including market share discounts; and so on.

Why do supply chain contracts take the forms that they do? Which contracts should firms adopt in a given market environment? Existing surveys of the theory underlying these questions approach the area from either the management science perspective (e.g., [33]) or from the perspective developed in the economics literature (e.g., [97]).[1] This separation reflects a division in the community of scholars working in the area, coming largely from either management science or economics. In reference to one of the principal economic approaches to the question, the transactions-cost-economics (TCE) approach, Williamson [200] writes "...some cross-referencing between the TCE and supply chain literatures notwithstanding, these two are mainly disjunct. Arguably, the complementarities and tensions between them should be more fully worked up [and] this could be the beginning of a constructive conversation."

1.2 The Aim of This Monograph

Like other areas in operations management and management science, the theory of supply chain decisions has expanded from its traditional domain of operations optimization by a single decision maker towards the coordination of the incentives and decisions of multiple firms. The theme of this monograph is that as supply chain management moves from a focus on optimization problems to issues of coordination, a closer link to the underlying economic foundations is essential. We offer a synthesis of the economic foundations of supply chain contracts.

A review of all of the relevant economics would require volumes. Accordingly, our treatment is selective, incorporating elements of economic theory that we believe will be of most value to our intended readers, students and scholars of management science and operations management. "Economic foundations" herein refers to the alignment

[1] Our focus is on the theory of supply chain contracts. Lafontaine and Slade [116] offer an excellent review of the empirical evidence on inter-firm contracts.

of incentives by contracts to achieve maximum benefits for firms along supply chains.

1.3 Our Approach

The conceptual starting point for understanding contracts along a supply chain is, ironically, a theoretical benchmark under which there are *no* contracts. Consider a good being produced in a perfectly competitive upstream market, and distributed by a perfectly competitive downstream market. (Let us call the firms "manufacturers" and "retailers" for concreteness.) The only decisions in this ideal market environment are quantity decisions, as prices are outside the control of any single firm. And these quantity decisions are coordinated perfectly along the supply chain by the price system. Indeed, in an economy with perfect markets the price system alone conveys all the information that is needed for individuals to make decisions that are in the best interest of the society as a whole [52, 178]. The efficiency of simple price-mediated exchange along a supply chain under perfect market conditions is Adam Smith's "invisible-hand theorem" writ small. With the price system acting as an invisible hand there is no need for any inter-personal or inter-firm contracts at all. Firms along a supply chain make the same decisions *as if* they costlessly met and decided on each detail of their production plans. Contracts play no role.

The price system in reality, not just in theory, provides the central mechanism for coordination of decisions along a supply chain. Consider, for example, the decisions of all parties who produce inputs into a pencil: the graphite miners, the lumberjacks, the mill operators, the final producers and the retail distributors. These parties are not members of a single, huge multi-party contract or planning committee. They interact anonymously for the most part, each maximizing profit given prevailing prices [159]. In maximizing its own interest given the prices along the supply chain, each party makes more or less efficient decisions.

But not perfectly efficient decisions. If the price system functioned as well in reality as in the abstract theory then we would see no contracts. The invisible hand theorem, in other words, tells us that we must depart from the perfect market benchmark to explain contracts.

Introducing any particular deviation from the perfect market setting gives rise to specific incentive distortions, i.e., specific failures of the price system. The economic theory of supply chain management can be thought of as a mapping from "imperfections" in the economic conditions, to incentive distortions, and then to contracts that optimally resolve the incentive distortions:

$$\text{Market imperfections} \quad \rightarrow \quad \text{Incentive Distortions} \quad \rightarrow \quad \text{Contracts}$$

In the domain of this mapping lie a large number of potential market imperfections. One is the *market power* that firms have in setting prices. Firms rarely take prices as outside their control. A second departure from the ideal world arises from *uncertainty*. Demand and costs are never entirely predictable. This would create no incentive problems with a sufficiently rich set of futures markets and insurance markets, as in the Arrow–Debreu model.[2] In reality, this set of markets is incomplete. The price system fails, and the coordination problem arises, whenever some markets are missing.[3]

One source of missing markets is *asymmetric information*. Downstream firms such as retailers are often better informed about the state of demand in their market than upstream producers. Alternatively, as in the case of innovators with special knowledge of the value of their innovations, information may be superior at the upstream stage. Exchanges cannot be made contingent upon events that are not jointly observed. Consumer information is also limited. Consumers may be influenced in their demand by retailers' actions such as sales effort, or the provision of information. Outlets must attract consumers through advertising and other product promotions because of limited consumer information. These are merely some examples of departures from the ideal of perfect markets.

[2] Specifically, the economic theory of exchange in perfectly competitive markets assumes a complete set of markets for contingent commodities [52] or a complete set of securities markets with ex post markets for goods [11]. A contingent commodity is a good specified not just by its physical characteristics but also by the date and "state of the world" in which it is to be delivered.

[3] The phrase "missing markets" encompasses all imperfections including market power, which is the absence of competitive markets.

Moving from the *domain* of the mapping to its *range* (observed contracts), contractual relations can be thought of as falling along a spectrum representing the degree of centralization. At one end of this spectrum is uniform pricing: a contract in which the seller states a price and the buyer chooses a quantity. At the other end is the vertically integrated or centralized firm.

In the supply chain literature both within the theory of management and traditional neoclassical economics, the concept of vertical integration or a centralized firm owning the entire supply chain has come to mean, usually implicitly, a contract in which *all* decisions are taken in the interest of the integrated firm. For example, the literature often asks when particular contracts can achieve the "first-best" profits that would be earned by a vertically integrated firm coordinating all decisions at zero cost.[4]

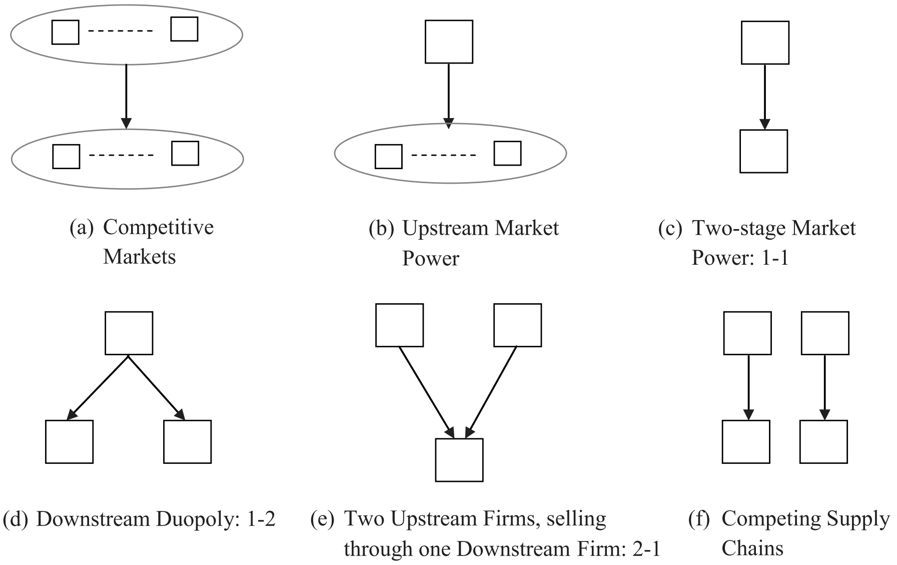
Against the benchmark of the centralized firm one can assess the performance of "minimally intrusive" or "minimally sufficient" contracts that restrict a smallest subset of the actions of contractual parties. The search for the simplest contracts that can achieve the centralized solution is an implicit recognition of costs of writing and enforcing complex contracts. Minimal contracts are more easily enforced than contracts dictating *all* the actions of agents involved. For example, in the classic paper by Pasternack [153], a contract providing for the buyback of a retailer's inventory at an agreed upon buyback price will elicit first-best retailer decisions on inventory. Implicit in this theory is the assumption that contracting directly over inventory is infeasible, or problematic for reasons outside the model.

In systematically analyzing contracts in this monograph, we will identify the source of the failure of the simple price contract in terms of the externalities introduced by the market conditions. An externality,

[4] The hypothetical costless, complete contract represented by the centralized firm is a very useful benchmark. In reality, however, incentive distortions arise even *within* the firm. The decision between undertaking a particular set of transactions through a market or within a firm involves a tradeoff between the costs of imperfections in the market — *transactions costs*, in economic terminology — with the transactions costs of allocations within the firm. While the main focus of this monograph is on the set of inter-firm contracts that align incentives along a supply chain, we will synthesize as well key contributions to the Coasian (1937) question of whether to transact in a market or within a firm.

in our context, is a failure of a firm (or individual) to capture the full benefits and costs of its decisions to contract partners along a supply chain. Having identified particular externalities at the heart of the "market failure" of the price system, we then design contracts that resolve the distortions in a minimally intrusive way. Note that while perfectly competitive markets yield the socially efficient outcome, contracts that arise as a response to missing markets will be structured to achieve privately efficient outcomes for the firms with market power. The use of contracts to achieve privately efficient outcomes may or may not increase social surplus. The focus of this monograph is on the private incentives for coordination and not on the appropriate policy towards restrictions on contracting. Iacobucci and Winter [89] offer an overview of the law and economics of placing imposing restrictions on the contracts that market participants may enter.

We organize the synthesis of supply chain contracts according to a single dimension of the market environment: the market structures for which contracts are designed. These are depicted in Figure 1.1.



(a) Competitive Markets

(b) Upstream Market Power

(c) Two-stage Market Power: 1-1

(d) Downstream Duopoly: 1-2

(e) Two Upstream Firms, selling through one Downstream Firm: 2-1

(f) Competing Supply Chains

Fig. 1.1 Market structures.

For brevity, we refer to the market structures as 1–1 for the two-stage monopoly; 1–2 for the upstream monopoly facing a downstream duopoly; and 2–1 for the upstream assembly problem and common agency problem in which two upstream firms face a single downstream firm. In our review, we take the market structure as exogenous for the most part.

1.4 Plan

As practical background, we provide in Section 2 an overview of evidence on the nature and frequency of specific supply chain contracts. We then offer in Section 3 some brief remarks on methodology concerning the application of economic theory to supply chain contracting. The basic setting, perfect markets, is reviewed in Section 4 of the monograph. The simplest departure from perfect markets is the introduction of market power, which we examine in Section 5 via the assumption of a single monopolist upstream, facing a competitive downstream market. Section 6 considers contracts in a standard framework: one firm operates at each of two levels of a supply chain. Section 7 adds imperfect competition downstream. Section 8 considers contracts in a setting with a single downstream firm and multiple upstream firms, including the case of a single incumbent firm facing potential entry. Section 9 reviews the role of contracts in competing supply chains (each chain with a single firm at each level). In Sections 10 and 11 we offer overviews of the dynamics of supply chain contracting as well as an explicit asymmetric information approach to contracting. Section 12 reviews the key contributions to the fundamental issues of vertical integration, investment in specific assets, and long run or relational contracting. We consider as well the economic theory on the role of reputational forces in resolving incentive distortions. Section 13 concludes the monograph with an overview of additional issues in the economics of supply chain contracting.

2

Background: Supply Chain Contracts in Practice

This section delineates the range of contracts — pricing strategies, sharing strategies, restrictions and additional contractual rights and options — that are adopted in practice. We review a wider range of contracts that has been analyzed in the management science literature to date.

The simplest contract between a buyer and a seller involves *uniform pricing*. Under this contract, an upstream firm sets a price for an input at which downstream firms are free to purchase any quantity they want. At the other end of the spectrum of contractual relations is complete vertical integration, in which all aspects of a transaction are internalized within a single firm. We review contracts lying along this spectrum. We also discuss the feasibility of various contracts in terms of monitoring requirements, enforceability and legal requirements. (We refer to any contract beyond uniform pricing as a *complex contract.*)

2.1 Pricing Strategies

While uniform pricing is the simplest supply chain contract, "pricing" captures a wide range of strategies. We discuss here two dimensions along which pricing contracts vary: the mechanisms for adjustment of prices over time and the various forms of non-linear pricing.

Price adjustment clauses: As Lafontaine and Slade [116] note that even uniform pricing contracts can be long term, and often allow for the adjustment of the prices to changing circumstances. In *cost-plus contracts*, prices automatically adjust to changing costs. The wholesale price may be set at a specified fraction of the retail price. In general, clauses that specify how prices may be adjusted over time can be categorized into redetermination and renegotiation clauses [48, 116]. Redetermination clauses specify formulae that determine precisely the adjustment of prices. The price may increase over time at a fixed rate; it may be tied to a raw-materials price index; or to the spot price of an output. Renegotiation clauses, on the other hand, specify a process rather than a rule for determining an adjustment in the price. A renegotiation clause may come into force once costs or external prices have gone outside specified limits.[1] A renegotiation clause, or even a clause related to pricing of an input to be developed in the future, may contain seemingly vague phrases such as "on commercial terms" or "fair, reasonable and nondiscriminatory".[2] The clause may specify a process for price offers and responses during negotiation, and may allow for arbitration.

The price adjustments in a contract may be a function of prices in other contracts. For example, a contract between a buyer and a seller may dictate that if another buyer (e.g., a rival) receives a better price from the seller, the buyer in the contract is refunded the difference in prices. In other words, the buyer is guaranteed the lowest price for the seller's input. This is referred to as a *most-favored-nations* (MFN) clause in a contract. Parallel to the MFN contract is a contract that refers to future opportunities that may arise for the buyer in the market. A *meeting-competition clause* guarantees to the buyer that the seller

[1] For example, Goldberg and Erickson [71] in a study of contracts for petroleum coke, in which over 90% of the contracts contained a price adjustment, found some contracts that called for renegotiation when the crude oil price was above or below some limits. Other contracts allowed for price adjustment indexed to crude oil prices. After the oil crisis of 1973, when market volatility increased substantially, many indexing clauses were replaced by renegotiation clauses.

[2] Fair, reasonable, and non-discriminatory terms (FRAND) is a phrase often required by standard-setting organizations for members that participate in the organizations. (Standards ensure compatibility and interoperability of devices manufactured by different companies.)

will match future prices that become available to the buyer from other sources.

Nonlinear pricing: A *general nonlinear pricing schedule* is an increasing function $R(q)$ announced by a seller. Buyers react to this announced schedule with their individual choices of quantity and payment, $[q, R(q)]$. Non-linear pricing can thus be interpreted as a menu of quantity–payment pairs from which the buyer can choose.

The class of nonlinear pricing schemes includes many specific subclasses. The simplest class is *quantity discounts*. *Block pricing* refers to the setting of piece-wise linear revenue schedules, and is used in electricity pricing, for example. *Two-part pricing* involves the setting of a fixed fee, f, in addition to a constant variable price p, i.e., $R(q) = f + pq$. Two-part pricing is common in franchise contracts and observed throughout the economy [112]. Credit cards that charge an annual fee plus a per-transaction fee, membership discount stores that require an annual fee for admission, telephone service with a monthly fee and a call fee per minute, are all examples. Amusement parks charge for admission as well as for individual rides within the parks (hence the name "Disneyland Pricing" sometimes attached to two-part tariffs).[3] A generalization of two-part pricing is the offer to buyers of a *menu of two-part pricing schedules*, from which buyers may choose. Any convex nonlinear pricing schedule can be implemented via the offer of a menu of two-part prices [190].

The fixed fee in two-part pricing contracts need not be positive. *Slotting allowances* are fixed fees paid by grocery manufacturers for the right to have their products carried by stores. Sometimes the fixed payment is in exchange for the right to prominent shelf or display locations [172].

Nonlinear pricing includes as well *take-or-pay contracts*, in which the seller receives a minimum payment for a specified quantity, whether or not that quantity is taken by the buyer. In this case $R(q)$ includes a fixed fee, with $R'(q) = 0$ up to the minimum purchase quantity, and

[3] "Disneyland pricing" refers specifically to two-part pricing in which the variable price exceeds marginal cost [146]. We note that Disneyland Inc. no longer uses Disneyland pricing.

may be linear beyond that quantity. In a retail market, this is essentially equivalent to *quantity forcing*, in which a downstream retailer is required to sell a minimum amount of an input. Contracts are observed in the food industry and elsewhere in which the buyer has an option to purchase an amount up a given quantity, B, at a fixed price, and the supplier has the right to sell a minimum amount, $A < B$, at the fixed price. This is an example of explicitly incomplete contract, in which prices for amounts outside the range $[A, B]$ are left up to ex post negotiation.

We set aside, for the most part, the richest set of pricing strategies: strategies aimed at price discrimination. The incentives for price discrimination, reviewed by Varian [196], has not been central in the literature on supply chain contracting.

2.2 Sharing Contracts

By sharing contracts, we mean contracts in which the payment from a buyer to a seller is a function of variables other than quantity purchased. *Royalty payments* depend on the quantity sold or revenue earned downstream by an input purchaser. Franchises typically pay revenue royalties, and video rental stores enjoyed a surge in profitability when their contracts with movie distributors were switched to revenue royalties [38, 140]. Contracts that involve the transfer of rights, such as patent rights, often contain output royalties [66].

Share-based royalties depend on the share of the buyer's purchases of a class of inputs that the seller provides. For example, a discount in price may be offered if the buyer agrees to use the seller's input for more than 90 percent of its needs. This is one form of a *loyalty contract.*

2.3 Options

Contracts can contain many kinds of options that affect the ultimate payment by the buyer to the seller. A *right-of-first-refusal* is an option for the seller to meet the price offered to the buyer by another seller. A buyer or a seller may have the option to purchase or to sell additional units at specified prices. In the case of the buyer's option, this is simply part of a nonlinear pricing schedule; the seller's option represents is distinct from nonlinear pricing.

Inventory buybacks and returns are often observed in supply chains. Wholesale suppliers of books, magazines, newspaper and other products often accept returns of unsold inventory, at a contractually agreed upon buyback price. Beyond printed material, returns policies are used for recorded music, computer hardware and software, greeting cards, and pharmaceuticals [152]. Buyback options are used most often when the product has a very low marginal cost of production relative to market value, as these examples suggest. In Japan, returns policies are used more often than in the U.S., and for products including consumer durables, furniture and even soft goods such as apparel [131]. Under a *consignment* contract, the manufacturer maintains control over retail inventory levels and full responsibility for all unsold inventory. A *quantity flexibility* contract obligates a manufacturer to meet orders within specified lower and upper limits, essentially placing a bound on the amount of unsold inventory for which the manufacturer is responsible.

Channel rebates are arrangements where a manufacturer reimburses a retailer based on the units sold. Under a *linear rebate* scheme the retailer is paid for each unit sold while a *target rebate* offers a per-unit reimbursement for each unit sold above a target level. *Capacity reservation* contracts are option contracts allowing a firm to reserve production capacity at its supplier. In a model with uncertainty and inventory, the amount that a retailer sells can differ from the amount that the retailer buys. All the contracts discussed fall within the class of (possibly nonlinear) payments to the manufacturer as a function of the amount purchased and amount sold at the retail level.

2.4 Vertical Restraints

Vertical price restraints: Moving beyond the payment form itself, supply chain contracts often involve restrictions on the actions of parties, the downstream firm in particular. *Vertical price restraints*, or *resale price maintenance* involve restrictions (most often a price floor) imposed by an upstream firm on resale prices set by a downstream distributor. Vertical price floors have been imposed on retailers of a wide cross-section of products: clothing, skis and other sports equipment, watches, jewelry, luxury goods of all kinds, candy, beer, bread, floor wax, furniture polish, milk, toilet paper, cereal, canned soup, books, shoes, mattresses, large appliances, and automobiles, to name a few [90, 150]. Products in virtually every category have been subject to resale price maintenance at one time or another, and estimates of the proportion of retail sales that have been subject to resale price maintenance range as high as 25% in the U.K. and 4%–10% in the U.S. [168, p. 549]. In Canada, before the law prohibiting resale price maintenance was enacted in 1951, an estimated 20% of goods sold through grocery stores and 60% sold through drugstores were 'fair-traded' [150, p. 153, 155].

Restraints may be imposed on *relative* prices rather than absolute prices. A restraint may restrict the size of a discount that can be offered on one product variety, for example. Or a restraint may specify that a downstream retailer not sell the upstream firm's product at a higher price than the price for a product of rival firms. For example, restrictions by some credit card companies prohibit merchants from charging more for transactions under one credit card than under another.[4]

Customer restrictions involve restraints on the set of buyers to whom a reseller may sell a product. Sometimes, for example, an upstream supplier may retain large customers to deal with directly rather than have those customers contract with a downstream distributor of the supplier's products.[5] *Territorial restrictions* involve contracts

[4] Restraints in contracts are traditionally viewed as strictly enforced constraints, designed with the expectation that parties to the contract will maximize profit subject to the restraint. The property-rights approach to contracts [75, 80, 82] reviewed in Section 12 of this monograph takes a different perspective on some restraints. To take an example, consider an exclusivity restraint providing that a dealer be the sole reseller within a given geographical area. The property-rights theory (PRT) interprets this as the assignment of decision rights — the right to add a second dealer to the area — to the dealer in the contract. The efficient decision as to the entry of a second dealer will be taken ex post, as the upstream manufacturer and first dealer will maximize joint profits, but a larger share of the gains from the additional dealer accrue to the first dealer under the territorial "restraint". The restraint is more accurately seen as an initial assignment of decision rights or property rights to the extent that renegotiation is a possibility. See ref. [136] for the empirical test of a PRT model applied to vertical territorial restraints.

[5] The White Motor Company, a truck manufacturer, prevented downstream resellers from selling to large customers; see *White Motor Co. v. United States, 372 U.S. 253 (1963).*

on where a retailer (e.g., a dealer) may locate or sell. In some cases, territorial restrictions go as far as disallowing sales to customers located outside the dealer's designated area. Territorial restrictions represent a combination of a restriction and protection. Because they are also imposed on other resellers, the restrictions offer the reseller protection against horizontal competition. In some contracts, e.g., franchise contracts and automobile dealership contracts, this protection is provided by a slightly different form of territorial restriction: that the upstream manufacturer cannot sell a new franchise or dealership within a specified distance from the franchise under contract. In some cases, territorial and price restraints may be combined: in *Eastern Scientific Co. v. Wild Heerbrugg Instruments*[6] the defendant distributed its product through a system where each distributor was bound by a price floor outside its territory but not within.

Bundling of two products refers to the sale of two products as a package, in a fixed ratio of quantities. *Mixed bundling*, a variation on this, refers to discounts offered if products are purchased as a bundle, so that the buyer has the option of purchasing a single product or a package. Mixed bundling is familiar in fast food restaurants, for example.

Requirements tying, or simply *tying*, stipulates that the buyer purchase all of its requirements of an input B from the seller if the buyer is to purchase input A from the seller. IBM, for example, required purchasers of its adding machines to buy all of their requirements of cardboard cards from IBM as well. Standard Stations supplied gasoline to dealers on the condition that the dealers purchase their entire requirements of other inputs from a specified set of suppliers.[7] Franchisors have sometimes required that franchisees purchase all input requirements from them when this restraint has been legal.[8]

Exclusivity restrictions may be imposed on downstream buyers or on upstream suppliers in a contract. That is, buyer's purchases from another seller may be ruled out, or sales to rivals of the buyer may be prohibited. Of retail sales through independent retailers, more than one third was found to be subject to some form of exclusive dealing

[6] F.2d 883 (lst Cir.), cert. denied, 99 S. Ct. 112 (1978).

[7] *Standard Stations of California v. U.S.*

[8] See *Chicken Delight, Inc.*, 448 F.2d 43 (9th Cir. 1971).

in a recent study [114, 194]. A well-known example of exclusivity is in the *Standard Fashions* case,[9] in which the supplier of dress patterns reduced its wholesale price, by half in some cases, in return for exclusivity on the part of department stores.

2.5 Feasibility of Various Contracts

2.5.1 Monitoring and Arbitrage

This monograph explains the *incentives* for contracts of various forms in supply chains, but the choice of contracts depends of course on which restraints or payment schemes are feasible. Some discussion of the feasibility of various contracts is therefore important.

If a seller has no control over arbitrage, that is, if resale of its products among any parties can take place at zero transaction costs, then uniform pricing is the only feasible contract. No more complex contract can be implemented. An attempt to implement a nonlinear pricing scheme, for example, would lead arbitrageurs to purchase the amount that minimized the average price paid per unit, then re-sell the units to downstream firms at this average price. And any attempt to impose a restraint on a downstream firm would be undone by resale to firms that had no contractual relationship with the supplier and were therefore free of the restraint.

If the monitoring ability of the upstream supplier is limited to detecting sales of a product by agents outside a designated set of purchasers (but the manufacturer cannot monitor resale among the designated agents) then two-part pricing is the most complicated contract feasible. Any attempt at a more complex nonlinear pricing scheme would lead to arbitrage among the downstream firms that have paid the fixed fee.

Contracts with restrictions require monitoring even beyond the prevention of arbitrage. The manufacturer can impose a restraint on prices, territories, tying, customer restrictions only if it can monitor whether the restraint is being followed. For example, if a manufacturer can monitor whether or not a downstream reseller is selling its rival's products

[9] See *Standard Fashion v. Magrane-Houston Co., U.S., 1922.*

(but not necessarily how much of those products it sells) then it can enforce exclusivity contracts. Similarly for upstream exclusivity contracts. If the manufacturer can determine from the reseller's accounts not just whether the retailer is selling rivals' products but how much of its sales are in rivals' products, and the share of the reseller's sales that it captures, it can enforce market share contracts. In the case of inventory buybacks of paperback books or magazines, the retailer may be required simply to rip off the front cover and return it (as proof of inventory not sold).

In some cases, the monitoring itself is decentralized. Resale price maintenance, in the case of a vertically imposed price floor, is monitored by competing retailers, who have the incentive to report violations of a resale price maintenance agreement. (Resale price maintenance cases often start with complaints by one retailer about violations of a resale price maintenance agreement by another retailer.) The same cannot be said of vertical price ceilings.

2.5.2 Legal Restrictions

The history of antitrust restrictions on vertical restraints is long, complicated and the subject of many papers and books itself. The trends in antitrust law in this area have been heavily influenced recently by economic analysis of the efficiency roles of restraints in supply chain contracting. The central issue in legal cases involving vertical restraints is increasingly whether the restraint in question is designed to serve the role of efficiently coordinating incentives along a supply chain, or whether it is designed to lessen competition among rivals at some stage of the supply chain.

For scholars analyzing the design of optimal supply chain contracts, we summarize briefly the law on vertical restraints. First, territorial restraints have been effectively per se legal since *Continental Television v. GTE Sylvania*, 433 U.S. 36 (1977). Vertical price *ceilings* have been effectively legal since *State Oil Co. v. Khan*, 522 U.S. 3 (1997) overturned *Albrecht v. Herald Co.*, 390 U.S. 145 (1968).

Minimum resale price maintenance, or vertical price *floors* are by far the most important vertical restraint [207]. The law on vertical price floors recently changed in the U.S. with the Supreme Court decision in *Leegin Creative Leather Products, Inc. v. PSKS, Inc., 551 U.S. 877 (2007)*. In this case the Court reversed a doctrine from 1911 that vertical price restraints were illegal per se under Section 1 of the Sherman Act.[10] *Leegin* replacing the older doctrine with the rule of reason under which the efficiency of the restraint and its potential anticompetitive effects must both be considered. The burden of proof in resale price maintenance cases requires further clarification in the common law, but appears to rest on the defendant, who must demonstrate that the practice has a legitimate role.[11]

While the status of resale price maintenance is generally described as per se illegal prior to *Leegin*, there was (and remains) an important exception to the restriction under the *Colgate* doctrine.[12] The Colgate doctrine, which has varied in scope over time, allows a manufacturer to establish a vertical price floor providing this is done *unilaterally* as part of the manufacturer's design of a distribution system. A manufacturer risks conviction under Sherman Act if it discusses the price floor with retailers (even prior to terminating a retailer) since this could be interpreted as a conspiracy to raise prices in violation of the Sherman Act. But is free to design a vertical price floor in a way that satisfies unilaterality. In practice, even after *Leegin* lawyers continue to design "Colgate programs" for their corporate clients for the purpose of ensuring that vertical price floors are covered by the Colgate exemption.

The upshot for scholars in this area is that in designing optimal contracts, vertical price floors should not be dismissed as a potential instrument for aligning incentives on the basis that it is prohibited. The law places some restrictions on the use of this instrument, but resale price maintenance is, after *Leegin*, more easily implemented. And the Colgate doctrine continues to offer significant scope for implementing the practice. For scholars focussed on the positive economic question of why firms adopt resale price maintenance, there is a very large set of cases available as evidence (see, e.g., [90]).

[10] *Dr. Miles Medical Co. v. John D. Park & Sons Co.*, 220 U.S. 373 (1911).

[11] See, e.g., Federal Trade Commission ruling Modifying the Order in the Nine West Resale Price Maintenance Case (May 6, 2008).

[12] See Supreme Court's ruling in *U.S. v. Colgate & Co., 250 U.S. 300 (1919).*

With respect to other vertical restraints such as tying or exclusive dealing and supply chain contracts in general, the law is, increasingly influenced by economic theory and analysis of the incentive-alignment role of supply chain contracts. Judging from recent legal history, scholars in both economics and management science can potentially influence the continued trend in the law towards stronger economic foundations. The analysis of supply chain contracts can add not just to the normative management question of the optimal design of contracts for supply chain efficiency, or the positive economic question of explaining existing contracts, but for a third issue: the normative policy question of the socially optimal legal or regulatory restrictions on the set of contracts.

3

Remarks on Methodology

The fundamentals of supply chain management draw on the economics of contracts and game theory in general. It is helpful to discuss at the outset a number of methodological issues regarding the application of the theory. The first of these concerns theorems on the existence of Nash equilibrium in supply chain management models.[1] It is common in applied microeconomics to adopt very stylized models. In industrial organization, the field of economics that the supply chain coordination literature draws upon, models often yield existence of equilibria for some parameter values but not for others. (The canonical model of Bertrand competition with differentiated products is an example. For this model a pure strategy equilibrium usually fails to exist if the degree of differentiation is positive but small.) Models in industrial organization including supply chain organization are useful simply to capture the logic and intuition of economic arguments. As long as existence of equilibrium can be assured for a wide range of parameters, the model

[1] Examples of papers that develop existence theorems include refs. [22, 45, 125].

is considered useful. Theorems on the general existence of equilibrium within supply chain models we believe to be of limited interest.[2]

A related methodological point is our application of what is called the "first-order" approach in contract theory. This approach assumes that first-order conditions are not only necessary but sufficient for the agent's optimal decisions within a principal–agent model. In setting up optimal contracting problems under this approach, incentive compatibility conditions are replaced by the agent's first-order conditions. Rogerson [163] and Grossman and Hart [74] provide conditions under which this approach is valid in the principal–agent model. The approach is common in applied incentive theory generally.

The first two methodological points, existence and the first-order approach, are connected by a common assumption of quasi-concavity of payoff functions. Quasi-concavity is sufficient for existence of pure strategy equilibrium in game theory models [61] and as a property of agent utility is sufficient for the first-order approach in incentive theory.

A third remark on methodology concerns the formulation of the optimal contracting problem. We assume for the most part in our synthesis that lump sum transfers are possible among firms and that the firms are risk-neutral. An optimal contract under these conditions and symmetric information at the time of contracting, can be characterized as maximizing the sum of expected profits. The assumption of wealth transfers is essentially that parties can set prices for *inframarginal* units different from the prices for *marginal* units. In practice, we observe many mechanisms to implement wealth transfers. Two-part pricing, quantity discounts, slotting fees — all of these instruments act to transfer wealth between parties without distorting incentives by changing marginal prices. The assumption that contracts maximize combined wealth, which dates back at least to Coase [43], is the natural first step to take in understanding the role of complex contracts. Any model that rules out lump sum transfers must — to be internally consistent — build in the reasons that such lump sums are not feasible. (One such reason would be the combination of uncertainty and

[2] In general equilibrium theory, in contrast, the most basic theorem of economics is the existence of an equilibrium that can support a Pareto optimal allocation of resources [52]. In general equilibrium models, existence theorems are fundamental.

limited wealth on the part of firms.) We offer some remarks at various points, especially in the section on asymmetric information, about the robustness of results to relaxing our assumption of transferable utility.[3]

Since our focus is on the simplest contracts that can achieve efficiency, and because we assume transferable utility, we do not focus on the allocation of profits among different firms in the supply chain. In contrast, bargaining and cooperative game theory approaches in supply chain management are concerned precisely with the issue of how profits are allocated in supply chains. In a recent review, Bernstein and Nagarajan [24] discuss both non-cooperative and cooperative models in supply chain management. The review of cooperative game theory and bargaining models presents results from the "atomic model" (our 1–1 structure) and then discusses bargaining solution concepts in more complex supply chain structures. With its focus on the allocation of profits in a supply chain, the bargaining approach is complementary to the non-cooperative approach that we follow. The approach we review in this monograph, mapping market conditions to observed contracts by characterizing externalities and distortions, can generate testable implications. For example, resale price maintenance can elicit efficient inventory levels in the 1–2 market structure under uncertainty, and would dominate uniform pricing.

A final remark on methodology concerns what we could call implicit versus explicit assumptions of asymmetry in information between a downstream firm and an upstream firm. In most of the models that we will discuss, the upstream and downstream firm can contract on a range of variables — price, quantity, etc. A complete contract would simply specify *all* decisions (as assumed to happen in a "centralized solution"). When we ask whether efficient inventory, downstream pricing or other decisions, can be elicited in the model, we are asking whether a contract *simpler* than the complete contract can achieve first-best incentives.

[3] Many papers in the supply chain contracting literature do not adopt the framework of joint wealth maximization and transferable utility. Lariviere and Porteus [117] restrict attention to a price-only contract and analyze the efficiency of supply chain decisions under this constraint on contracts — assuming a single manufacturer and a single retailer. Perakis and Roels [154] characterize the efficiency of various other supply chain configurations including assembly systems and supply chains with a single manufacturer and multiple retailers.

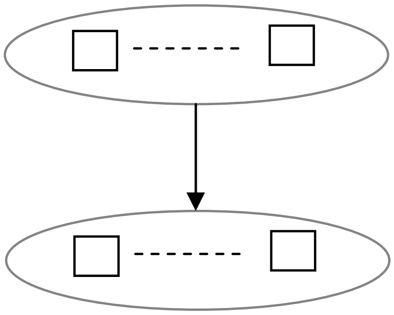
By assuming that the simpler contract is feasible but the full contract is not, we are implicitly relying on some form of asymmetric information or non-contractibility.[4]

All of these methodological points follow from our perspective on the basic purpose of economic theory in supply chain management. None of the models that we survey has been or could be used as a completely realistic and operational description of real-world supply chains. None could be used to estimate numerically the optimal parameters of a supply chain contract. Management science has a strong tradition of applying models to find optimal solutions to practical problems. In the area of supply chain contracting, however, as in the economics of industrial organization generally, the main role of models is to provide intuition and testable implications. To explain any particular incentive distortion and associated contract in supply chains, the simplest model is always preferred. The simplest model illuminates the forces at work with the greatest clarity.

[4] The distinction between complete and incomplete contracting has come to take on a second meaning in the contracting literature. A complete contract in the second sense is a contract that is the most general one possible under the information assumptions of the model. See ref. [191]. An incomplete contract involves an exogenous, arguably *ad hoc*, restriction of some kind on the set of feasible contracts. For example, assuming away any kind of lump-sum transfers is generally *ad hoc*.

4

The Benchmark: Perfectly Competitive Markets



The starting point for understanding the optimality of complex contracts in coordinating decisions along supply chains is, ironically, a world in which contracts are completely unnecessary: perfectly competitive markets. As we explain in the introduction, the most fundamental of all economic principles is under that ideal market conditions, the price system alone elicits incentives perfectly. This section sets out this theorem.

The perfect markets benchmark is abstract and a far cry from the complexities of real world supply chains. But as a conceptual starting point the theorem is essential. The benchmark set of conditions is an ideal from which departures will be taken to understand why supply chains contracts are organized the way that they are.

A perfectly competitive market is one with zero transactions costs, a homogenous product produced by all firms, perfect knowledge on the part of market participants, and zero market power in the sense that each firm takes prices as given. Consider a pair of vertically related, perfectly competitive markets; see Figure 1.1(a). In the downstream market, a large number of firms compete. All firms share the same production technology, described by a production function $q = F(x_0, x_1, \ldots, x_n)$ that gives the firm's output q as a function of the quantities of inputs $\boldsymbol{x} = (x_0, x_1, \ldots, x_n)$.[1] We assume throughout that F is differentiable and that it obeys Inada conditions: as x_i approaches 0, $\partial F/\partial x_i$ approaches ∞. (The Inada conditions guarantee interior solutions.) It will be convenient to assume that F is homothetic, i.e., that $F = h(G(\boldsymbol{x}))$ for some linear homogenous function G and strictly monotonic function h. We can allow for *variable proportions*, i.e., substitution among inputs. Our consideration of supply chain contracts in subsequent sections will often involve *fixed proportions*.[2]

Given input prices $\boldsymbol{w} = (w_0, w_1, \ldots, w_n)$, the cost function for a firm is defined as $C(q) = \min[\boldsymbol{w} \cdot \boldsymbol{x} | F(\boldsymbol{x}) = q]$. The average cost curve is defined as $c(q) = C(q)/q$. The demand side of the market is represented by a demand function, $Q(p)$, and we let $P(q)$ represent the inverse of this demand curve.

Technology is often assumed to exhibit constant returns to scale.[3] Constant returns to scale in the production function leads to flat average cost curves. Or the production function may be assumed to yield U-shaped average cost curves (which indicate returns to scale that are initially increasing, then decreasing). Economists often generate U-shaped average cost curves with an assumption that some factor, such as CEO managerial capacity, is inherently fixed. This captures the fact that firms cannot in reality simply duplicate their production processes without limit. To generate U-shaped cost curves we shall take the first factor, 0, as the fixed factor, giving it a value of 1 for each firm. The firm technology is therefore described by $q = F(1, x_1, x_2, \ldots)$ where the function F exhibits constant returns to scale. The U-shaped average cost curve is due entirely to the fixed amount of the first factor.

[1] The production function is entirely a description of the physical process available for production; markets or any considerations of the organization of production do not enter.

[2] A production function with fixed proportions is characterized by a vector $(a_1, \ldots, a_n)$, with

$$F(x_0, x_1, \ldots, x_n) = \min(a_1 x_1, \ldots, a_n x_n)$$

For example, the assembly of an automobile from parts requires 4 tires, 1 steering wheel and so on. Note that in the case fixed proportions, inputs are perfect complements.

[3] A production function $F(x_0, x_1, \ldots, x_n)$ involves constant returns to scale if, for any $\lambda > 0$, $F(\lambda x_0, \lambda x_1, \ldots, \lambda x_n) = \lambda F(x_0, x_1, \ldots, x_n)$.

In a competitive market, the profit maximization problem faced by any firm is[4]

$$\max_{x_1,\ldots,x_n} \pi = pF(1, x_1, x_2, \ldots, x_n) - \sum_{i=1}^{n} w_i x_i \tag{4.1}$$

A competitive equilibrium in the downstream market is a price equal to the minimum average cost, $p^* = \min_q c(q)$; a quantity for each firm $q^* = \arg\min c(q)$; a total quantity $q^* = Q(p^*)$ and a number of firms $m^* = Q^*/q^*$. As is common, we ignore the constraint that the number of firms in the market must be an integer. In this equilibrium, the downstream firms all operate at minimum average cost and source each input from an arbitrary set input suppliers. Firms earn zero *economic* profit, but since costs include a normal rate of return to capital, the *accounting profits* of firms are positive. The notion of a "supply chain" does not apply in a competitive setting, since each downstream firm simply purchases inputs from an input supplier without the need for a "relationship" of any kind with the supplier.[5]

We can easily demonstrate the efficiency of the price system in coordinating vertical incentives. Consider a single upstream firm and a single downstream firm. The two firms may operate separately in the competitive markets or as an integrated firm. Does vertical integration — that is, complete contracting at zero transactions and enforcement costs — yield any gain in profits relative to price-mediated exchange? If not, then the price system achieves first-best efficiency because even complete contracting would not improve profits.

Let the upstream firm be the supplier of input 1 and, for simplicity, assume that the upstream firm (and its competitors) produce at constant unit cost c_1. Suppose that the upstream firm integrates with a downstream firm. The integrated firm solves the following problem

$$\max_{x_1,\ldots,x_n} \pi = pF(1, x_1, x_2, \ldots, x_n) - c_1 x_1 - \sum_{i=2}^{n} w_i x_i \tag{4.2}$$

[4] We can ignore the cost w_0 in the profit maximization.

[5] We will sometimes use the result that a competitive market behaves as a single, constant-returns-to-scale, price-taking firm with all $n + 1$ inputs.

Comparing this equation with (4.1), note that for input 1, the upstream cost, c_1, rather than the market price w_1, enters the expression for profit. But since in the competitive equilibrium without integration $w_1 = c_1$, this problem is identical to the maximization problem solved by a downstream firm prior to integration. The choices taken by the downstream firm are, trivially, identical to the integrated firm. Incentives are unchanged by integration. There is no incentive for vertical integration. This result is both simple and deep. The price system alone coordinates decisions perfectly along a supply chain under ideal conditions.[6]

Uncertainty and competitive markets: Since so much of supply chain economics and design involves decisions under uncertainty, especially inventory decisions, it is important to extend the perfect-competition benchmark to incorporate uncertainty. With the introduction of uncertainty, the perfect-competition benchmark in economic theory becomes even more abstract or distant from the practice of supply chain management than the framework outlined above. But it remains important as a benchmark, so that we can understand incentives and contracts as consequences of the gap between the real world and the theoretical ideal.

Uncertainty is introduced into perfect competition, preserving the fundamental theorem of the efficiency of competitive markets, via the assumption of a complete set of contingent markets [52, Chapter 7]. Imagine — and it does take some imagination — that for every possible future contingency or “state of the world” there is an ex ante market today for delivery of each product contingent upon the state in the future. For example, suppose that one can trade today in a frictionless market for the delivery of lithium, at 9:43 am on November 14, 2028, contingent upon the weather, state of innovation, population growth, the election results in Kazakhstan, and any other exogenous variable.

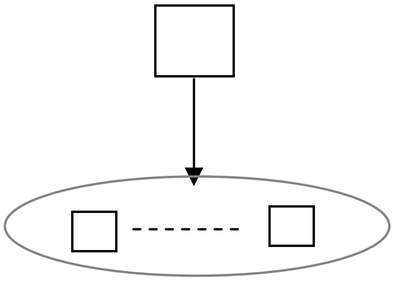
[6] The proof that the price system leads to efficient decisions under ideal conditions of perfect competition is extremely simple, in the supply chain context, as we have shown. The simplicity follows from the *partial equilibrium* nature of the model: prices outside the markets of focus are taken as fixed. The Arrow–Debreu theorem on the efficiency of competitive prices is much deeper in involving *general equilibrium*, the determination of all prices in a market economy.

With the introduction of these contingent markets, the economy with uncertainty is "reduced to the previous case" of certainty simply by extending the concept of a product to include delivery contingent upon a particular state of the world. With a complete set of contingent markets, and retaining assumptions of zero transactions costs and symmetric information, and various other assumptions such as convexity, a perfectly competitive economy continues to achieve the ideal allocation of resources [52]. As in the case of certainty, a version of this principle holds for a supply chain: the price system alone elicits efficient quantity decisions for all agents along the supply chain.

The perfectly competitive, complete-contingent-markets economy is, to emphasize once again, useful not as a realistic description of real world conditions but as a benchmark from which to analyze the impact of departures from this ideal. In the following section, we start with the simplest departure from the perfectly competitive economy.

5

Upstream Market Power



The perfectly competitive market setting analyzed in Section 4 includes the assumption that many firms offer identical products at each vertical stage. In this setting, if any single firm raised its price by even one penny, the firm would lose all demand. Individual firm demand, in other words, is perfectly elastic.

The simplest departure from this benchmark, to begin our exploration of supply chain contracts, is to consider a firm facing demand that is downward sloping. In other words, the firm has market power. We take the simplest assumption of an upstream monopolist for one of the inputs, input 1. The downstream market, and the markets for inputs, $i = 2, \ldots, n$ remain competitive.[1]

The incentive distortion introduced by market power is immediate. The input monopolist, with downward-sloping demand, sets price $w_1 > c_1$ to maximize profits. Under the variable proportions assumption, downstream firms then substitute away from the monopolized input towards other inputs. These firms face the monopoly price for

[1] We set aside consideration of market power in the downstream market, i.e., monopsony power. Monopsony power involves a distortion analogous to the variable proportions distortion.

the input, as their opportunity cost for purchasing additional units, rather than being guided by the "correct" price which is the upstream firm's marginal cost. The input distortion, known as the *variable proportions distortion*, is perhaps the most basic of all incentive distortions in supply chains.

We elaborate in this section on the variable proportions distortion and explore potential contractual resolutions to the distortion in a framework of certainty. We then introduce uncertainty in market demand into the framework and review the consequences of market power for decentralization of optimal inventory decisions.

5.1 Certainty: The Variable Proportions Distortion

5.1.1 The Distortion

Consider, as in Section 4, a good produced by firms in a competitive market with demand function $Q(p)$. All have the production function $F(1,x_1,\ldots,x_n)$. The first input into production, x_0, is fixed at 1 unit for each firm.[2] The second input, x_1, is monopolized. All other inputs $i = 2,\ldots,n$ are competitively supplied at prices equal to their unit cost, $w_i = c_i$.

The standard formulation of the upstream monopolist's problem starts with the definition of *derived demand*. The derived demand for input x_1 at a price w_1 is the amount of input x_1 that would be purchased by the downstream competitive market facing input prices $(w_1,c_2,\ldots,c_n)$ for inputs $1,\ldots,n$. For a formulation of the input demand in terms of the technology F and the demand function $Q(p)$, we refer to [195]. It suffices here to represent the derived demand function as $\tilde{q}(w_1)$. The monopolist's optimal input price, w_1^*, satisfies the well-known Lerner condition

$$\frac{w_1^* - c_1}{w_1^*} = \frac{1}{\eta} \tag{5.1}$$

[2] The variable proportions distortion can be developed in a model with constant returns to scale at the firm level, but our analysis below of contractual resolutions to the problem benefits from the assumption of U-shaped average cost curves.

where η is the elasticity (in absolute value) of the monopolist's derived demand, $\tilde{q}(w_1)$.

We analyze the variable proportions distortion by comparing the decisions of a decentralized input monopolist, in which only w_1 is set by the monopolist, with the decisions of a centralized or vertically integrated firm.

One could formulate the decentralized firm's problem as $\max_{w_1}(w_1 - c_1)\tilde{q}_1(w_1)$. Instead, we formulate the decentralized firm's problem in a way that facilitates comparison with the efficient solution. The upstream firm maximizes *total* system profits: since downstream firms earn zero economic profits, this is equivalent to maximizing its own profits. The decentralized firm's problem is represented as the choice of w_1, the downstream price p, and the entire vector of inputs $\boldsymbol{x}$ for each competitive downstream firm — but is subject to incentive-compatibility-type constraints that $\boldsymbol{x}$ and p be chosen to be consistent with the competitive equilibrium downstream that results from w_1. This formulation parallels the contract-theoretic approach that we adopt throughout this monograph.

$$\max_{w_1,\boldsymbol{x},p} pQ(p) - c_1x_1 - \sum_{i=2}^{n} w_ix_i \tag{5.2}$$

subject to

$$\begin{aligned}(x_1,\ldots,x_n) &= \arg\max_{(y_1,\ldots,y_n)} pq(y_1,\ldots,y_n) - w_1x_1 - \sum_{i=2}^{n} w_iy_i\\ pq(1,x_1,\ldots,x_n) &= w_1q_1 + \left(\sum_{i=2}^{n} w_ix_i\right)\end{aligned}$$

Note that in this maximization problem we effectively *pretend* that the firm has control over p and $\boldsymbol{x}$ as well as w_1, but constrain the firm to select only the $(p,\boldsymbol{x})$ compatible with w_1.

To understand the distortion introduced by market power, we compare the solution to this problem with the solution to the centralized-firm (first-best) problem. *The centralized firm solves the problem (5.2) without the constraints.* In identifying distortions in the price system, in other words, we are asking whether the incentive compatibility constraints are binding.

The failure of the price system to achieve coordination is immediate from a comparison of the first-order conditions in the constrained versus the unconstrained problem. The efficient use of x_1, via the unconstrained problem, equates the *marginal rate of technical substitution* between factor 1 and each of the other factors, j, to the ratio of opportunity costs

$$\frac{\partial q/\partial x_1}{\partial q/x_j} = \frac{c_1}{c_j} \tag{5.3}$$

The first-order conditions of the constrained problem (5.2) however imply

$$\frac{\partial q/\partial x_1}{\partial q/x_j} = \frac{w_1}{w_j} = \frac{w_1}{c_j} \tag{5.4}$$

We have our first failure of the price system. As a consequence of market power and the ability of downstream firms to substitute away from a monopolized input, distortions arise in the choice of input mix. The downstream firms choose the "wrong" mix of inputs because they see w_1 as an opportunity cost rather than being guided by the true opportunity cost, c_1.

The source of the incentive distortion is a *vertical externality*: each downstream firm ignores the gains to the upstream monopolist that flow through the wholesale margin, $(w_1 - c_1)$, with the purchase of each additional unit of input x_1.

A simple example of the variable proportions distortion is the case of the upstream supplier of material into a production process. Downstream firms substitute inefficiently into other inputs as a result of the upstream firm's monopoly pricing. For example, if the aluminium industry is monopolized, car manufacturers can substitute away from aluminium to other materials.

The variable proportions distortion, we would argue, also applies (broadly interpreted) to manufacturers supplying retail markets. Consider a monopolist producer of high-end paint supplying the market for hardware stores. Hardware stores can be considered as competing in the market for a single "generalized" product, the basket of items required to maintain a home. A hardware store aims to offer the best generalized product at the lowest cost, and will substitute away from a high-margin input (the high-end brand of paint) towards less expensive inputs or brands. This inefficient substitution is a source of lost profits for the entire supply chain. The variable proportions distortion problem is therefore a source of potential benefits from the use of more complex contracts. The depiction of retail markets as perfectly competitive is, of course, unrealistic. But the aim of the theory is to focus on the variable proportions distortion by adopting the single departure from the ideal economy that gives rise to the distortion.

5.1.2 Contractual Responses to the Variable Proportions Distortion

Vertical integration: The variable proportions distortion has been discussed in the economics literature mainly as an incentive to vertically integrate [199]. In theory, perhaps, a paint supplier could purchase all hardware stores. In practice, vertical integration is often not a feasible solution, as this example illustrates. What kinds of contracts, short of vertical integration, can resolve the problem? We outline several contractual responses to the input-mix distortion that have been discussed in the economics literature, and suggest additional contracts that potentially resolve the distortion.

Bundling and requirements tying: A well-known article by Burstein [31] shows that **tying** (or tied-sales contracts) can resolve the variable proportions distortion. *Tying* comes in two flavors: *bundling* is a strategy that requires a buyer to purchase two or more inputs in fixed proportions. For example, the buyer may be required to purchase two units of input 1 for each unit of input 2. A bundling contract simply dictates the mix of inputs that a downstream firm must purchase. Consider, for example, a supply chain in which the only substitute for x_1 is the second input, x_2. By bundling the two inputs in the efficient proportion, the upstream supplier resolves the variable proportions distortion directly.

A *requirements tying* contract differs from bundling. A requirements contract constrains the buyer of the monopolized input 1 to purchase *all* of the buyer's requirements of input 2 from the monopolist as well.[3] A requirements contract allows the monopolist to raise the price of input 2, while lowering the price of input 1, and thus to collect monopoly rents over two inputs instead of 1. This allows the monopolist to mitigate the distortionary substitution away from input 1 into input 2. If inputs 1 and 2 are uniquely substitutable, with no other input substituting for either, then requirements contracts can eliminate completely the distortion.

To make this argument formally, assume that there are only two variable inputs x_1 and x_2. The production function is $F(1, x_1, x_{2,})$. An example is the supply of plastic, high quality (x_1) and low quality (x_2), to the production of a machine part. A requirements tying contract in this case allows the manufacturer to set both prices w_1 and w_2 instead of w_1 alone. The two prices can be set in the efficient ratio:

$$\frac{w_1}{w_2} = \frac{c_1}{c_2} \tag{5.5}$$

The input mix chosen by the downstream purchasers (auto parts producers), which is governed by (5.4), is efficient under this requirements tying restriction because under (5.5), (5.4) and (5.3) are equivalent. The marginal rate of technical substitution, set by input demanders to equal the input price ratio, is non-distorted. The requirements restraint is necessary for the implementation of efficient pricing because if the supplier sets the price $w_2 > c_2$ without the requirements restraint, buyers would simply purchase from a competitive firm instead.

Output royalty: A simple contract that is equivalent to tying *all* inputs used in a downstream competitive market is an output royalty. We usually think of a firm as collecting revenue in proportion to the amount of an input that it sells — i.e., charging a price for input that it provides — but there is no reason that payment streams in contracts should be so restricted. *Any* action of the buyer could be "taxed" in this sense. Setting an output royalty, i.e., a tax on total *output* of the downstream

[3] Alternatively, the monopolist may require that the downstream firm purchase alternative inputs from a set of input suppliers who are being compensated by the monopolist for inclusion in the contract.

firms, in addition to the efficient marginal cost pricing of the input provided, is a strategy that eliminates the variable proportions distortion. When Microsoft sold the Windows operating system to O.E.M.'s in the 1990s, it charged a price based on all personal computers sold rather than on the subset incorporating its operating system, a strategy referred to as "per-processor pricing," it was adopting the strategy of an output royalty — arguably in response to a variable proportions distortion.[4] The marginal cost of its input was zero, and charging a markup on its input into the final product, sales of personal computers with software included, necessarily involved substitution away from the input relative to an efficient mix.

Market-share discounts: In addition to bundling, tying and output royalties as efficient responses to the variable proportions distortion, a fourth contract that can resolve the distortion is a market-share requirement or discount. A market share requirement in a contract with an upstream input supplier is a restraint that the downstream firm use the supplier's input for a minimum share of its requirements of some class of inputs. Pricing based on market share can, in principle, eliminate the variable proportions distortion just as bundling does.

Market-share pricing is not uncommon. Examples of this practice, are found in antitrust cases, *AMD v. Intel* (2005); *Masimo v. Tyco Health Care* (2004), *Concord Boat v. Brunswick* (2000) and *U.S. v. Microsoft* (1998).[5] As suggested by our reference to antitrust cases, the variable proportions distortion is only one motivation for market-share contracts. Other motivations for market-share based contracts (as for requirements tying contracts) are more contentious.[6]

Two-part pricing: Given that the variable proportions distortion operates through a positive wholesale margin, $(w_1 - c_1)$, it is natural to ask if the monopolist can avoid the distortion by simply collecting

[4] In the 1995 Consent Decree with the U.S. Government, Microsoft agreed to end per-processor pricing, but was unconstrained in offering quantity discounts.

[5] For a discussion of these cases and an excellent analysis of market share based contracts, see ref. [151].

[6] The other potential uses of market share contracts, including at the extreme exclusivity contracts, is the exclusion of other firms from a market. We discuss the use of supply chain contracts for exclusionary purposes in Section 8.

revenue through a fixed fee alone, leaving an efficient variable price $w_1 = c_1$. Eliminating the wholesale margin would remove the vertical externality.

In the competitive setting with U-shaped average cost curves, setting a fixed fee f for each downstream firm changes the total cost for each downstream firm from $C(q)$ to $C(q) + f$. The average cost function changes from $c(q)$ to $c(q) + f/q$, which reaches a minimum point at a *higher output level.* (See Figure 5.1.) Since competitive firms produce at the output where average cost is minimized, the fixed fee introduces another distortion: too few downstream firms operate in equilibrium. In other words, there is too little reliance by the aggregate competitive firm (or industry) on the first input, x_0.[7] The optimal two-part price balances the two distortions but does not in general yield a first-best optimum.[8] Ordover and Panzar [148] offer a full analysis of two-part pricing by a monopolist facing a downstream competitive market.

In subsequent sections of this monograph, we will find that a two-part pricing strategy is often first-best because it allows a "residual claimancy" contract. In the context here, however, two-part pricing does not resolve completely the variable proportions distortion. The upstream firm benefits from two-part pricing, but not to the extent of eliminating the distortion.

Resale price maintenance and two-part pricing: Two-part pricing alone is beneficial, but as we have seen it introduces another distortion in inducing excessive output levels by each competitive firm downstream. We suggest that this new distortion can be corrected with resale price maintenance. Consider a downstream competitive market operating under a binding price floor constraint. The price floor exceeds minimum average cost. Additional downstream firms will enter the market until the point where the price floor equals average cost, as in Figure 5.1. (Given the excess supply in the market, each firm would like to supply more that it is allocated in this equilibrium; we are assuming that the

[7] Recall that x_0 is set at 1 for each firm in the market, and note that the competitive market can be modeled as a single, aggregate, price-taking firm.

[8] The optimal two-part price is equivalent to optimal multi-product *Ramsey pricing* of the inputs 0 and 1 to the aggregate competitive firm. (For a discussion of *Ramsey pricing*, see ref. [134].)

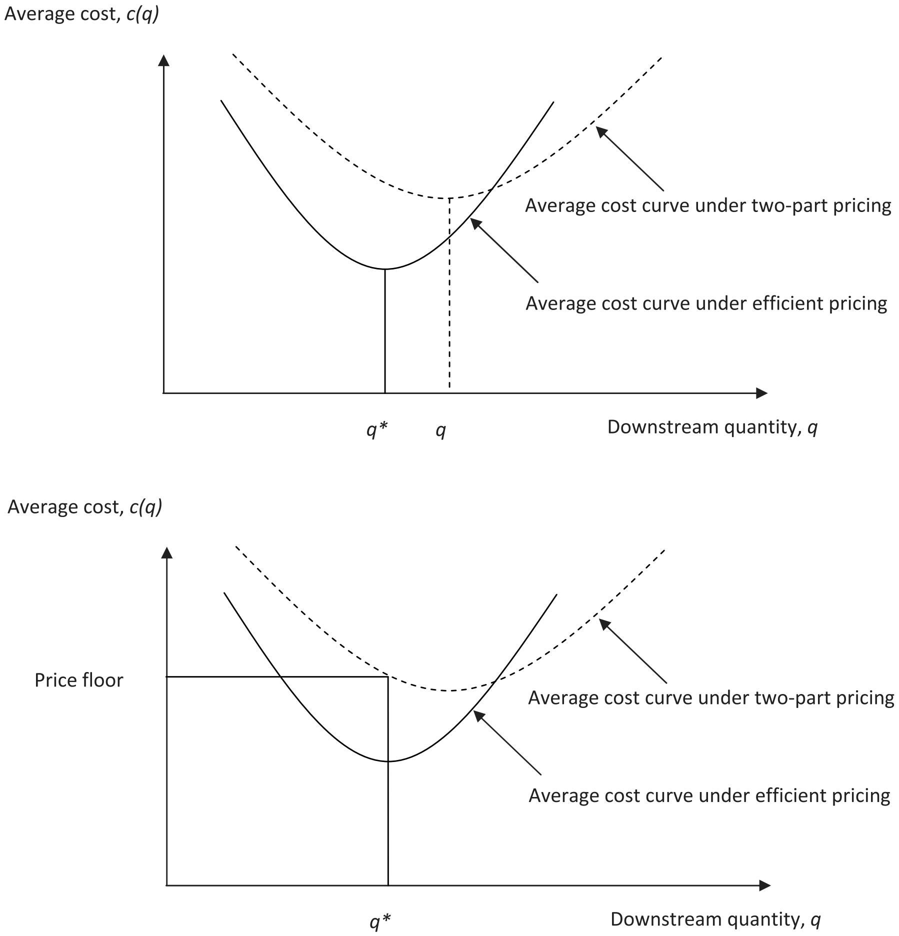


Fig. 5.1 Resale price maintenance and two-part pricing.

input is equally distributed among firms.) That is, free entry guarantees that each retailer produces at the point where the average cost, $c(q) + f/q$, equals the maintained price. Setting the price floor at the appropriate level elicits the efficient output on the part of each downstream firm — this is the output level where the average cost curve $c(q)$ (as opposed to $c(q) + f/q$) is minimized.

Note that at this price, each firm would prefer to sell additional quantity since its marginal cost lies below average cost and therefore below the price floor. The firm would like to undercut the price set by its rivals, but is unable to do so because of the price floor. Setting the minimum price appropriately elicits the efficient output level for each firm downstream. Thus, two-part pricing and resale price maintenance together allow first-best profits.

Recall that the variable proportions distortion can be interpreted as applying to a monopolist selling into a competitive retail market. The resolution of resale price maintenance and two-part pricing therefore applies potentially to the use of resale price maintenance in retail markets. Interpreted in a retail market context, this theory is consistent with the common argument that price floors are profitable as a means of attracting more outlets to the market — the "outlets hypothesis" [72].

5.2 Uncertainty: Inventory Incentives

Along with prices, inventory levels — i.e., quantities — are the most fundamental decisions that must be coordinated along a supply chain. Maintaining the single departure from the perfect competition ideal, upstream market power, we address the vertical coordination of inventory incentives under demand uncertainty.

Anecdotal evidence suggests strongly that incentives to maintain efficient inventory levels cannot be elicited without complex contracts. When Corning Glass Works was constrained to simple contracts following a case brought by the Federal Trade Commission in 1975 which ruled out resale price maintenance, one of the most detrimental effects to the firm was the cutback in inventory by downstream distributors and the loss of its smaller outlets [91, p. 325]. Hourihan and Markham [87] examined a set of cases also involving the prohibition of complex contracts. These authors found as well that with the restriction of firms to simple contracts, in three of the four cases where the restriction was binding, the availability of the product to buyers dropped.[9] This drop in availability the authors attributed to a decline in inventory levels and in downstream decisions to carry the product at all.

[9] The reader may wonder why the law sometimes restricts the set of contracts available to firms. The answer is that the law is sometimes wrong. In Section 2.4, we discussed the development of the law on resale price maintenance.

In these case studies, the contract that had supported superior inventory levels was resale price maintenance. The task for economic and management theory in illuminating these cases is to explain the role of resale price maintenance in generating adequate inventory incentives. The key papers on this topic, in a competitive downstream setting, are [54, 55]. We present here simple models that capture the ideas of Deneckere et al. [55].

The starting point to any inventory incentive theory is the foundational newsvendor model. A firm, facing a random demand, must choose inventory before the realization of demand. All unsold inventory perishes. The conventional assumption is that price is exogenous. This corresponds to an implicit assumption that all consumers value the product at some common value, v. The demand facing the firm is then perfectly inelastic up to a price equal to v, i.e., demand is rectangular; only the size of this demand is uncertain. Taking the usual notation, we let the random demand be x, with a distribution F and continuous density f. The firm produces at a cost $c < v$. Thus we have a monopolist with uncertain, perfectly inelastic demand, whose product has a perfect substitute available at price v. The monopolist solves

$$\max_y \pi(y) = vE\min\{y,x\} - cy = vy - vE\max\{y - x, 0\} - cy \quad (5.6)$$

which yields an optimal condition

$$[1 - F(y)]v = c \quad (5.7)$$

The marginal benefit of an additional unit of inventory, $[1 - F(y)]v$, is equated to marginal cost, c, at the optimum. This yields the classic fractile solution:

$$1 - F(y^*) = c/v$$

at the optimal inventory, y^*.

Ex post pricing: Suppose, however, that the monopolist must distribute its product to final buyers through the intermediation of a competitive retail sector. The monopolist simply sets a wholesale price w and sells to retailers, which bear no cost other than w. The retailers must choose their purchases of inventory prior to the realization of random demand. After the realization of demand, price is determined in an ex post market. The ex post market consists of the realized demand curve (rectangular out to the realized value of x) and a supply curve that is perfectly inelastic at the amount y of aggregate inventory chosen by firms in the sector. For any realization of demand up to y, inelastic supply exceeds inelastic demand, and the market-clearing price in this ex post market is 0. For realizations of demand greater than y, the market-clearing price is v, the maximum amount that buyers will pay. That is, given y and realized x, the market price

$$p(x,y) = \begin{cases} 0\colon & x \leq y, \\ v\colon & x > y \end{cases} \tag{5.8}$$

Industry profits are given by:

$$[1 - F(y)]vy - wy = vy - vyF(y) - wy \tag{5.9}$$

Compare the classic newsvendor profit function (5.6) with the industry profits under competitive intermediation (5.9). The inventory problem under competitive intermediation is equivalent to an optimal inventory problem *in which the monopolist faces the constraint that if its inventory turns out to be excessive (even by one unit), the entire revenue from the inventory must be abandoned*. For the supply chain as a whole, retail competition destroys all profits if inventory is excessive even by one unit.

The downstream competitive market will order inventory, y, up to the level at which each of the price-taking firms (firms take the price, conditional upon any realization of x, as given) foresees zero expected profits. The aggregate zero-profit level is where the expected revenue from each unit equals its cost, i.e., $[1 - F(y)]v - w = 0$.[10]

Incorporating this competitive market reaction as a constraint, we write the monopolist's problem as follows

$$\max_{w,y}(w - c)y \tag{5.10}$$

subject to

$$[1 - F(y)]v - w = 0$$

[10] This equation expresses the zero expected profit condition, although it is identical to the optimal inventory condition for a downstream firm facing an upstream price w.

Substituting the constraint into the objective function, the monopolist's problem can be written as $\max_y([1 - F(y)]v - c)y$. The first-order conditions to this problem yields the following characterization of the optimal inventory level, y^*:

$$[1 - F(y^*)]v = c + f(y^*)vy^* \tag{5.11}$$

Compare (5.11) with (5.7). Competitive intermediation *adds a term* to the classic fractile characterization of optimal inventory — the additional term $f(y)vy$ on the right-hand side of (5.11) as compared to (5.7). When an additional unit of inventory is added, the probability of losing the entire revenue vy is increased by the amount $f(y)$, and this additional marginal cost is reflected in the first-order condition. This means that the optimal inventory is lower as a result of the intermediation by competitive firms. Profits are reduced.

Setting aside vertical integration as a solution, we turn to the simplest contracts that resolve the incentive distortion. The natural strategy to consider, given case evidence, is (as Deneckere et al. [55] suggest) the adoption of resale price maintenance (RPM).

Consider the imposition of a price floor at an arbitrary price $\underline{p}$. Holding $\underline{p}$ constant leads to a one-to-one relationship between the instrument w and the outcome y, as firms are induced to raise inventory to the point where the aggregate profits in the competitive retail sector are zero. This relationship is given by the following:

$$\underline{p}y - \underline{p}E\max\{y - x, 0\} - wy = 0$$

Under RPM, the price does not fall to zero as soon as x is slightly below y. The monopolist's problem becomes the following[11]:

$$\max_{w,y,\underline{p}}(w - c)y \tag{5.12}$$

subject to

$$\underline{p}y - \underline{p}E\max\{y - x, 0\} - wy = 0$$

Solving the constraint in this problem for w and substituting it into the objective function (5.12) *replicates the original newsvendor maximization problem for the vertically integrated firm.* The monopolist can set

[11] This formulation assumes that the price floor is binding; this is easily demonstrated.

$\underline{p} = v$, and set w so that the downstream retailers choose the first-best inventory level. (Note that the w that elicits first-best inventories will be greater than marginal cost c.) Adopting RPM, therefore, allows the monopolist to achieve the first-best system profits. This is the simplest theory of the observed role of resale price maintenance in supporting inventory levels.

Ex ante pricing: In a contemporaneous paper, Deneckere et al. [54] develop a similar model with demand uncertainty and inventory choices, but in which retailers must commit to prices *before* the realization of demand. In the simple-contract game (with no vertical restraints), the manufacturer sets a wholesale price, w, and the retailers then each choose price and inventory levels simultaneously. In many markets, including video markets and fashion good markets that Deneckere et al. [54] describe, the resolution of uncertainty is short, and retailers must choose prices ahead of the realization of demand. The pricing cycle is as long or longer than the production/inventory cycle.

We can capture the basic idea of Deneckere et al. [54] with a model in which — as before — all consumers share the same value v for a unit of demand, but in which the number of consumers is random, and represented by a smooth distribution, $F(x)$.[12] But now prices are set ex ante. We assume, for simplicity, that each retailer provides one unit or none. The "inventory" decision is thus whether to carry the product or not.[13] Consumers are assumed to enter the market sequentially, rather than simultaneously.

The result is that the retail market is characterized by the following equilibrium, also studied by Dana [49]. With a continuum of buyers and a continuum of retailers, the result is an equilibrium with *price dispersion*: sellers in equilibrium set a variety of prices and buyers enter the market compare prices perfectly, each buying (upon entry) from the lowest-priced supplier still available. Retailers set prices along a continuum, trading off a higher probability of sale (at a low price) with

[12] Deneckere et al. [54] present a numerical example with unitary demand and common value, as we are doing here, but assume three possible states of demand.

[13] This assumption is inessential.

a higher realized profit conditional upon sale (at a higher price). Along the continuum, retailers each earn zero expected profits.

Consider the retail market facing a given wholesale price, w, and no other costs. For a given distribution $G(p)$ of retailers charging various prices up to p,[14] a retailer charging a particular price $\widehat{p}$ will sell if $x \geq G(\widehat{p})$, an event that happens with probability $1 - F(G(\widehat{p}))$. The retailer's expected profit from selling one unit is therefore $[1 - F(G(\widehat{p}))]\widehat{p} - w$. An equilibrium price distribution is a distribution that results in zero retailer profits at each price, and thus is the solution G^* to the following functional equation:

$$[1 - F(G^*(p))]p - w = 0 \tag{5.13}$$

It is straightforward to show that G cannot have atoms, and that the support of G must be $[w,v]$. Note that $G^*(v)$ is the total number of retailers.

At a given w the total inventory ordered by retailers, i.e., the input demand function facing the upstream monopolist, is given by $G^*(v)$, the total number of retailers. The profit of the monopolist is given by $(w - c)G^*(v)$, where G^* is the solution to the retail equilibrium equation.

Consider, instead of the simple contract, a contract in which the upstream monopolist imposes a price floor at v. In this case, the inventory purchased from the monopolist is the amount that yields zero profits downstream, as earlier. That is, y is the solution to the following:

$$vy - vE\max\{y - x,0\} - wy = 0 \tag{5.14}$$

Note that instead of a distribution of downstream prices on $[w,v]$, all the downstream firms set price v. In terms of economics, if the $G^*(v)$ retailers under the simple contract suddenly face a price floor at v, the quantity demanded under each realization of uncertainty is unchanged but price is higher. The higher price elicits more entry ex ante (via either additional retailers or additional inventory at each retailer). Resale price maintenance thus increases the demand for the upstream firm's product. As before, the manufacturer can set w to elicit first-best inventory and profit levels.

[14] We omit the dependence of G on w to keep the notation simpler.

The Deneckere–Marvel–Peck models, the essence of which is captured in our two simple settings, are in our view fundamental contributions to the foundations of supply chain organization. The models capture the failure of the price system to convey the optimal incentives for inventory with the introduction of the smallest departure from perfect competition, market power upstream. And this smallest departure is enough to explain the role of resale price maintenance in correcting the incentive distortions. Marvel and Wang [132] add a buy-back policy to resale price maintenance and show in the model of inventory with pricing flexibility (the second model analyzed above) that addition of this second instrument yields first-best profits. The manufacturer need not even know the distribution of demand uncertainty in implementing the optimal policy, but can instead rely on retailers' inventory incentives. Marvel and Wang [133] extend the analysis to the additional instrument of revenue sharing, and show that any two of the three instruments — a constant wholesale price, buy-backs and revenue sharing — can achieve first-best coordination.

5.3 An Aside: Perfectly Competitive versus Monopoly Choices of Inventory in One Market

The focus of this monograph throughout is on the vertical control of decisions such as the choice of inventory under conditions of market power. The setting that we adopt in various forms has two vertically linked markets. But is useful as an aside to consider a more basic question: what is the impact of market power on the optimal inventory choice in a single market (i.e., a single-stage setting rather than a vertical setting)?

Rectangular demand: We show first that when demand is rectangular, i.e., all consumers share the same value v for the good, a competitive market and a monopolized market yield the identical equilibrium inventory level. Just as in the textbook case of equilibrium under rectangular demand with certainty, market power has no impact on quantity.[15] Consider the standard newsvendor setting for the optimal

[15] A monopoly will set price at v and quantity equal to the number of consumers. A perfectly competitive market will set price equal to cost and sell to the same number of consumers.

inventory problem. The monopolist produces an inventory y at cost c per unit, prior to the realization of demand. Demand consists of a random number of consumers x with common value v for a product. The distribution of x is $F(x)$ with density $f(x)$. The monopolist will set price v. (This is irrespective of whether price is set ex ante or ex post; we take the assumption of ex post pricing.) The monopolist's expected profit is $\pi = vE\min(x,y) - cy$. The monopolist's optimal inventory satisfies $F(y^*) = (1 - c/v)$, the classic fractile solution. This can also be written $1 - F(y^*) = c/v$.

Consider next a competitive market, i.e., market with free entry, in which prices are flexible (i.e., there is market-clearing ex post). The amount of inventory set by a competitive market will be the level y_c that yields zero profits. Profit depends on the price that is set ex post. The demand curve ex post is rectangular up to x at a price equal to v; and the supply curve is perfectly inelastic at a quantity y_c. The equilibrium price is therefore v if $x \geq y_c$ and 0 if $x < y_c$. Expected industry profit can be expressed as $[1 - F(y_c)]vy_c - cy_c = 0$, or $1 - F(y_c) = c/v$. The competitive market and the monopoly market produce the same inventory when prices are flexible and demand is rectangular. The difference in the market outcomes between competition and monopoly with rectangular demand, as under certainty, is in which parties capture the market surplus. In the monopoly case, the single firm captures the entire surplus; and under perfect competition, consumers gains total surplus.

Downward sloping demand: When demand is *downward sloping* and prices continue to be set ex post, then perfectly competitive markets yield a higher quantity than monopoly, as in the textbook setting with known demand, under reasonably regularity conditions. To show this, let $q(p,\theta)$ be a demand curve with the standard property that elasticity, $\varepsilon(p,\theta) = d\ln q(p,\theta)/d\ln p$, is increasing in price (for every θ). Let $P(q,\theta)$ be the inverse demand. Define $\tilde{\varepsilon}(\hat{q},\theta)$ as the demand elasticity at the market clearing price (i.e., at the price where $\hat{q} = q(p,\theta)$) and assume that $\partial\tilde{\varepsilon}(\hat{q},\theta)/\partial\theta > 0$.[16] The monopolist's problem in this case can be

[16] This regularity condition holds for multiplicative demand uncertainty, $q(p,\theta) = \theta\tilde{q}(p)$ for some function $\tilde{q}(p)$.

written as the following

$$\max_{y,\tilde{y}(\theta)} E_\theta[P(\tilde{y}(\theta)) - c]\tilde{y}(\theta)$$

subject to

$$\tilde{y}(\theta) \leq y$$

This maximization problem allows for the fact that the monopolist will gain ex post by disposing of some output, rather than selling it, if θ is realized below the critical value $\hat{\theta}$ defined by $\tilde{\varepsilon}(y,\theta) = 1$. In the range $\theta < \hat{\theta}$, the elasticity of demand is less than 1 at the price at which y would be sold, and the monopolist gains more revenue by raising price to the point where $\varepsilon(p,\theta) = 1$ than by selling all units.[17] The monopolist's first-order condition for this problem, with respect to y, is

$$\int_0^{\hat{\theta}} -c dF(\theta) + \int_{\hat{\theta}}^{\infty} P(\tilde{y}(\theta),\theta) - c + \tilde{y}(\theta) \cdot \frac{\partial P[\tilde{y}(\theta),\theta]}{\partial \tilde{y}(\theta)} dF(\theta) \qquad (5.15)$$

At each θ in the first integrand, where the monopolist is disposing of some output, the realized change in profits from increasing y includes no additional revenues. In this region, the constraint in the maximization problem is not binding. The second integrand is the familiar derivative of monopoly profit with respect to quantity.

At the inventory level set in a competitive market equilibrium, y_c, expected profits are zero. Therefore, $E_\theta[P(y_c,\theta) - c] = 0$, which we re-write as

$$\int_0^{\hat{\theta}} [P(y_c,\theta) - c]\, dF(\theta) + \int_{\hat{\theta}}^{\infty} [P(y_c,\theta) - c]\, dF(\theta) = 0 \qquad (5.16)$$

Evaluate the monopolist's first-order condition (5.15) at $y = y_c$. The first integrand in (5.15) is less than the first integrand in (5.16) since $P(y_c,\theta) > 0$. That the second integrand in (5.15) is also less than the second integrand in (5.16) is easily shown using the fact that the elasticity of demand is less than 1 in this region, i.e., that $|p/y \cdot dp/dy| < 1$.

[17] A monopolist will never sell quantity in a region of the demand curve where elasticity is less than 1.

The monopolist, therefore, sets an optimal inventory level less than the competitive market, in this case of ex post pricing.

Note that in terms of the impact of market power on total surplus in this setting, the monopolist is inefficient not only in setting inventory too low ex ante, but also in disposing of valuable output ex post under some realizations of demand. The image that is brought to mind is the bulldozing of agricultural products in European countries once the government has arranged price at a collusive level. Thus, with ex post price flexibility, the undergraduate textbook result that monopoly involves lower output and lower welfare extends to the simple inventory model with uncertain demand, providing that prices are flexible, being set after the realization of demand.

Downward sloping demand and ex ante pricing: Monopoly may be efficient. When prices are set ex ante, the comparison is not so straightforward. With rigid prices in the face of demand uncertainty, and a fixed level of inventory as well, demand will be rationed ex post in the states where inventory falls short of quantity demanded at the fixed price. Two kinds of rationing mechanisms can be considered.[18] Suppose that individuals with the highest value for the product are those who are served (perhaps they walk faster on their way to the outlet) — the *efficient rationing mechanism* — then the conventional results of lower output and lower welfare under monopoly are preserved.

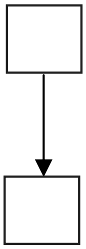
If instead rationing is *proportional*, with the probability of facing a stock-out being independent of willingness-to-pay across consumers, then the monopoly may yield higher surplus. Suppose, for example, that some individuals have extremely high willingness-to-pay for the product, but that a very large number are willing to pay only a penny above unit cost. Randomness is in the number of the latter type of consumers. A monopoly would price to serve only the high willingness-to-pay consumers, and all of these consumers would be served. A competitive market, on the other hand, would price at cost. This would unleash

[18] Tirole [190, p. 212]. See also ref. [125] for an analysis of various rationing schemes under a different market structure (duopoly).

the low-demand consumers to take up demand with high probability in rationing states, crowding out (randomly) some of the high-value consumers for whom successful transactions generates the most surplus. Total expected surplus falls with a move from monopoly to competition in this example.

6

Two-stage Market Power: The 1–1 Market Structure



We introduced market power at one stage of the supply chain in Section 5 for purposes of understanding the incentive and contract implications from the smallest departure from the world of perfect competition. But market power is ubiquitous. What are the contracting implications of market power at *both* stages of a two-stage supply chain? Recall that with perfect competition downstream, market power upstream creates incentive problems in a world of certainty only when technology exhibits variable proportions. This is no longer true with market power at two stages. Even with fixed proportions at two stages, incentive distortions arise.

We take the simplest setting with a monopoly at each of two levels. (See Figure 1.1(c).) The upstream monopolist provides the only input, for example a product that is resold downstream. This is a fixed proportions technology, which sets aside the variable proportions distortion covered in the previous section of this monograph.

6.1 Certainty

6.1.1 Price Decisions Only: The Double Markup Problem

Among the first contributions to the analysis of supply chain incentives was Spengler [181], who pointed out that vertical integration between two vertically related firms with market power will reduce the price. Spengler assumed certainty in demand, but restricted firms to the adoption of uniform-pricing contracts. The incentive issues with two-stage market power arise, in the simplest setting, where the two firms with market power are assumed to have no access to lump sum transfers. Under this assumption, the objective function of the upstream firm in setting its price will be its own profit.

Consider a two-stage supply chain with one firm operating at each stage. The final demand curve, $q(p)$, is downward sloping. To describe Spengler's double mark-up problem, we start with the benchmark of a vertically integrated (centralized) firm. In setting the price of the final product, this firm faces a straightforward maximization problem

$$\max_p \Pi(p) = (p - c)q(p) \tag{6.1}$$

We assume that the absolute value of the elasticity of demand, $\varepsilon(p) = |d\ln q/d\ln p|$, is nondecreasing in p. (This is a standard assumption and sufficient for the second-order conditions in (6.1)). Let $\varepsilon_I = \varepsilon(p^*)$ where p^* is the optimal price for the integrated firm. The first-order condition for the maximization problem (6.1) yields the familiar Lerner equation:

$$\frac{p^* - c}{p^*} = \frac{1}{\varepsilon_I} \tag{6.2}$$

The price p^* is the efficient (profit-maximizing) price for the supply chain.

In the two-stage, non-integrated case, the upstream firm faces a derived demand $\tilde{q}(w) = q(\hat{P}(w))$ where $\hat{P}(w)$ is the price that the downstream firm will set given w. Expressing $\hat{P}(w)$ formally: $\hat{P}(w) = \arg\max_{\widetilde{p}}(\widetilde{p} - w)q(\widetilde{p})$. The upstream firm sets w to solve the following problem

$$\max_w \pi_u(w) = (w - c)\tilde{q}(w) = (w - c)q(\hat{P}(w))$$

We denote $\tilde{\varepsilon}(w) = |d\ln\tilde{q}(w)/d\ln(w)|$ be the elasticity of the derived demand function. We denote also $\varepsilon_u = \tilde{\varepsilon}(\hat{w})$ where $\hat{w}$ is the optimal price for the upstream, non-integrated firm. The optimal price $\hat{w}$ satisfies the following Lerner equation

$$\frac{\hat{w} - c}{\hat{w}} = \frac{1}{\varepsilon_u} \tag{6.3}$$

Given $\hat{w}$, the downstream firm solves

$$\max_p \pi_d(p) = (p - \hat{w})q(p)$$

which yields, for the downstream firm, an optimal price $\hat{p}$, that satisfies

$$\frac{\hat{p} - \hat{w}}{\hat{p}} = \frac{1}{\varepsilon_d} \tag{6.4}$$

where $\varepsilon_d = \varepsilon(\hat{p})$. From (6.2) and (6.4) we have

$$p^* \frac{\varepsilon_I - 1}{\varepsilon_I} = c \quad \text{and} \quad \hat{p}\frac{\varepsilon_d - 1}{\varepsilon_d} = \hat{w} \tag{6.5}$$

Defining $G(p) = p \cdot [\varepsilon(p) - 1]/\varepsilon(p)$ (6.5) can be rewritten as $G(p^*) = c$ and $G(\hat{p}) = \hat{w}$. By (6.3) we know that $\hat{w} > c$, and we know that G is increasing (from the assumption that $\varepsilon(p)$ is nondecreasing). It follows that $\hat{p} > p^*$.

This captures the classic double markup problem arising from two-stage market power. The downstream firm marks up price above the wholesale price (which has already been marked up by the upstream firm) instead of the opportunity cost for the system, c.

Economists focussed initially on vertical integration as a solution to the double markup distortion. If two firms along the same supply chain have market power, then vertical integration of the firms will reduce prices, increasing not only the profits of firms involved, but consumer surplus as well.[1]

The source of the incentive distortion is a *vertical externality*: in setting the price, the downstream firm does not consider the negative impact on upstream profits of an increase in the price. When price is increased by 1 unit, the upstream profits drop by $(w - c)q'(p)$.

[1] The important relationship between firms for which a merger is presumptively price-reducing is that the firms produce *complements*. Vertically related firms are merely an example of this.

The externality, working through the wholesale margin $(w - c)$, drives a wedge between the downstream firm's objective function and the collective interest of both firms.

6.1.2 Resolution: Two-part Pricing and the Residual Claimancy Principle

Under the assumptions of this simple model, the incentive problem is easily resolved. A two-part price, with the variable price equal to marginal cost, eliminates the variable wholesale markup and therefore eliminates the incentive distortion. No more complicated contract is necessary, since this contract yields first-best profits. Under the assumption of certainty, any other nonlinear pricing scheme, such as quantity discounts [94], would work as well: any contractual outcome implemented with a nonlinear price can under certainty be implemented with a two-part price.

This two-part pricing solution to the potential double mark-up problem, is an example of the *residual claimancy principle*: if a single agent is taking a decision and a contract leaves the entire marginal profits from the decision with the agent, then the contract eliminates all externalities and therefore all incentive distortions. The agent's decision under a residual claimancy contract is first-best.

Three factors prevent the residual claimancy solution to contracting problems in practice. Consider the set of principal–agent problems. The first set of circumstances preventing the residual claimancy solution is the combination of uncertainty and limited wealth on the part of the agent. A residual claimancy solution is effectively the sale of the project or venture to the agent. The principal takes a lump sum and is no longer involved; all externalities are internalized. But if the return is uncertain and the agent's wealth is limited, the agent will in general be unable to promise an adequate lump sum payment to the principal. The limited wealth constraint will bind. The second set of circumstances precluding residual claimancy contracts is the combination of uncertainty and agent risk aversion. Even if the agent's wealth or limited liability constraint is not binding — so that a residual claimancy contract is feasible — the contract will not be desirable. The residual claimancy contract leaves all risk with the risk-averse agent, whereas sharing of the risk by the principal is part of a first-best solution. In this case, the second-best optimal solution balances the incentive benefits of allocating residual claim to the agent with the risk-allocation benefits of leaving some risk with the principal.[2] Finally, if there is more than one agent, the single residual claim must be split among agents (in the absence of a budget breaker [84].[3]

6.1.3 Extension to Multi-task Agency

It is natural to ask how incentives and contracts change when the downstream firm's decisions involve not just price but sales effort as well. ("Sales effort" is interpreted here and throughout the monograph as any action affecting demand.) Retailers do more than set prices. Retailers provide point-of-sale information, a comfortable shopping environment, adequate numbers of sales staff and cashiers, and so on.

The extension of the two-stage monopoly problem to the case where the downstream retailer takes more than one action — an example of the multi-task agency problem [86] — is straightforward. The residual claimancy principle is unaffected by the inclusion of multiple tasks on the part of the single agent. The residual claimancy contract internalizes the entire profit-maximization problem for the agent, and whether the agent has one task (pricing) or many, the first-best outcome is achieved. In other words, when a downstream retailer undertakes multiple tasks a two-part pricing scheme, with variable price equal to marginal cost, continues to elicit the first best.

6.1.4 Failure of Residual Claimancy: Bilateral Agency in the 1–1 Market Structure

The multi-task nature of decision-making along a supply chain characterizes both downstream distributors and upstream firms.

[2] "Second-best" refers to optimality under the constraints rather than the unconstrained first-best contract.

[3] These three sets of circumstances ruling out the simple residual claimancy contract give rise to three classes of principal–agent models: the risk-averse agent models [83, 176] the limited-liability principal-agent model [166] and the bilateral agency model [84].

Manufacturers invest in product quality, advertising and in product support activities such as honoring warranties; retailers undertake local promotion, sales effort, dimensions of point-of-service product quality, and so on. Not all of these actions can be guaranteed by contract but will instead be chosen once contracts are established. Hence the need to design a contract that will optimally elicit incentives not just downstream but also at the manufacturer's level. The 1–1 principal agent problem becomes a bilateral agency problem: the contract must be designed to account for the post-contractual incentives of both parties to the contract. The manufacturer is no longer a "principal" but becomes instead an agent as well — an agent being defined (as always) as any contractual party that takes a decision influenced by the contract.

Leaving either party with the full residual claim, and zero for the other agent, would not in general be optimal. The optimal contract leaves each party with positive, though incomplete, incentives to undertake effort. Lafontaine [112] concludes, for example, from an empirical analysis of franchise contracts that "results are broadly consistent with a two-sided hidden-action or moral-hazard explanation of franchising, suggesting that there are incentive issues on both sides." Bhattacharyya and Lafontaine [25] cite additional literature with evidence supporting the theory of bilateral moral hazard. In the context of franchising, for example, the payoff depend not only on the efforts of the franchisee but also on inputs of the franchisor. This conclusion surely extends from franchising to supply chains generally.

The supply chain coordination problem arises from the non-contractibility of the actions taken by the two agents. The question that naturally arises here is the nature of the second-best optimal contract given the bilateral agency constraint against residual claimancy. Bhattacharyya and Lafontaine [25] and Romano [164] analyze the bilateral-agency supply chain model. In these models, an upstream firm and a downstream firm both take actions to enhance the value of the product. One interpretation is that an upstream firm invests in product quality and a downstream firm invests in promotion or sales effort.

In simple bilateral agency models, with risk-neutral agents (or no uncertainty), two-part sharing rules implement any solution achievable with general nonlinear rules. That is, a rule consisting of a fixed fee and a constant share of output implements the second-best solution.[4]

In the Romano [164] model, the constant shares are implemented via a constant wholesale price. We offer here a simple version of his model. Romano shows that a vertical restraint on the *prices* set by the downstream firm — the one action of the firm that *is* contractible in his model — is second-best optimal. With the downstream price as the only contractible action, the contract consists of a wholesale pricing function $W(q) = wq + F$ and a retail price, p, imposed on the downstream firm. Let c be the constant production cost of the upstream firm, y denote the upstream expenditure on quality and x the downstream expenditure on promotion or sales effort. Final demand is a function $q(p, y, x)$. The upstream and downstream profits (disregarding any fixed fee transfer) are $\pi^u = (w - c)q - y$ and $\pi^d = (p - w)q - x$. These functions are assumed to satisfy standard regularity properties.[5]

In any contracting problem under certainty, in which lump sums are freely transferable between contracting parties (so that the contracting game is one of transferable utility), the objective function in designing the contract is the sum of profits. The bilateral supply chain contracting problem therefore becomes[6]

$$\max_{w,p,y,x,q} (p - c)q(p, y, x) - y - x$$

subject to the incentive compatibility constraints

$$\begin{aligned} y &= \arg\max_{\tilde{y}} (w - c)q(p, \tilde{y}, x) - \tilde{y} \\ x &= \arg\max_{\tilde{x}} (p - w)q(p, y, \tilde{x}) - \tilde{x} \end{aligned}$$

[4] The sufficiency of a two-part sharing rule is clear. Under any general nonlinear sharing rule, each agent at the optimum equates his expected marginal return to his marginal cost. Replacing the general rule with a two-part rule with constant shares equal to the same expected marginal returns leaves each agents' incentive unchanged. The two constant shares will add to 1.

[5] See (a1)–(a7) from Romano [164].

[6] Romano's description of this problem is more complicated in that (a) he includes only the upstream firm's profit, and (b) imposes a participation constraint downstream. With a lump sum transfer as part of the contract, we can reformulate the objective as maximizing combined profits.

Given that the price p is set in the contract, the wholesale price becomes purely an instrument to divide the residual claim, $(p - c)$ for each unit of demand, between the upstream and downstream parties, with $(w - c)$ going to the upstream firm and the remainder, $(p - w)$, going to the downstream firm. Intuitively, the greater share of residual claim will go to the party whose non-price actions are (a) most important to the overall enterprise — here $q(p, x, y)$ — and (b) most sensitive, via the incentive compatibility constraint, to the allocation of residual claim. Constraining the retail price to be higher than the level that would be voluntarily chosen by the retailer has the benefit of raising the total residual claim, $(p - c)$, available to entice greater non-price instruments — but at the cost of demand being discouraged by the higher price. At one extreme, if non-price variables are not very important, the optimum contract would be close to the two-part pricing contract, with $w = c$ and no price restraints, that allocates all residual claim to the downstream firm since it makes the one decision, on p, that matters in this case. Romano shows that the contract will constrain the retailer to setting a specified retail price if non-price variables matter.[7]

In reality, the impacts of the manufacturer's decisions on demand and the retailer's decisions on demand are different in terms of the timing. An upstream manufacturer invests in quality and brand name promotion to build reputation — a capital stock. If the manufacturer were to shirk on reputation by cutting back quality, it would lose an asset of substantial value: its reputation as a high-quality, prominent brand. This is an asset that is valuable not just for the current period or current transactions but well into the future. A downstream retailer shirking on effort in setting x to sell the current inventory often faces a less severe reputational effect. This suggests that actual contracts should be geared towards concern over downstream shirking and incentives more than upstream incentives. This is exactly what we see in franchise contracts [78, 99]. In Section 12, we explore more generally

[7] The exception is the (coincidental) case where $E \equiv \theta(-\varepsilon_p + \varepsilon_x\eta_x) + (1 - \theta)\varepsilon_y\eta_y = 0$, where θ is $(w - c)/(p - c)$, the manufacturer's share of residual claim, the ε variables are elasticities of demand with respect to p, x and y; η_x is the elasticity of the provision of x with respect to the equilibrium price; and simply for η_y.

the impact of reputational and related long-run factors on supply chain coordination.

6.2 Uncertainty

Within the market structure of two-stage monopoly of this section, we now introduce uncertainty. Uncertainty is represented in the simplest way by a demand function that depends not only on price, p, but also on a random variable, θ: $q(p, \theta)$. We divide our synthesis of contracts under uncertainty into three cases: in the first case, the only downstream decision is retail price; in the second case, inventory (quantity) is the only downstream decision and; in the third case, both price and inventory decisions are made downstream.

As discussed earlier, uncertainty by itself does not preclude a first-best contract. If the downstream firm is risk-neutral and has sufficient wealth, it can pay the upstream firm a lump sum equal to the expected profits from a first-best contract and, because it then holds the entire residual claim, will have internalized the incentive problem. First-best incentives would follow. The simple residual-claimancy resolution is ruled out, however, if either the agent has limited wealth, and therefore cannot commit to paying the upstream firm a sufficiently high fixed fee, or if the agent is risk averse. In analyzing the pricing decision, we outline the case of limited agent wealth, assuming that the agent's wealth is equal to 0. With zero wealth, the agent cannot purchase the entire supply chain to implement the residual-claimancy contract.

6.2.1 Price Decision Only

6.2.1.1 Optimal Two-part Pricing

The upstream firm offers a contract ex ante. Uncertainty is realized and then the downstream firm accepts or rejects the contract and, if it accepts, chooses retail price p. We assume as before that demand uncertainty takes a multiplicative form $\theta q(p)$, with θ having a continuous probability density with support $[\underline{\theta}, \overline{\theta}]$. With respect to the feasible contracts, we consider first a two-part pricing contract, $W(q) = F + wq$. We then briefly discuss the structure of the more general nonlinear pricing contract. We subsequently consider contracts that include restrictions on the retail price.

In the two-part pricing problem formulation, the contract offered ex ante by the upstream firm is (F,w). The multiplicative form of demand uncertainty, $\theta q(p)$ means that the price chosen by the agent is independent of θ, since demand elasticity is independent of θ. The agent's incentive compatibility constraint is therefore particularly simple: $p = \arg\max_{\tilde{p}}(\tilde{p} - w)q$ or $(p - w)/p = 1/\varepsilon$ where $\varepsilon = d\ln q/d\ln p$ is the elasticity of demand. The multiplicative form of demand uncertainty thus leads to a degenerate class of contract problems, in which the agent's first-best action is independent of the state.

The individual rationality constraint, that the agent not be made worse off by participating in the contract, applies ex post. That is, the agent has the option, upon realization of θ, to refuse the contract. For a given contract (F,w), the agent will consider ex post whether to remain in the contract on the basis of whether her profits are positive. It is straightforward to show that the agent will accept the contract ex post if θ is not less that a critical value $\hat{\theta}$. The optimal two-part price contract solves the following:

$$\max_{F,w,\hat{\theta},p} \int_{\hat{\theta}}^{\overline{\theta}} F + (w - c)\theta q(p)\, d\theta$$

subject to

$$p = \arg\max_{\tilde{p}}(\tilde{p} - w)q(\tilde{p}) \qquad \text{(ICC)}$$

$$(p - w)\hat{\theta}q(p) - F \geq 0 \qquad \text{(IR)}$$

This is a "Disneyland pricing" problem [146]. In the solution to this problem, the optimal variable price, w^*, exceeds c and the agent earns rents in every realization $\theta > \hat{\theta}$.[8] The positive wholesale margin, $w^* - c$,

[8] The proof that $w^* > c$ can be sketched as follows. Suppose that $w^* \leq c$. The manufacturer can raise w^* slightly above c and reduce the fixed fee F by an amount that leaves the marginal type $\hat{\theta}$ indifferent. This ensures that the set of agents accepting the contract, $[\hat{\theta}, \overline{\theta}]$, remains unchanged. All agents of type $\theta > \hat{\theta}$, who buy more than the marginal agent $\hat{\theta}$, end up paying more under the new contract. Since the individual rationality constraint is not binding for any of these agents, the manufacturer's profit increases.

is used by the manufacturer to extract additional rents at high realizations of demand, at the cost of forgoing rents in states $\theta < \hat{\theta}$, and at the cost of inefficiently low quantities of sale in all states. Importantly, this is a second-best solution. Two distortions are involved. First, each retailer sets a price that is inefficiently high at the optimum because $(w - c) > 0$ yields a vertical externality — a benefit to the upstream firm that is not taken into account in the agent's decision. Second, retailers below $\hat{\theta}$ drop out whereas all retailers would sell in the first-best optimum. The distortions that remain under Disneyland pricing raise the question, addressed below, as to whether more elaborate contracts can yield greater profit.

A more general class of strategies available to the upstream manufacturer is nonlinear pricing. A general nonlinear pricing scheme specifies a payment from the retailer to the manufacturer as a function of the amount of the input purchased: $R(x)$. The optimal such scheme can be characterized via application of the *revelation principle*. This principle states that an optimal mechanism can always be formulated in terms of a report by the agent as to the state of demand (i.e., the optimal mechanism is a *direct mechanism*) with incentive compatibility constraints that it is optimal for the agent to announce the truth (an *incentive compatible* direct mechanism). This is a standard nonlinear pricing problem which in general yields higher profits than two-part pricing. We refer the reader to ref. [190] or [205].

6.2.1.2 Vertical Restraints

We turn now to contracts involving not just pricing but also vertical restraints on the actions of the downstream agent. In this setting, under the multiplicative demand assumption, a strategy for the complete resolution of incentive problems is immediate. The first-best outcome is easily achieved via a contract that includes a restraint that the downstream firm set price to maximize $(p - c)q$, i.e., that the agent set price to solve $(p - c)/p = 1/\varepsilon$. This achieves maximum total profits.

The ease of this solution is misleading, however. It is an artifact of the multiplicative form that the first-best price is independent of the realization of demand. More generally, if elasticity varies with the realization of θ the optimal contract will involve a second-best outcome, even with some vertical restraints on price, or payments that depend upon retail price.

6.2.2 Inventory Decision Only

Let us move from the coordination of pricing decisions to the second of the basic supply chain decisions, quantity. With uncertainty in demand, quantity becomes an economic decision that is distinct from pricing. Quantities, in reality, must (at least to some degree) be determined prior to the resolution of uncertainty. Hence the classic inventory problem: selecting the optimal inventory ex ante to the realization of uncertain demand.

With price exogenously determined, the solution to this problem in the simplest setting for a single firm is well-known. As explained in Section 5.2, in the newsvendor setting, the firm knows the distribution of demand, $F(x)$, and chooses inventory y to maximize profits, $pE\min\{x,y\} - cy$. From this, the fractile rule for optimal inventory y^* follows directly from the first-order condition: $1 - F(y^*) = c/p$. Equivalently, $F(y^*) = (p - c)/p$.

In a vertical setting, with a uniform wholesale price, w, and a contractually enforced retail price, p, the downstream firm chooses inventory to satisfy $1 - F(y) = w/p$. If $w > c$, the downstream firms choice of inventory is inefficient. Again, we run into the vertical externality: the downstream firm ignores the profits going to the upstream firm in the form of the upstream margin, $(w - c)$ with each unit of inventory added. The downstream firm therefore orders an inefficiently small number of units.

The most direct solution to this incentive problem (in a world without wealth constraints) is a two-part pricing scheme, with a variable price that reflects the true marginal cost to the supply chain, c, of the inventory. Once again, this is an implementation of the residual claimancy principle.

Another celebrated solution is suggested by Pasternack [153]: an inventory buyback at a rate b. Consider the design of a contract (w, b) by an upstream firm, consisting of the wholesale price w and the buyback rate, b. Under such a contract, the inventory problem solved by the downstream firm is

$$\max_y p \int_0^y x dF(x) - [w - bF(y^*)]y \tag{6.6}$$

The appropriate buy-back contract can elicit elicitation of the first-best inventory y^*, defined by $F(y^*) = (p - c)/p$. To see this, note that the agent's optimal y under the contract, from the first-order condition corresponding to (6.6), yields

$$F(y) = (p - w)/(p - b) \tag{6.7}$$

Choose b to solve the following

$$b = \frac{w - c}{p - c} \cdot p \tag{6.8}$$

Then (6.7) and (6.8) imply $F(y^*) = (p - c)/p$. Thus, the right choice of b yields the first-best solution.

One of Pasternack's insights is that there is a degree of freedom left in choosing w and b to solve (6.8). *Any* increase in the two instruments that preserves the equality (6.8) yields the efficient solution and distributes a greater share of profit upstream. Thus a buyback policy allows both first-best efficiency and arbitrary division of rents.

But the question arises, what problem does the Pasternack contract really solve? One problem it solves is eliciting optimal inventory by a single downstream agent when price is fixed exogenously, as we have explained. But this problem already has an easy solution: the two-part pricing scheme with the efficient transfer price, c. We suggest in Section 11 that the Pasternack solution is uniquely robust to the introduction of asymmetric information in the knowledge of the demand distribution F in a way that the residual claimancy contract is not.

6.2.3 Coordination of *Both* Price and Inventory Decisions

6.2.3.1 Restatement of Price-setting Newsvendor Problem for a Centralized Firm

We consider next the coordination of price and inventory decisions. Consider first the case where the downstream firm, facing uncertain demand, chooses both price and inventory. For a centralized supply chain, this problem, the "price-setting newsvendor," has been studied extensively (see ref. [155] for a review). We reformulate the price-setting newsvendor problem before addressing incentive coordination issues.

A long establishment observation in the price-setting newsvendor literature within management science is that additive and multiplicative uncertainty have different qualitative effects on the optimal price. Specifically, with additive uncertainty the optimal price is less than the optimal "riskless price," while this inequality is reversed under multiplicative uncertainty.[9]

Salinger and Ampudia [165] explain this observation by applying a modified version of the Lerner rule to the price-setting newsvendor problem. From an economist's point of view, it is natural to think of the Lerner rule as applying to this context: a price-setting newsvendor is simply a monopolist facing demand uncertainty. Recall that the Lerner rule follows from the optimality condition that marginal revenue equals marginal cost. The usual definition of marginal revenue is the additional revenue from *selling* one more unit while the definition of marginal cost is the additional cost from *producing* an extra unit. In the monopoly pricing problem under certainty, the quantity produced is always equal to the quantity sold, but this is not so under uncertainty. One can adapt the Lerner rule to the price-setting newsvendor by ensuring that marginal revenue and marginal cost are measured on a consistent basis: either per expected unit sold or per unit produced.

We reformulate the Salinger–Ampudia argument following Krishnan [107]. Denote the number of sales by $t(p,y;\theta) = \min\{q(p;\epsilon),y\}$. The expected profit function is

$$E\Pi(p,y;\theta) = pEt(p,y;\theta) - cy \tag{6.9}$$

Note that each unit in inventory will either be sold or will be an "overstock." Denote the number of "overstocks" as $O(p,y;\theta) = [y - q(p;\theta)]^{+}$. The inventory level $y = ES + EO$, i.e., the available inventory

[9] The optimal riskless price is the optimal price for a monopolist facing a demand function where the random variable representing uncertain demand is replaced by its mean value.

is equal to the expected sales plus the expected overstocks. The profit function can be rewritten as:

$$E\Pi = pES - c(ES + EO) \tag{6.10}$$

The optimal price for the firm is characterized by the following first-order condition[10]:

$$\begin{aligned} \frac{\partial E\Pi}{\partial p} &= \frac{\partial[pES - c(ES + EO)]}{\partial p} \\ &= p\frac{\partial ES}{\partial p} + ES - c\left(\frac{\partial ES}{\partial p} + \frac{\partial EO}{\partial p}\right) = 0 \end{aligned} \tag{6.11}$$

We can rearrange (6.11) as follows:

$$\frac{p - \left(1 - \frac{\frac{\partial EO}{\partial p}}{-\frac{\partial ES}{\partial p}}\right)c}{p} = \frac{1}{-\frac{\partial ES}{\partial p}\frac{p}{ES}} \tag{6.12}$$

Denote the denominator of the RHS of Equation (6.12), $\frac{\partial ES}{\partial p}\frac{p}{ES}$, by η_A. Now consider the second term in the numerator of the LHS of Equation (6.12), $\left(1 - \frac{\frac{\partial EO}{\partial p}}{-\frac{\partial ES}{\partial p}}\right)c$. This term can be rewritten as $(\frac{\partial ES + \partial EO}{\partial ES})c$. Because $y = ES + EO$, this can further be simplified as $\frac{\partial y}{\partial ES}c$. Denote this term by C. Substituting $\eta_A = \frac{\partial ES}{\partial p}\frac{p}{ES}$ and $C = \left(1 - \frac{\frac{\partial EO}{\partial p}}{-\frac{\partial ES}{\partial p}}\right)c$ into Equation (6.12) gives us the Lerner rule modified for the price-setting newsvendor:

$$\frac{p - C}{p} = -\frac{1}{\eta_A} \tag{6.13}$$

Note that $\eta_A = \frac{\partial ES}{\partial p}\frac{p}{ES}$ is "the elasticity of the average quantity sold with respect to price." And $C = \frac{\partial y}{\partial ES}c$ is the "marginal expected cost of an expected unit sold."[11] By expressing both the elasticity and

[10] Another first-order condition, for the inventory decision y, is needed to fully characterize the price-setting newsvendor's problem. Here, however, the focus is only on the pricing decision.

[11] This is because for each expected unit sold, we will have to produce some extra inventory, which will end up as an expected overstock. The term $\frac{\partial y}{\partial ES}c$ exactly captures the extra cost required (through investment in inventory) for one extra unit of expected sales.

marginal cost in terms of each unit *sold*, the above equation yields a modified Lerner rule for the price-setting newsvendor. Thus the price-setting newsvendor problem is reduced to the most fundamental of all economic formulations of monopoly power: the Lerner equation.

6.2.3.2 Coordination of Incentives in a Decentralized Firm

We now move to the coordination of price and inventory decisions in our context of a 1–1 supply chain. Recall that total profits are given by $\Pi(p,y;\theta) = pt(p,y;\theta) - cy$. Maximizing $E\Pi(p,y;\theta)$ with respect to p and y yields the following first-order conditions as a characterization of the centralized optimum:

$$Et(p,y) + \frac{\partial Et(p,y)}{\partial p}p = 0 \tag{6.14}$$

$$p\frac{\partial Et(p,y)}{\partial y} - c = 0 \tag{6.15}$$

Let (p^*, y^*) denote the first best (p,y) that solves these two conditions. Under a uniform pricing contract (i.e., a wholesale price only), the retailer will maximize $E\pi(p,y;\theta)$ with respect to p and y, given w. The retailer's first-order conditions are:

$$Et(p,y) + \frac{\partial Et(p,y)}{\partial p}p = 0 \tag{6.16}$$

$$p\frac{\partial Et(p,y)}{\partial y} - w = 0 \tag{6.17}$$

Anticipating the comparison of centralized, or efficient, choices with decentralized incentives throughout this monograph, it is useful to describe the difference between decentralized and centralized marginal returns as follows:

$$\frac{\partial E\pi(p,y;\theta)}{\partial p} = \frac{\partial E\Pi(p,y;\theta)}{\partial p} \tag{6.18}$$

$$\frac{\partial E\pi(p,y;\theta)}{\partial y} = \frac{\partial E\Pi(p,y;\theta)}{\partial y} - \overbrace{(w-c)}^{\text{vertical externality}} \tag{6.19}$$

It is clear, from a comparison of (6.15) and (6.17), that the uniform pricing contract cannot achieve the first-best outcome. This is because of the vertical externality on inventory leads to a divergence between these two optimality conditions: this is the vertical externality captured in (6.19). Note, however, that in (6.18) there is no vertical externality on price: *the double-markup effect is at work only on the inventory decision.* This follows from the fact that given a level of inventory, the retailer will choose price to maximize revenue, which is exactly what the centralized firm would do at the same level of inventory.

The incentive problem can be resolved easily with a standard residual claimancy contract, a two-part price with variable price equal to c. Unlike the Pasternack inventory-only problem, however, here an instrument consisting of w and a buy-back price b (and no fixed fee) cannot elicit first-best decisions. To show this, note that under a buy-back policy the retailer's profit function is $\pi(p,y;w,b,\theta) = pt(p,y;\theta) - wy + bO(p,y;\theta)$ and the agent's first-order conditions in choosing p and y are:

$$Et(p,y) + \frac{\partial Et(p,y)}{\partial p}(p-b) = 0 \tag{6.20}$$

and

$$p\frac{\partial Et(p,y)}{\partial y} - w + b(1 - \frac{\partial Et(p,y)}{\partial y}) = 0 \tag{6.21}$$

Setting $w > c$ in order to collect revenues under a (w,b) contract necessarily creates a vertical externality in the agent's inventory decision. As in the Pasternack inventory-only model, this externality can be offset by setting $b > 0$ to induce the retailer to hold more inventory. Here $b > 0$ now *creates* a vertical externality in price. Specifically, as a result of $b > 0$, the retailer will raise price (and sell fewer units) because the cost of raising price is partially borne by the manufacturer. The externality decomposition with $b > 0$ is the following:

$$\frac{\partial E\pi(p,y;\theta)}{\partial p} = \frac{\partial E\Pi(p,y;\theta)}{\partial p} - \overbrace{b\left(\frac{\partial Et(p,y)}{\partial p}\right)}^{\text{vertical externality}} \tag{6.22}$$

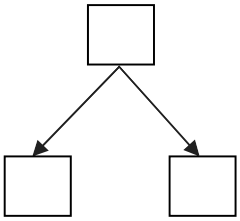
$$\frac{\partial E\pi(p,y;\theta)}{\partial y} = \frac{\partial E\Pi(p,y;\theta)}{\partial y} - \overbrace{(w-c) + b\left(1 - \frac{\partial Et(p,y)}{\partial y}\right)}^{\text{vertical externality}} \tag{6.23}$$

A buyback price, b^*, that sets the vertical externality in (6.23) to zero would achieve first-best inventory *conditional* upon the retail price. But the retail price p would vary with the buyback price b. The decomposition in Equation (6.22) reflects this fact. Since $b > 0$ introduces a new vertical externality on price, this means that the buy-back policy must be accompanied by a price restraint to yield the first-best outcome. A price *ceiling* will constrain the retailer's incentive to raise price. The prediction, in this 1–1 market structure, is that buy-backs and price *ceilings* are complementary instruments.

The above arguments can be extended to the coordination of non-price decisions such as effort. Taylor [188] and Krishnan et al. [108] consider the coordination of inventory and effort decisions in a supply chain. In Taylor's model, effort is exerted ex ante while the Krishnan, Kapuscinski and Butz model considers ex post effort. Both models show that a buy-back alone will, as in the price and inventory argument above, distort the retailer's incentive to exert effort by subsidizing unsold inventory. In this multi-task environment, a first-best outcome requires that buy-back contracts be accompanied by additional instruments such as sales rebates or markdown allowances which subsidize retailer effort.

7

Downstream Duopoly: The 1–2 Market Structure



A very rich setting for explaining supply chain contracts is a market structure in which two competing downstream firms purchase from an upstream monopolist. (See Figure 1.1(d).) To begin, we show below that two-part pricing in the simplest version of this new market structure continues to eliminate the double markup problem. This was true in the 1–1 market structure, but here two-part pricing includes a variable price that now exceeds marginal cost. At the heart of all of the new results in this section is the effect of moving from a setting with a single potential externality (the *vertical* externality) to a setting with two externalities, the *vertical* externality and a *horizontal* externality. A horizontal externality is one imposed by an agent on a competitor within the same organization, e.g., contractually linked to the same upstream firm.

In developing the implications for contracts of the 1–2 market structure, we draw upon the "balancing-externalities" methodology.[1] Contractual instruments such as the wholesale margin, royalties and buy-back prices affect the mix of vertical and horizontal externalities.

[1] See refs. [110, 137, 206].

Consider a downstream firm taking many actions. If the externalities on a marginal change in any particular action by the firm are *balanced* in the sense of summing to zero, then the agent's action maximizes combined profits for the entire set of firms in the contract. To achieve first-best profits, a contract need restrict only those actions for which externalities cannot be balanced.

The balancing-externalities methodology is first set out under conditions of certainty. When we turn subsequently to the analysis of contracts under uncertainty, we shall argue that an important failure of the price system to coordinate incentives arises from an inherent "missing externality" in downstream firms' decisions. The consequence of the missing externality, we shall argue, is the impossibility of balancing externalities within a set of contracts containing only prices. This creates a role for vertical restraints.

7.1 Certainty

7.1.1 Price as the Only Downstream Decision

7.1.1.1 Uniform Pricing

Consider a supply chain in which an upstream firm supplies to two symmetric downstream firms, both of which sell to the same set of buyers. The downstream firms are imperfect substitutes, competing in prices (i.e., as Bertrand competitors), with demand curves given by $q_i(p_1,p_2), i = 1,2$. The downstream firms incur no cost other than the wholesale price, w, set by the upstream firm, which in turn has a constant per-unit cost c.

The upstream firm's choice of optimal w can be expressed as

$$\max_{w,\boldsymbol{p}} \pi(\boldsymbol{p},w) = (w-c)(q_1(\boldsymbol{p}) + q_2(\boldsymbol{p})) \tag{7.1}$$

subject to the (incentive compatible) constraint that (p_i,p_j) is an equilibrium in the downstream market, given w.

$$p_i = \arg\max_{\widetilde{p}} (\widetilde{p} - w) q_i(\widetilde{p},p_j) \tag{7.2}$$

Assuming a symmetric equilibrium, the pair of incentive compatibility constraints (one for each downstream firm) can be reduced to a single constraint (7.2). We can re-express the constraint (7.2) in terms of the optimal Lerner markup rule. The symmetric decentralized price, p_d, is given by:

$$\frac{p_d - w}{p_d} = \frac{1}{\varepsilon} = \frac{1}{\varepsilon_m + \varepsilon_{ij}} \tag{7.3}$$

where ε is the elasticity of demand in the symmetric equilibrium and ε_{ij} is the cross-elasticity of demand. Equation (7.3) uses the fact that, in a symmetric duopoly, the elasticity of demand facing a single firm is the sum of the elasticity of market demand and the cross-elasticity of demand[2]: $\varepsilon = \varepsilon_m + \varepsilon_{ij}$.

The benchmark against which this decentralized equilibrium is assessed is the centralized firm. The centralized firm's problem results in an optimal downstream price, p^*, given by

$$\frac{p^* - c}{p^*} = \frac{1}{\varepsilon_m} \tag{7.4}$$

Note that if $\varepsilon_{ij} = 0$, i.e., each downstream firm is a local monopolist, the double mark-up problem is identical to the previous section. Compare (7.3) and (7.4) in the case where $\varepsilon_{ij} > 0$. If w were set equal to c, then the comparison shows that p_d would be less than p^*. Total profits would not be maximized because price is too low. As w is raised above c, not only does the upstream monopolist increase its share of the pie (starting from a zero share) it increases the size of the pie. Suppose that the monopolist continues to raise w to the value, w^*, that elicits p^*. Total profits are maximized at this wholesale price. As w is raised slightly above w^*, however, the first-order impact on total profits is zero by the envelope theorem. And the value to the monopolist of raising its share further is a positive first-order effect. (The monopolist's share of profits is strictly increasing in w.) The implication is that, in this uniform pricing case, the monopolist's optimal wholesale price exceeds w^*. The equilibrium retail price exceeds p^*. The double markup problem is mitigated by imperfect downstream competition, but remains.

[2] Market demand elasticity is defined as $|d\ln q(p,p)/d\ln p|$, and cross-elasticity of demand is defined as $|d\ln q_i(p_i,p_j)/d\ln p_j|$.

To express this mathematically, note that the manufacturer's profits can be written as the product of its share of profits, and total profits:

$$\Pi_u = \frac{(w-c)}{(p(w)-c)}(p(w)-c)q[p(w)] \tag{7.5}$$

Evaluating the derivatives of each term at w^* (the value of w that elicits p^*), $d[(w-c)/(p(w)-c)]/dw > 0$ and $d[(p(w)-c)q[p(w)]]/dw = 0$. Hence, evaluated at w^*, $d\Pi_u/dw > 0$. The outcome under uniform pricing involves prices above the first-best price level, just as in the basic Spengler double mark-up problem.

7.1.1.2 Two-part Pricing

Two-part pricing, as in the two-stage market power case, eliminates double marginalization. Here, however, the optimal variable wholesale price is the wholesale price w^* that elicits p^*. This is the same w^* that elicits p^* in the uniform pricing case. (But note that under two-part pricing, w^* is optimal, whereas under uniform pricing, equilibrium w and p exceed w^* and p^*.) The optimal two-part pricing scheme is characterized by the maximization (with respect to w and p) of the sum of all profits in the supply chain system subject to the incentive compatibility constraint that w elicit p:

$$\max_{w,p} 2q(p,p)(p-c)$$

subject to

$$p = \arg\max_{\widetilde{p}} (\widetilde{p} - w)q(\widetilde{p}, p)$$

The objective function is maximized at the first-best price, p^*, given by

$$\frac{p^*-c}{p^*} = \frac{1}{\varepsilon_m} \tag{7.6}$$

The incentive compatibility constraint can be written as

$$\frac{p-w}{p} = \frac{1}{\varepsilon_m + \varepsilon_{ij}} \tag{7.7}$$

Achieving the first-best with two-part pricing is simply a matter of setting w so that the right-hand sides of (7.6) and (7.7) are equal. The required wholesale price, w^*, exceeds c from the fact that $\varepsilon_{ij} > 0$.

The retailer's pricing decision in this setting is the first decision in our framework that is distorted by more than one externality. We can set out the impact of the externalities to illustrate our balancing-externalities approach in the simplest way. The downstream firm's first-order condition, maximizing π_1, and the centralized firm's first-order condition, maximizing Π, can be compared as follows:

$$\frac{\partial \pi_1}{\partial p_1} = \frac{\partial \Pi}{\partial p_1} - \overbrace{\frac{\partial q_1}{\partial p_1}(w-c)}^{\text{vertical externality}} - \overbrace{\frac{\partial q_2}{\partial p_1}(p-c)}^{\text{horizontal externality}} \tag{7.8}$$

The vertical externality works through $(w - c)$ and the impact on quantity q_1. The horizontal externality, which works through the impact on q_2, contains a "purely horizontal effect," operating through $(p - w)$ and an additional vertical component, operating through $(w - c)$. The externalities have opposite signs, and the optimal wholesale price, w^*, balances the two externalities exactly, setting the sum of the last two terms in (7.8) equal to zero.

7.1.2 Price and Effort

We begin to appreciate the richness of supply chain contracts once we expand firms' strategy space beyond price. In actual retail markets, firms do much more than set price. Virtually anything that retailers do, from keeping stores tidy to adequate staffing with cashiers to providing sales advice and promotion, adds to demand.

We continue to assume that the demands for the product downstream at the two retailers are symmetric, with the demand at retailer 1 now given by $q_1(p_1, s_1; p_2, s_2)$ where p_i and s_i are the price and sales effort or service at the two outlets.[3] Effort is measured in units of the dollar cost of effort.

We move directly to two-part pricing (skipping the uniform pricing case). Retailers bear costs given by the wholesale price, w, paid to the

[3] This discussion draws upon Winter [206].

upstream manufacturer, a fixed fee paid to the manufacturer for the right to carry the product, and expenditure s_i on sales effort. The profit for retailer i gross of the fixed fee is denoted by π_i, and the total profit, for the manufacturers and both retailers is denoted by Π. These profit functions are given by

$$\pi_1(p_1, s_1; p_2, s_2) = q_1(p_1, s_1; p_2, s_2)(p_1 - w) - s_1 \tag{7.9}$$

$$\pi_2(p_1, s_1; p_2, s_2) = q_2(p_1, s_1; p_2, s_2)(p_2 - w) - s_2 \tag{7.10}$$

and

$$\begin{aligned}\Pi((p_1, s_1; p_2, s_2) &= q_1(p_1, s_1; p_2, s_2)(p_1 - c) \\ &\quad + q_2(p_1, s_1; p_2, s_2)(p_2 - c) - s_1 - s_2\end{aligned} \tag{7.11}$$

Assume that Π is maximized at a symmetric set of prices and effort levels and denote this optimum by (p^*, s^*). Assume further that the profit functions are concave. Can the symmetric (p^*, s^*) be elicited with a single instrument, w, or are more complex contracts required?

The key to understanding any incentive distortions in retailer decisions is to isolate and decompose the difference between the marginal gain in *individual* profit and the marginal gain in *total* profit from a change in either p_i or s_i. From Equations (7.9)–(7.11), it follows that at a symmetric configuration ($p_1 = p_2 \equiv p$ and $s_1 = s_2 \equiv s$), this difference can be expressed for retailer 1 (and similarly for retailer 2) as follows:

$$\frac{\partial \pi_1}{\partial p_1} = \frac{\partial \Pi}{\partial p_1} - \overbrace{\frac{\partial q_1}{\partial p_1}(w - c)}^{\text{vertical externality}} - \overbrace{\frac{\partial q_2}{\partial p_1}(p - c)}^{\text{horizontal externality}} \tag{7.12}$$

$$\frac{\partial \pi_1}{\partial s_1} = \frac{\partial \Pi}{\partial s_1} - \overbrace{\frac{\partial q_1}{\partial s_1}(w - c)}^{\text{vertical externality}} - \overbrace{\frac{\partial q_2}{\partial s_1}(p - c)}^{\text{horizontal externality}} \tag{7.13}$$

The individual retailer's private optimum in setting p_1 is distorted from the collective optimum by the two externalities: when p_1 is raised, the manufacturer collects the wholesale markup, $(w - c)$, on a smaller demand through retailer 1. This is the *vertical* externality. The effect of this externality is to distort the retailer's price upwards. The second externality operates through the cross-elasticity of demand between the two retailers. When p_1 is raised, the competing retailer collects the retail markup $(p - w)$ on an additional $\partial q_2/\partial p_1$ units and the manufacturer collects the wholesale markup $(w - c)$ on the same additional units; these add up to the term labeled *horizontal* externality in Equation (7.12). The effect of this externality is to distort the retailer's price downwards. The same two externalities distort the sales effort decision. For each instrument, the vertical and horizontal externalities act to distort the decentralized decision in opposite directions.

When will the right choice of w alone elicit the optimum p^* and s^* at both outlets? That is, when does the price system elicit (privately) efficient incentives in this model of a distribution system? This efficiency property will hold when the value of w that renders the sum of the last two terms of (3) to zero at the optimum $(p^*, s^*; p^*, s^*)$ also renders the sum of the last two terms of (4) to zero. The externalities must balance in both first-order conditions at the same value of w. If they do not then the price system fails. The externality-balancing condition can be expressed as $\epsilon_p^r/\epsilon_s^r = \epsilon_p^m/\epsilon_s^m$, where ϵ_p^r and ϵ_s^r are retailer elasticities with respect to p and s and the right-hand side terms are market elasticities.[4]

Let ε_p^i and ε_s^i be the elasticities of demand with respect to price and service that each of the downstream firm faces. These are the percentage impacts on demand of a 1 percent change in price or service respectively. The ratio $\varepsilon_p^i/\varepsilon_s^i$ measures the marginal rate of substitution between price and service in "producing" a given level of demand.[5] Let ε_p^m and ε_s^m be the same elasticities of demand but at the market level, i.e., as seen by a centralized firm. Then a simple manipulation of (7.12)

[4] This condition for the efficiency of the price system can also be derived as a direct implication of the classic theorem by Dorfman and Steiner [58] that optimal choice of p and s leads to equality of the elasticity ratio with the ratio of revenue to s.

[5] Our terminology is borrowed from production theory or preference theory. Given a production function $F(x_1, \ldots, x_n)$, the marginal rate of technical substitution between x_i and x_j, when these variables are measured in logarithms, is given by

$$\frac{\partial \ln F}{\partial \ln x_i} \Big/ \frac{\partial \ln F}{\partial \ln x_j}$$

This is the percentage change in the input x_i that must accompany a 1 percent change in x_j in order to leave total output unchanged. In our context, we are assuming that demand is "produced by" inputs of price and service.

and (7.13) shows that at the centralized optimum, there exists a value of w, w^*, that simultaneously renders the last two terms of these equations equal to zero.

This is intuitive. The decentralized firms are *biased* towards excessive price competition (compared to the centralized firm's objective function, total system profit) if $\varepsilon_p^i/\varepsilon_s^i > \varepsilon_p^m/\varepsilon_s^m$ and if the inequality is reversed, the firms are biased towards excessive effort or service competition. Only if the marginal rate of substitution for the decentralized firm is identical to that of the centralized firm is first-best efficiency achievable through decentralization. No restraints are necessary and the price system itself results in perfect coordination in this case.

This reduces the problem of why firms would have the incentive for complex contracts (the failure of the price system) to an inequality. But why would a decentralized downstream firm in this case be biased towards or away from price competition?

Three main arguments have been given that explain a bias towards price competition and away from service competition. The first is a traditional free-riding argument [189]. Consider a stereo store providing expert sales advise, listening rooms and a comfortable shopping environment — and in the same area of a town or city, another store provides the identical products (model numbers) but "in the box" with no service whatsoever. Consumers can choose the model that they want from the up-market store, then walk to the other store and purchase the model at a low price. The low-priced store free rides on the provision of information at the high-priced store, thus compromising or eliminating the high-priced store's incentive to provide information at all. The sales of the product suffer because the price system alone has failed to elicit the right incentives for information provision.

A second possibility for market failure is in the case where sales effort may be product promotion or information provision that can be targeted to a local area or selected set of potential consumers for the retailer [137]. Lowering price attracts consumers from other retailers, expanding the retailer's market area — but raising promotional effort works on expanding potential set of consumers *within* the retailer's area. The cross-elasticity of demand (between the retailers) with respect to price is positive, but the cross-elasticity of demand with respect to promotion is zero. Letting the cross-elasticities be ε_p^{ij} and ε_s^{ij}, we have (for the case of a symmetric duopoly)

$$\begin{aligned} \varepsilon_p^i &= \varepsilon_p^m + \varepsilon_p^{ij} \\ \varepsilon_s^i &= \varepsilon_s^m + \varepsilon_s^{ij} \end{aligned} \tag{7.14}$$

If $\varepsilon_p^{ij} > 0$ and $\varepsilon_s^{ij} = 0$, then it follows from (7.14) that $\varepsilon_p^i/\varepsilon_s^i > \varepsilon_p^i/\varepsilon_s^i$, and the market is again biased, from the perspective of total profit maximization, towards excessive price competition.

The third argument for bias towards excessive price competition (and failure of the price system) is that consumers are heterogenous in their willingness to shop and in the value that they put on service [206] — and that there is a positive correlation across consumers between the willingness to search and the tolerance for low levels of service. An individual retailer's strategy is designed around attracting consumers both into the market and away from other retailers, whereas the centralized firm (collective) profits are optimized by focussing only on attracting consumers into the market. The retailer's marginal rate of substitution between the two instruments, $\varepsilon_p^i/\varepsilon_s^i$, is greater than the manufacturer's ratio $\varepsilon_p^m/\varepsilon_s^m$: given the positive correlation between willingness to shop and tolerance for low service, to attract consumers away from rival outlets, the retailer relies more on low prices than high service. Again, the retailer bias is towards excessive reliance on low prices, compared to the ideal, centralized firm-optimum (p^*, s^*).

We have outlined three reasons why prices alone cannot elicit efficient retailer incentives. Efficient retailer incentives (in price and service) can be achieved, however, with a vertical restraint on price. Take the case of a bias towards too much price competition, i.e., where the sum of externalities in (7.12) is negative at the optimal wholesale price. Suppose that the manufacturer establishes a price floor at p^*. Then lowering w has the effect of raising the retail margin $(p - w)$, which is the retailer's marginal benefit to investing in service. (Price is prevented from following the wholesale price downwards by the price floor constraint.) By increasing the marginal retailer benefit to offering service, lowering w will elicit greater service. The wholesale price will be lowered to the point where service reaches its efficient level, s^*. Mathematically, taking price as fixed at p^*, w is set to the level that makes the last two terms in (7.13) sum to zero. This is a mechanism that *directly* raises incentives for greater service.[6]

Klein and Murphy [106] suggest an additional, indirect mechanism by which a price floor elicits greater service. Contracts between a manufacturer and retailers may, explicitly or implicitly, constrain service to an efficient level. Monitoring may be possible, but costly. In any incentive setting where agents are undertaking decisions that are monitored infrequently and with cost, it pays a principal (here the manufacturer) to protect rents among the agents: the rents provide a "carrot" that will be lost by any agent that is terminated. Protecting retailer rents through resale price maintenance (and low or zero fixed fees), in combination with monitoring by the manufacturer, elicits greater service. (This kind of argument is known as an "efficiency-wage" argument; see ref. [175].)

The three theories outlined above for explaining the source of retailer bias towards excessive price competition — and the prediction that price floors can resolve the incentive distortion — solve the puzzle of why vertical price *floors* have been so popular, relative to vertical price ceilings. A vertical price ceiling is easily explained as resolving the double markup problem that was discussed in Section 6. Yet vertically imposed price floors, when legal, have been much more popular than price ceilings (see, e.g., [90]).

7.2 Uncertainty

Uncertainty is central to inventory decisions. A vast literature has developed in both management science and economics on the coordination of supply chains with *competition* among downstream firms. We begin this section with a review of selected articles from this literature, and then offer an integrated model.

A building block to the theory of competing firms within a vertical supply chain is the analysis of the horizontal duopoly game between firms making inventory decisions. A standard reference to this problem

[6] If, on the other hand, there is a bias towards too little price competition in favor of too much service competition, i.e., if at the optimal wholesale price the sum of the elasticities in (7.12) is positive, then a price ceiling elicits first-best incentives.

is [125]. This paper considers two competing newsvendors who interact in a market. The setting includes the following: uncertain demand faced by the two competing newsvendors; a fixed (exogenous) price; spillover of demand from one firm to the other in the event of a stockout.

In the Lippman–McCardle model, the competing newsvendors face random quantities of demand, q_1 and q_2, which are (in alternative variations) independent, identically distributed random variables or derived from a random aggregate demand that is allocated to the two firms according to one of various sharing rules. The exogenous price is denoted by v. The authors consider a game in which the firms choose inventories (y_1, y_2). If there is a stockout at firm j, with $q_j > y_j$, then an exogenous fraction a_i of the stockout is added to the demand at firm i. The quantity demanded from firm i is thus[7]

$$R_i = q_i + a_i(q_j - y_j)^+$$

The description of the game is completed with the specification of the payoff to each firm i, in terms of expected profit[8]:

$$\pi_i(y_1, y_2) = vE\min\{R_i, y_i\} - cy_i$$

Lippman and McCardle offer two main results. First, a pure strategy Nash equilibrium exists in this model. Second, if $a_i = 1,\ i = 1,2$ then the total industry inventory is always greater than in the case of a monopoly.[9] That is, competition increases inventories. This is a version of the standard economic result that competition in markets increases quantities.

The Lippman and McCardle model is purely horizontal, rather than set in a supply chain context. Bernstein and Federgruen [22], in another prominent contribution, add an upstream manufacturer. Bernstein and Federgruen also endogenize prices, so that coordination of both decisions is at issue. The authors propose a general formulation of demand, but for tractability rule out demand spillovers entirely.

Bernstein and Federgruen begin with independent (non-competing) retailers, which effectively yields (in our terminology) a 1–1 market

[7] Throughout, $(x)^+$ means $\max\{0, x\}$.

[8] Lippman and McCardle include as well a constant rate at which excess inventory can be salvaged, as well as a constant penalty cost per unit of demand not met.

[9] This result is easily extended to the more general case of $a_i > 0$.

structure. In this case, we can write the demand functions $q(p,\theta)$ as depending on own-price as well as a random shock, θ. The payoff functions for the case of an integrated firm can be expressed as

$$\Pi(p,y) = pE\min\{y, q(p,\theta)\} - cy = (p-c)y - pE[y - q(p)]^+$$

We let the total-profit-maximizing, or efficient, values of the decision variables be (p^*, y^*). Can (p^*, y^*) be elicited in a decentralized supply chain? In the decentralized case, Bernstein and Federgruen allow the manufacturer to offer a buy-back price, b, for retailers. As well, in their model, the wholesale price w is linked to the retail price via a functional relationship (a "price discount scheme") $w = W(p)$. Bernstein and Federgruen restrict W to be linear, with an intercept at c: $W(p) = c + ap$. The retailer's payoff function, given the contract $(b;a)$ is given by

$$\pi(p,y;b,a) = [p - W(p)]y - (p-b)E[y - q(p)]^+$$

The manufacturer's payoff in the decentralized case is

$$\pi_m(p,y;b,a) = [W(p) - c]y - bE[y - q(p)]^+$$

Bernstein and Federgruen, in the spirit of Pasternack [153], show that with the instruments b and a, the first-best profits can be achieved, i.e., (p^*, y^*) elicited, with any desired distribution of profits between the retailer and the manufacturer.

With competing downstream retailers, the combination of perfect coordination plus complete flexibility in implementing desired sharing of profits between upstream and downstream firms can again be achieved, but the required functional dependence of w upon all prices is now more complex. Bernstein and Federgruen select a particular functional form for $w_i = W(\boldsymbol{p})$, and show that with this additional form (and their other assumptions, including the absence of spillovers), that first-best coordination is achieved. Again this includes the flexibility to implement an arbitrary division of total system profits between upstream and downstream players.

Narayanan et al. [143] also develop a model with an upstream monopolist and competing retailers, for the purpose of investigating contracts that will coordinate inventory and pricing incentives in the supply chain. The Narayanan–Raman–Singh (NRS) model is somewhat restrictive, in that it assumes (as do Bernstein and Federgruen) that there are no demand spillovers in the event of stockouts. Demand is assumed to be perfectly correlated between retailers as well in this model. NRS do allow, however, lump sum transfers from retailers to manufacturers with the result that the supply chain coordination issue is about implementing the efficient (total profit maximizing) choices of price and inventory.[10] NRS show that the two instruments, wholesale price and buy-backs, are sufficient to elicit the first-best choices of price and inventory.

Butz [32] makes a similar point. Price, however, is flexible in the Butz model, being chosen ex post. But the parallel argument holds: buy-backs combined with a wholesale price instrument ensure optimal prices and inventories. Butz shows that the alternative instrument of price restraints also achieves the efficient outcome, and that the optimal form of vertical price restraints is a price floor, as most often observed in practice.

Krishnan and Winter [110]: We turn next to the externality-balancing approach to the coordination problem in the case of uncertain demand. This approach we applied in the certainty case above, in the case of price only and price and service. With uncertain demand, we consider incentives in two actions, prices and inventories.

The following model is developed in ref. [110]. A single manufacturer sells a product through two competing retailers. The demands for the product at the two outlets are $q_i(p_i, p_j; \theta), i, j = \{1, 2\}, i \neq j$ and θ is a random variable. Before the realization of uncertainty, the retailers set prices and choose inventory levels, y_1 and y_2.

Let $\lambda_i(p_i, p_j; \theta)$ be the proportion of any excess demand from outlet i that would spill over to outlet j in the event of a stock-out at outlet i, and similarly for λ_j. Note that the proportion of inventory that "spills

[10] The Pasternack consideration of dividing profits between upstream and downstream firms is set aside.

over" in the event of a stockout depends on the prices at the two outlets, but not on the inventory levels.[11]

The timing of the game is as follows. The manufacturer offers contracts; the two retailers observe the contract offers and simultaneously choose to accept or not; the two retailers simultaneously choose price and inventory levels from some intervals $[0,\overline{p}]$ and $[0,\overline{y}]$; uncertain demand is realized; and finally, consumers make purchase decisions.

Given prices $(p_1,p_2) \equiv \mathbf{p}$ and inventories $(y_1,y_2) \equiv \mathbf{y}$ and the realization of θ, the demand at outlet i is given by $D_i(\mathbf{p},y_j;\theta) = q_i(\mathbf{p};\theta) + \lambda_j(\mathbf{p};\theta)(q_j(\mathbf{p};\theta) - y_j)^+$. The "transactions" or sales at outlet i are given by

$$\begin{aligned} t_i(\mathbf{p},\mathbf{y};\theta) &= \min\{y_i, D_i(\mathbf{p},y_j;\theta)\} \\ &= y_i - (y_i - D_i(\mathbf{p},y_j;\theta))^+ = y_i - O_i(\mathbf{p},\mathbf{y};\theta) \end{aligned} \tag{7.15}$$

where $O_i(\mathbf{p},\mathbf{y};\theta)$ is the number of "overstocks" at outlet i.

The realized profit of outlet i can be expressed as

$$\pi_i(\mathbf{p},\mathbf{y};w,\theta) = p_i t_i(\mathbf{p},\mathbf{y};\theta) - wy_i - F \tag{7.16}$$

The realized profit of the entire supply chain is

$$\Pi(\mathbf{p},\mathbf{y};\theta) = p_1 t_1(\mathbf{p},\mathbf{y};\theta) - cy_1 + p_2 t_2(\mathbf{p},\mathbf{y};\theta) - cy_2 \tag{7.17}$$

Uniform pricing and the missing externality: We can now derive the first-order conditions, from (7.17) and (7.16), and compare the individual incentives and collective efficiency in price and inventory decisions. By taking the difference in first-order conditions we get following equations that are parallel to Equations (7.12) and (7.13) in Section 7.1.2:

$$\frac{\partial E\pi_1}{\partial y_1} = \frac{\partial E\Pi}{\partial y_1} - \overbrace{(w-c)}^{\text{vertical externality}} - \overbrace{p_2\frac{\partial Et_2}{\partial y_1}}^{\text{horizontal externality}} \tag{7.18}$$

$$\frac{\partial E\pi_1}{\partial p_1} = \frac{\partial E\Pi}{\partial p_1} - \overbrace{p_2\frac{\partial Et_2}{\partial p_1}}^{\text{horizontal externality}} \tag{7.19}$$

[11] In ref. [110], a structured model is provided which, when developed from first principles, satisfies this condition.

A fundamental feature of this model, in contrast to the model in Section 7.1.2, is the “missing externality” in the last equation. The missing externality captures the fact that, once we control for the level of inventory, *pricing decisions are not subject to a vertical externality.* In other words, the vertical externality imposed by the manufacturer’s wholesale mark-up is reflected through the inventory decision alone. Under a simple pricing contract, once the inventory choices are made, the manufacturer has no direct (vertical) interest at all in the price at which the inventory is resold: the wholesale revenue, costs and hence profits are completely determined by the inventory purchase.

But the simple pricing contract does not implement the efficient outcome (p^*, y^*). Because of the missing vertical externality on price, the horizontal externality distorting the price decision cannot be balanced off with an offsetting vertical externality. This result relies only on the necessary first-order conditions for the aggregate optimum.

Contracts that achieve coordination: More complex contracts, such as those that allow for vertical price restraints or inventory buy-backs, can compensate for the missing vertical externality on price. To see the role of price restraints, note that if profit functions are quasi-concave, the efficient (p^*, y^*) can be elicited with a price floor at p^*, a fixed fee F, and a linear wholesale price w^*. The role of the price floor is clear. The missing externality in the outlet’s first-order condition on price leaves the outlet with the incentive to drop price below p^*.[12] The floor constrains the retailer against pricing below the first-best. The wholesale price can then be set such that the outlets, each facing a newsvendor problem, have the right incentives to choose y^*, the efficient inventory levels.

In contrast to a price restraint, a buy-back policy resolves the incentive problem by *creating* a vertical externality: once the manufacturer agrees to take back unsold inventory this affects the pricing decisions of the retailers. The buy-back price can be chosen to be of the right magnitude to precisely offset missing externality. The analytics are

[12] A technical condition on the sign of the derivative dEt_2/dp_1 is imposed, but this is satisfied for common distributions [110].

straightforward [110]. The right buyback price is the buyback price that creates exactly the vertical externality that satisfies the balancing condition.[13]

This model yields the insight that RPM and buyback policies are substitutes, as in Butz [32]. Other instruments can "fill in the missing externality" to solve the coordination problem. A price-discount scheme, for example, links the wholesale price to the retail price and therefore gives the manufacturer a direct interest (through its revenue receipts) in the price charged by the retailer, even holding inventory constant.

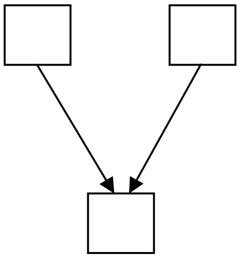
The externality-balancing approach is a promising framework for investigation of the coordination of decisions beyond price and inventory, along a supply chain. Consider, for example, the incorporation of sales effort decisions into a supply chain model with the market structure of one firm upstream and two downstream. The potentially complex analytics of this model can be clearly organized around the balancing conditions. And the approach suggests an answer to a puzzle regarding another form of vertical restraints, territorial exclusivity. Why would a manufacturer ever establish exclusivity — local monopolies — in a downstream market? In the simplest of models, creating market power downstream exacerbates the double-markup problem and reduces total profits. Yet we commonly observe this practice, in contexts ranging from franchising to automobile dealerships. One answer to the puzzle is that territorial protection mitigates (and in the extreme case, eliminates) horizontal externalities. This leaves us with the two-stage market power model in which coordination is easily achieved with two-part pricing — no matter how many demand instruments there are. Territorial exclusivity allows the residual claimancy principle to be employed to resolve incentive incompatibilities through the use of efficient transfer prices.

[13] The expression for the optimal buyback price is

$$b^* = -\frac{\epsilon_{pc}^t}{\epsilon_p^t}\, p^*.$$

8

Two Upstream Firms, Selling Through One Downstream Firm: The 2–1 Market Structure



The 2–1 market structure (Figure 1.1(e)) covers three general areas in the economics of supply chains. The first is the *assembly problem* in which a downstream firm must strike contracts that provide for the provision of complementary inputs and adequate inventories of the inputs upstream. The second is the *common agency* problem in which multiple upstream manufacturers compete in contract offers to a common downstream retailer and in sale, through the retailer. The third context, which is in the mainstream of the economics literature rather than the management science literature, is *entry deterrence*: the design of supply chain contracts to gain a strategic, entry-deterring advantage over a potential entrant.

8.1 Assembly Models

The assembly problem involves coordinating the supply of multiple, complementary, inputs required (in fixed proportions) in the production or assembly of a final product. The assembly problem has been studied extensively in the supply chain literature. Our discussion here approaches this problem through the lens of balancing externalities.

Netessine and Zhang [145] have followed this approach. While the Netessine–Zhang model deals with a 1–2 market structure — the distribution of complementary goods purchased from a single upstream firm — and the assembly models deals with production of complementary goods that is used in assembly downstream, both classes of models have the same incentive structure. Firms producing (or distributing) complementary goods at one level deal with a firm at the other level.

Importantly, all firms in the assembly structure are producing complements. This is in contrast to the 1–2 market structure where the upstream and downstream activities are complements (both manufacturing and retailing are necessary inputs into the final product), but the downstream firms' actions are substitutes. As we discussed in Section 7, it is this combination of vertical complementarity and horizontal substitutability that allowed the externalities to balance out, under a two-part pricing contract. When all actions in the supply chain are complementary, as we shall see, the distorting externalities all have the same sign and cannot balance.

Upstream inventory decisions: Consider a simple model with a downstream assembler trying to coordinate the inventory decisions of upstream suppliers. Assembly models typically assume that the assembler offers take-it-or-leave-it contracts to the suppliers. We adopt this assumption and start with simplest case. One unit each of two upstream inputs, 1 and 2, is needed by a downstream manufacturer m for assembling a final product. Let the downstream firms' market price, p, be determined exogenously, and the demand for the final product be $q(\theta)$, where θ is a random variable, and let the marginal production costs of the inputs be c_1 and c_2. We start with an analysis of a simple two-part pricing contract, where the manufacturer sets a variable wholesale price, w_i, for each retailer and (because of the fixed transfer) optimal contracts can be characterized as maximizing total system profits. Under this contract, each input supplier faces a newsvendor problem, and chooses inventory levels, y_1 and y_2. The number of units of the final product sold depends on three factors: the random demand, and the inventory levels chosen by each input supplier. The assembly of the inputs into the final product is flexible in the sense that it takes place ex post. Define $t(q(\theta), y_1, y_2) = \min\{q(\theta), y_1, y_2\}$ as the number of transactions or sales of the final product.

The expected profit of the manufacturer is:

$$E\pi_m(w_1, w_2) = (p - w_1 - w_2)Et(q(\theta), y_1, y_2)$$

The profit of input supplier 1 is:

$$E\pi_1(w_1, y_1; y_2) = w_1 Et(q(\theta), y_1, y_2) - c_1 y_1 \tag{8.1}$$

and supplier 2's profit is similar.

The profit of the centralized supply chain is:

$$E\Pi(y_1, y_2) = pEt(q(\theta), y_1, y_2) - c_1 y_1 - c_2 y_2 \tag{8.2}$$

As before, we can derive the first-order conditions from (8.2) and (8.1), and compare the individual incentives and collective efficiency in inventory decisions. Decomposing the first-order conditions, we get:

$$\frac{\partial E\pi_i}{\partial y_i} = \frac{\partial E\Pi}{\partial y_i} - (p - w_i)\frac{\partial Et}{\partial y_i} \tag{8.3}$$

Note that $\frac{\partial Et}{\partial y_i} \geq 0$; an increase in inventory at one of the input suppliers cannot decrease total sales. The centralized firm will set inventory levels such that the marginal cost of each unit of additional inventory is equal to the marginal expected revenue from eventual sales. Equation (8.3) shows that, under a simple pricing contract, each supplier will under-invest in inventory, since some of the revenue from additional units sold will be captured by both the other supplier and the manufacturer. Conditional on sufficient demand and sufficient inventory at supplier j, supplier i's investment in one additional unit of inventory will yield w_j to supplier j and $p - w_i - w_j$ to the manufacturer. Investments in inventory by a supplier will have a positive externality on *both* the manufacturer and the other supplier. In other words, both the vertical and horizontal externalities are positive. Unlike the 1–2 market structure case, the vertical and horizontal externalities on inventory cannot be rendered offsetting with a two-part pricing contract.

Contractual resolutions: An offsetting externality can be created, however, by a contract that commits the manufacturer to paying not just for the products that the manufacturer buys from the supplier, but also for each unit of inventory that the supplier produces. This inventory (or capacity) "pre-commitment" contract is analogous to a buy-back contract: a firm in the supply chain encourages another firm to invest in the efficient resource level. Consider a contract where the manufacturer pays the supplier a fee, s_i, per unit of inventory produced. For each unit of inventory that is eventually consumed, the manufacturer pays an additional w_i. Supplier 1's profit function under this pre-commitment contract is:

$$E\pi_1(w_1, s_1, y_1; y_2) = w_1 Et(q(\theta), y_1, y_2) - (c_1 - s_1)y_1$$

Following our familiar method, we can decompose the first-order conditions:

$$\frac{\partial E\pi_i}{\partial y_i} = \frac{\partial E\Pi}{\partial y_i} - (p - w_i)\frac{\partial Et}{\partial y_i} + s_i$$

The inventory fee, s_i, creates an incentive for the supplier to invest in more inventory. This fee, if chosen appropriately, can offset the negative externality that depresses inventory investment. For the two effects to cancel out, we need to set s_i and w_i such that:

$$\frac{s_i}{(p - w_i)} = \frac{\partial Et^*}{\partial y_i} \tag{8.4}$$

where Et^* is the transactions at the efficient outcome.[1] Note that (8.4) involves the selection of two instruments to satisfy one constraint. The implication of this fact is that (s_i, w_i) can be chosen not only to resolve the incentive distortion but to distribute profits as desired. The fixed fee becomes redundant (as in the Pasternack [153] and Marvel and Wang [133] solutions to the supply chain incentive problems).

We have sketched our own perspective on inventory incentives in assembly systems. The simple model presented above, with only upstream inventory decisions, can be extended to the case where the downstream manufacturer also makes a decision, either installing production capacity ex ante, or setting a retail price for the assembled

[1] The efficient outcome can also be achieved by simply letting the manufacturer dictate inventory levels to the retailer or can accept responsibility for the financial cost of upstream overstocks.

product. The management science literature offers a large number of recent contributions to this theory including refs. [19, 20, 95, 198].

8.2 Common Agency

We turn next to supply chain coordination of price and *promotion* (or effort) choices in a market structure with two firms upstream and one downstream. Real-world supply chains quite typically overlap in having a common retailer sell the products produced upstream. Virtually any retailer, in fact, sells the goods of multiple manufacturers. The 2–1 model is the simplest structure for uncovering and resolving the incentive issues that arise with common agency.

To this point in our monograph, a supply chain contract has been designed to achieve efficiency or maximum total profits for the contractual parties. In the common agency context, we have a setting where upstream firms compete in contract offers for representation in a downstream firm (a marketing agent or retailer). The key contributions here are refs. [17] and [18]. We begin by outlining the Bernheim–Whinston (BW) application of their common agency theory to a single retailer selling the products of competing manufacturers [17]. BW consider a setting in which contract offers are made by upstream suppliers to a number of agents, and in which the selection of the agent is therefore endogenous. To keep within the conventional supply chain context, however, we discuss the BW model in the case in which the agent is pre-determined — BW's *intrinsic* agency game.

In the BW model, there are two products, $i = 1, 2$, each produced by an upstream firm and sold through a downstream firm with a constant marginal cost c_i and possibly a fixed cost. The demand for product i depends upon its price, p_i, a level of marketing intensity, m_i, chosen by the downstream firm as well as a random variable, θ. The marketing intensity, m_i, is observable by the upstream firm and can be contracted upon. The two principals (manufacturers) offer contracts simultaneously to the single downstream agent. The contract from firm i specifies a reward schedule $I_i(x_i, m_i)$ as a function of x_i, which denotes the sales of product i, and m_i, the observable marketing effort by the downstream firm. Prices p_i are set by the upstream firms, and (p_1^c, p_2^c)

represents the cooperative, or collusive, prices that would be set by a fully integrated monopolist in the market.

BW show that there is an equilibrium for the intrinsic agency game in which principals set prices p_i^c and the reward schedules take the form of 100% of profits minus a lump sum:

$$I_i(x_i, m_i) = (p_i^c - c_i)x_i - K_i \tag{8.5}$$

That is, the downstream agent becomes the residual claimant for each upstream firm and therefore a residual claimant of profits for the entire industry. As (8.5) shows, dependence of the contract upon effort m_i is completely unnecessary. The intuition for this result is clear. Whatever the contract offered by the rival manufacturer, the residual claimancy contract between a manufacturer and the retailer must elicit a decision that maximizes the combined profits of these two parties. Residual claimancy contracts guarantee first-best profits because the agent undertakes effort decisions with zero externalities, internalizing all benefits from additional effort at the margin. With the agent collecting all marginal profit for the industry, it then pays each upstream firm i to set its price at p_i^c in order to maximize the value of the lump sum payment at which it effectively sells its profit stream to the agent.[2]

Cachon and Kok [36] analyze equilibrium contracts in exactly the common agency market structure (2–1). Cachon and Kok restrict the contract offers to one of three forms: uniform pricing, a quantity-discount contract, and a two-part tariff. The authors compare the welfare of the manufacturers and retailers *across* the games in which the contracts are restricted to the different forms. For example, allowing two-part tariffs, rather than uniform pricing can lead to more intense competition between the manufacturers with the result that manufacturers are worse off even though total industry profits improve.

Cachon–Kok and Bernheim–Whinston illustrate a difference that sometimes arises between the management science and economic approaches to the same problem. From an economist's perspective,

[2] The logic is maintained when the pricing decision as well as the marketing effort decision is allocated to the downstream agent.

the form of the contract should be determined *endogenously*, as in Bernheim–Whinston. No enforceable mechanism exists that would restrict the form of the contract (e.g., to either uniform pricing or to two-part pricing). Any restriction of the contract space is therefore *ad hoc*. To look at the issue in a different way, in any equilibrium of the Cachon–Kok uniform-pricing game, each contract is Pareto dominated for the contracting parties by a two-part pricing contract. Nothing internal to the model prevents the parties from simply moving to the superior contract.

8.3 Supply Chain Design with the Threat of Entry

To this point, we have focussed on contracts as coordinating the incentives of agents along a supply chain. Contracts can be designed to serve a completely different kind of role: contracts can have strategic value in competition against other manufacturers or other supply chains. While this role has not been featured in the literature on supply chain coordination, it is prominent in the contract theory foundations of antitrust policy. In this section, we offer an overview of one particular strategic role: the use of contracts to deter entry into a market.

The theory begins with the classic article by Aghion and Bolton [3]. An incumbent, in a market with a single downstream firm, faces the threat of entry by another firm. The downstream firm requires one unit of the input, which it values at a known value v. The incumbent's cost of producing this unit is c_I. Entry is uncertain because the entrant's cost, c, is random. Let $G(c)$ represent the distribution of the entrant's cost. The incumbent is assumed to have a first-mover advantage in contracting in the sense that it can contract with the downstream firm before the entrant's cost is realized. A contract can be expressed as a pair (p, d) consisting of a price that the buyer (the downstream firm) agrees to pay if it buys in the future as well as a stipulated damages, d, that the buyer pays if the decision is made not to purchase (i.e., too purchase from the entrant instead). Equivalently, the buyer pays d up front and an additional $(p - d)$ if the decision is made to purchase. In other words, the contract is a call option with an option price $p_o = d$ and an exercise price, $x = p - d$.

If a contract is not struck in the Aghion–Bolton model, the incumbent and the entrant compete in prices ex post. In the Bertrand equilibrium of this no-contract subgame, if $c < c_I$ then the entrant supplies ex post at a price equal to c_I; if $c > c_I$ then the incumbent supplies at a price $\min(c, v)$.

A contract yields a superior combined payoff to the buyer and incumbent than no contract, because if a contract is in place, then the entrant must meet the exercise price x of the contract option if it is to supply the buyer.[3] By lowering x below c_I, which is the price that a low-cost entrant would have to charge to attract the buyer in the absence of a contract, the incumbent and buyer as a pair are thus able to extract a lower price from the entrant (conditional upon the states of successful entry) than in the absence of a contract. The optimal choice of x satisfies

$$\frac{c_I - x}{x} = \frac{1}{\varepsilon} \tag{8.6}$$

where $\varepsilon \equiv d \ln G(c)/dc$ is the elasticity of the distribution function [96]. Since $G(p)$ is the probability that the entrant would be willing to supply at a price p, G can be interpreted as a supply curve of the entrant, and ε as the elasticity of supply. Equation (8.6) is the Lerner equation for a *monopsonist* (rather than the more familiar *monopolist* Lerner equation): the value that the incumbent–buyer pair obtain with the purchase from the incumbent is c_I, the incumbent's cost of production, because this is the opportunity cost avoided when the purchase is made. The incumbent and buyer, as a pair, sign a contract that extracts the optimal monopsony rents from the entrant.

A contract can thus be designed to extract a transfer from a party *outside the contract* — in the Aghion–Bolton model, a transfer from the entrant. Jing and Winter [96] suggest a variation on this theory. The variation involves actions by four parties: the incumbent, the buyers downstream from the incumbent, a set of entrants, and an input supplier further upstream from the incumbent. Suppose that entrant's costs are common but random. Instead of market power resting with

[3] If buyers have an option to buy at x, then an entrant can attract them only with a price offer less than x.

an entrant, the supplier of an input further upstream is a monopolist (with known cost). The incumbent anticipates future negotiations with the upstream supplier over its input price. The incumbent strikes a long term contract (with a relatively low exercise price) with the downstream buyer. The downstream contract makes the market less profitable for the entrants because they have to meet a low exercise price. This reduces the willingness-to-pay of the entrants for the upstream input (since it is now harder for them to cover costs upon entry). Since selling to an entrant is the upstream monopolist's best alternative to selling to the incumbent, the incumbent is therefore able to bargain for a lower input price. Thus, long-term contracts at one stage of the supply chain may be used by a firm to extract rents from a firm with market power at another stage of the supply chain. Jing and Winter apply this theory to an antitrust case in the supply of marketing information.

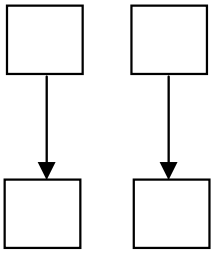
Let us return to the framework of a single potential entrant into a market with one incumbent. Consider now the possibility of many buyers. If the buyers (downstream firms) are all local monopolists, and the entrant has constant returns to scale, then the story is unchanged from Aghion–Bolton. If a potential entrant has a fixed cost, so that it needs to capture a substantial share of the market to cover the fixed cost and justify entry, then the theory of exclusionary contracts is strengthened. To enter the market with a given cost realization, an entrant needs to capture an adequate share of free buyers, those who are not committed to a contract or who would choose not to exercise an option contract. The incumbent can profitably sign up substantial numbers of buyers to long-term contracts because each buyer, in signing a contract, ignores the externality imposed on other buyers from its decision to accept the contract and consequent reduction in the probability of successful entry. Each additional contract reduces the number of free buyers, thus dampening the discipline that the incumbent faces in competing for the free buyers ex post from potential entry. Each contract with a buyer allows the entrant to extract additional profits from the free buyers. This theory is developed in refs. [158] and [169].

If buyers compete, however, the Aghion–Bolton theory can break down [62]. A single buyer downstream may hold-out from joining the contract, and then is available for the entrant. The entrant can charge a price that allows the single buyer to undercut the remaining buyers and capture the entire market. As Fumagalli and Motta demonstrate, this possibility eliminates the exclusionary effect of contracts when buyers compete downstream.[4] Simpson and Wickelgren [177] offer additional insights in analyzing exclusionary contracts in the case of competing downstream buyers.

[4] Wright [208] and Abito and Wright [1] offer an alternative analysis of the impact of buyer competition downstream in the Fumagalli–Motta model.

9

Competing Supply Chains



We have, to this point, been analyzing the role of contracts as coordinating incentives along a single supply chain, within the context of different market structures along the chain. The consideration of *competing* supply chains, Figure 1.1(f), expands on the strategic role for contracts introduced in the discussion of entry-deterrence in the last section. The observable choice of contracts by firms in a supply chain will influence the decisions that rivals, in other supply chains, take in the future. The design of contracts is again *strategic*, benefiting the firm by eliciting a change in rivals' behavior.

The central example consists of two supply chains each with a manufacturer and an independent retailer. The supply chains compete. In the absence of competition, the manufacturer would aim to elicit decisions on the part of downstream firms that duplicated the decisions of a centralized firm. A manufacturer that could control costlessly the actions of downstream firms through vertical integration would do so. A residual claimancy contract, i.e., two-part pricing with the variable price equal to marginal cost, would achieve the same result. When we move to considering competing manufacturer-retailer pairs (supply chains), however, it may be optimal for a supply chain to retain an element of decentralization. For a manufacturer competing in prices in a differentiated products setting, decentralization makes the supply chain a "softer" competitor. The supply chain's reaction curve, in competing against the rival firm, rises. The rival, observing this increase, sets a higher price. In other words, the rival responds by also adopting less aggressive pricing. The equilibrium in the pricing game is established at a higher price as a result of the commitment achieved through decentralization via contracts.

The key articles in the literature on strategic decentralization of vertical restraints are refs. [28, 138, 161]. The following is a simple model that illustrates the main point of this literature, and is drawn from Bonnano and Vickers. Two manufacturers, 1 and 2, sell products for which demands are $q_i(p_1,p_2), i = 1,2$, and produce at costs (c_1,c_2). The manufacturers each have the option of selling through the intermediation of a retailer. In the first stage of a two-stage game, the manufacturers each set w_i as well as a fixed fee (a franchise fee); in the second stage the retailers, one for each manufacturer, set prices. A standard residual claimancy contract would be one in which $w_i = c_i$. This contract, equivalent to vertical integration, would be optimal for either firm in the absence of the firm's rival. In the strategic setting, however, the optimal wholesale prices satisfy $w_i^* > c_i$, under standard regularity conditions on demand.[1] When w_i is set greater than c_i, retailer i is committed to a higher reaction curve. Figure 9.1 illustrates the logic. Whatever the decision of the other manufacturer as regards vertical integration or separation, it is optimal for each manufacturer to maintain separation since this elicits a higher price in the final retailer's pricing game.[2]

[1] These conditions are that demands be downward sloping, $\partial q_i/\partial p_i < 0$; concave, $\partial^2 q_i/\partial p_i^2 \leq 0$; and that products be substitutes, $\partial q_i/\partial p_j > 0$. Letting $R^i(p_1,p_2;w_i) = (p_i - w_i)q_i(p_1,p_2)$ be retailer profits, with subscripts as derivatives, further conditions are stability of the Bertrand retailer game, $R_{ii} + R_{ij} < 0$ and that the goods be strategic complements, $R^i_{ji} = q^i_j(p) + (p_i - w_i)q^i_{ji}(p) > 0$.

[2] When the strategic variables in the competing supply chain model are quantities, rather than prices, decentralization has no strategic value. In this case, the actions by competing supply chains are strategic substitutes rather than strategic complements. Reaction functions are downward-sloping rather than upward-sloping. In this case, committing to a soft reaction function through decentralization would be disadvantageous, in making the

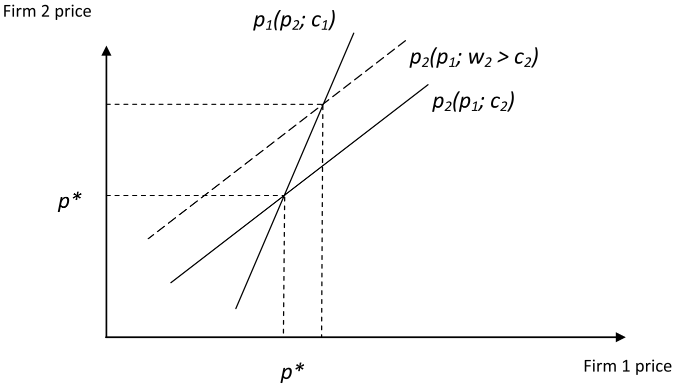


Fig. 9.1 Reaction functions under price competition.

It is important to note, as many authors have, that this conclusion depends on the ability of each manufacturer–retailer pair to commit to a contract that cannot be secretly renegotiated. Contracts that are renegotiated in secret will revert back to $w_i = c_i$, since contracts struck or renegotiated in secret lose the strategic motivation entirely. A commitment to leave town is essential for the principal.

Katz [98], however, notes that if the downstream firm is risk-averse (e.g., because it has low capitalization and wants to avoid bankruptcy) then the privately optimal contract, even struck in secret, retains the property that $w_i > c_i$. The strategic gains from separation are therefore retained even with secret renegotiation if downstream agents are risk averse.

Related papers: The McGuire and Staelin [138] model, which restricts manufacturers to offering uniform pricing contracts to retailers, shows that strategic decentralization by both manufacturers is an equilibrium when product differentiation is small. Coughlan [46] extends the McGuire and Staelin [138] model to more general demand functions, and also provides some preliminary empirical evidence in support of this model: firms entering a new more market were more likely to use an intermediary if the market was competitive. Gupta and Loulou [77]

rival supply chain more aggressive in the competition in quantities. (See refs. [30, 47, 190, p. 207].)

introduce a second consideration into the model: manufacturers invest in process improvements in order to reduce production costs. They show that decentralization remains an equilibrium even when investments in process improvement are accounted for. In addition, decentralized channels invest less in process improvement, further raising prices in the decentralized channel.

10

Dynamics

Many of the fundamental issues in management science require a dynamic rather than a static perspective. The basic management of inventories is inherently a dynamic problem due to fixed costs, production or procurement lead times, inventory perishability, nonstationarity of demand, and so on. Dynamic issues arise even with the ability of consumers strategically to delay purchases of durable goods and to stockpile goods in anticipation of future consumption and prices.

We begin this section with a review of the classic Coase conjecture on optimal pricing by a durable goods monopolist. We discuss in the same subsection a parallel topic that is tied even more closely to supply chain management, the *storable* goods monopoly problem. We then synthesize the basic dynamics of inventory and price incentives in a stationary infinite-horizon model. This model relaxes the assumption of complete perishability of inventory in the classic newsvendor model, adopting as the key departure from the static framework a simple assumption of inventory durability. We conclude this section with a brief discussion of other issues in the dynamics of supply chain incentives.

10.1 The Durable Good and Storable Good Monopoly Problems

The durable goods monopoly problem was identified by Coase [44]. Consider a monopolist selling a durable good to a fixed set of buyers. If the monopolist could commit to selling at one date only, it would choose the monopoly price (and quantity) to maximize profits. Consumers who value the product more than the monopoly price would buy at that price. Any consumers with values below the monopoly price would not buy. If the monopolist *cannot commit* to selling at one time only, however, consumers would be naïve to purchase at the monopoly price initially. This is because once the monopolist has sold to the high-valuation customers, it then has an incentive to reduce price and sell to lower-valuation customers. The high-valuation customers, anticipating the reduction in price, will rationally delay their purchases. The monopolists' pricing power is undone because sales in the second (and future) periods essentially compete with sales in the first period. The monopolist competes with its future self.

The Coase conjecture is that as the time between selling dates converges to zero, the monopolist's price converges to the competitive price, which is marginal cost. Bulow [29] reviews this topic and discusses related issues.

Dudine et al. [59] focus on a parallel intertemporal demand incentive. The monopolist's inability to commit again drives the dynamics. While the durable goods monopoly problem focuses on buyers' incentive to *delay purchases*, the "storable goods monopoly" problem focuses on buyers' incentive to *advance purchases* by stockpiling goods for future consumption. While the operations and inventory literature has analyzed numerous incentives to stockpile inventory (long production lead times, fixed costs, and so on), the stockpiling incentive in this monograph is driven purely by strategic considerations. The argument in Dudine et al. [59] is the following. Suppose a monopolist faces an identical linear demand curve in each of two periods; let $q_t(p_t) = 1 - p_t, t = 1,2$ where t refers to the time period. The demand is that of a single buyer. Assume that the marginal production cost is zero. If the monopolist could commit to prices in both periods, then it would simply charge the monopoly price in each period, i.e., $p_t^* = 1/2$ and $q_t^* = 1/2$ for $t = 1, 2$.

Now suppose that commitment to future prices is not feasible. This provides an opportunity and incentive for buyers to influence future prices by buying goods in the first period for use in the second. Suppose the monopolist sets $p_1^* = 1/2$. The buyer, acting strategically, buys $q_1 = 1/2 + \epsilon$, where $\epsilon > 0$. Because the buyer has stockpiled ϵ units for use in period 2, the demand function in period 2 is now $q_2(p) = 1 - \epsilon - p$. The monopolist in this second period will maximize profits at a price $p_2 = (1 - \epsilon)/2$. As a consequence of the buyers ability to store the good, the price in period 2 is lower than the price that the monopolist could have charged if the monopolist had been able to commit up-front to the second period price.

The monopolist's inability to commit allows the buyer effectively to lower the second period price by stockpiling inventory. This lower price benefits the buyer *on all units* purchased in the second period. As long as the cost of stockpiling inventory is sufficiently low, the trade-off between the cost of stockpiling inventory and the benefit of a lower price in the second period works to the buyer's advantage. In response to the buyer's strategic behavior, the monopolist will attempt to deter the stockpiling by raising prices in the first period. In the equilibrium, however, the inability to commit to the second period price will always leave the buyer with an incentive to stockpile inventories.

10.1.1 Related Papers

Both the durable goods and the storable goods problems have been analyzed in the supply chain management literature. In particular, ref. [182] addresses intertemporal pricing with customers delaying purchases (the durable goods effect) and ref. [183] deals with customer stockpiling (the storable goods effect). In ref. [182] the monopolist has a fixed amount of inventory that has to be sold over a finite time horizon. Customers vary along two dimensions: their valuation for the product and their willingness to wait. Su demonstrates that prices can decrease or increase over time, depending on whether the high-valuation customers are less or more patient than the low-valuation customers. In ref. [183], customers can stockpile inventory and the manufacturer can commit to a price path. However, the manufacturer may benefit from offering period price promotions as long as frequent shoppers are willing to pay more than occasional shoppers.

Su and Zhang [184] consider a newsvendor setting in which customers, anticipating ex post markdowns on unsold inventory, may strategically delay purchases. The monopolist's choice of low quantity, i.e., high price, can alleviate this problem, but the monopolist may not be able to commit credibly. Su and Zhang introduce a commitment device that we have explored earlier in this review: decentralization of decision-making. A decentralized supply chain with a simple uniform pricing contract can, by introducing a vertical externality on price and inventory decisions, credibly commit to keeping prices high and quantities low. Arya and Mittendorf [13] also argue that "channel discord," i.e., strategic decentralization serving as a commitment device, can mitigate the durable goods monopolists' problem.

Anand et al. [5] analyze the storable goods problem using a supply chain model with an upstream seller and a single downstream buyer facing a downward sloping market demand. The downstream firm can strategically hold inventory in order to lower future prices. Anton and Varma [8] consider an oligopoly version of this problem. Buyers' incentives to stockpile, the basic problem as perceived by the seller, are exacerbated here because each firm bears only part of the future cost of consumers stockpiling inventory and each firm has therefore low incentive to suppress stockpiling.

Several recent papers consider related dynamic issues. Most consider strategic incentives to *delay* purchases: see refs. [185, 186, 187]. The strategic incentives to *stockpile* inventory, however, have received less attention in the supply chain literature. Inventory dynamics are central to supply chain coordination in reality and deserve further attention in the theory literature.

The classic Clark and Scarf [41] article develops a serial multi-echelon model, i.e., a model with one firm at each of many stages. By establishing the concept of echelon stocks[1] Clark and Scarf show

[1] The echelon inventory level at any stage of the supply chain is the inventory available at that stage and at all downstream stages.

that the multi-dimensional problem can be solved as a sequence of one-dimensional problems. Lee and Whang [119], Chen [40], Cachon and Zipkin [39] and Porteus [157] all analyze the coordination problem of attaining the optimal centralized outcome through appropriate transfers for each echelon.

10.2 The Dynamics of Price and Inventory Incentives

Let us step back to the basic static model of coordinating price and inventory incentives of Section 7.2. We argued in that section that the conceptual key to understanding observed contractual resolutions (including vertical price floors) is a missing externality. The upstream manufacturer, in the static model, has no direct (vertical) interest at all in the price at which inventory is resold, given the level of inventory that is transacted. From the perspective of the balancing-externalities approach to vertical contracting, this missing externality implies that the price system alone cannot possibly coordinate incentives. More complex contracts are necessary and the missing externality translates into specific predictions about the contracts that emerge.

Introducing a simple assumption of inventory durability, rather than complete perishability, changes things. *A vertical externality is resurrected*: when inventory can be carried over for use in future periods, the manufacturer again has an interest in the retailer's price decision because this affects how much the retailer will buy in future periods.

The following infinite horizon model developed in Krishnan and Winter [111]. In this 1–2 market structure, demand in each period is uncertain and depends on the price and inventory level chosen by each retailer i. In each period t, at retail prices $\mathbf{p}_t = (p_{1t}, p_{2t})$ and inventory levels $\mathbf{y}_t = (y_{1t}, y_{2t})$, the demand at retailer i is $q_{it}(\mathbf{p}_t, \mathbf{y}_t, \boldsymbol{\phi}_t)$, where $\boldsymbol{\phi}_t$ is a real-valued random variable representing demand uncertainty.[2] The number of *transactions* by firm 1 in period t is $T_{it}(\mathbf{p}_t, \mathbf{y}_t, \boldsymbol{\phi}_t) = \min(q_{it}(\mathbf{p}_t, \mathbf{y}_t, \boldsymbol{\phi}_t), y_{it})$. The parameter $\boldsymbol{\phi}_t$ is a joint distribution for the

[2] In this model note that inventory levels influence demand, and not just sales. This recognizes the fact that consumers will be attracted to shop at stores where they have a higher likelihood of finding the product in stock. Models that consider the strategic role of inventory in attracting demand include Dana and Petruzzi [50], Balakrishnan et al. [15] and Bernstein and Federgruen [21, 23].

two retailers, with a continuous density and independent and is identically distributed over time.

Let γ denote a "durability factor" of the inventory, which represents the value that the product maintains from one period to the next. The factor γ can be interpreted as the outcome of discounting and an exponential rate of inventory decay. In other words, one unit of inventory left over at the end of the previous period is worth $\gamma(1-c)$ in the current period. Krishnan and Winter [110] show that Bellman equation characterizing the optimal dynamic price and inventory reduces to the following stationary (myopic) problem:

$$\begin{align} E\Pi^d(\mathbf{p},\mathbf{y},\phi) \equiv p_1ET_1(\mathbf{p},\mathbf{y},\boldsymbol{\phi}) + p_2ET_2(\mathbf{p},\mathbf{y},\phi) - cy_1 - cy_2 \tag{10.1}\\ +\gamma c[y_i - ET_i(\mathbf{p},\mathbf{y},\phi) + y_j - ET_j(\mathbf{p},\mathbf{y},\phi)] \tag{10.2} \end{align}$$

where $E\Pi^d(\mathbf{p},\mathbf{y},\phi)$ is the firm's expected profit in each period. The firm chooses $(\mathbf{p},\mathbf{y})$ to maximize $E\Pi^d$.

Similarly, each retailer chooses:

$$E\pi_i^d(p_i,y_i,\boldsymbol{\phi};(p_j,y_j)) = p_iET_i(\mathbf{p},\mathbf{y},\phi) - wy_i + \gamma w[y_i - ET_i(\mathbf{p},\mathbf{y},\boldsymbol{\phi})] \tag{10.3}$$

As before, we can decompose the first-order conditions to isolate the incentive distortions:

$$\begin{align} \frac{\partial E\pi_i^d}{\partial y_i} &= \frac{\partial E\Pi^d}{\partial y_i} - \overbrace{(w-c)\left(1-\gamma\left(1-\frac{\partial ET_i}{\partial y_i}\right)\right)}^{\text{vertical externality}}\\ &\quad - \overbrace{(p_j-\gamma c)\frac{\partial ET_j}{\partial y_i}}^{\text{horizontal externality}} \tag{10.4}\\ \frac{\partial E\pi_i^d}{\partial p_i} &= \frac{\partial E\Pi^d}{\partial p_i} - \overbrace{\gamma(w-c)\frac{\partial ET_i}{\partial p_i}}^{\text{vertical externality}} - \overbrace{(p_j-\gamma c)\frac{\partial ET_j}{\partial p_i}}^{\text{horizontal externality}} \tag{10.5} \end{align}$$

Note that as long as $\gamma > 0$ then, unlike in the static case, the vertical externality on price reappears. This reflects the fact that, conditional on inventory, the price chosen by the retailer affects the amount of inventory carried forward, which in turn affects the amount that the manufacturer can sell to the retailers in future periods. (If $\gamma = 0$, however, the static model of Krishnan and Winter [110] emerges which implies that retailers will be biased towards price competition.) The bias towards price competition holds for small values of γ. Krishnan and Winter show under some conditions that there exists a threshold value of γ below which retailer's are biased towards price competition and above which the bias flips and retailers are biased towards inventory competition. In other words, inventory durability versus perishability is a key factor in determining incentive coordinating supply chain contracts.

11

Asymmetric Information

The placement of a section on asymmetric information towards the *end* of a monograph on applied contract theory must, to the experienced reader, seem strange. Contract theory is built upon the foundation of asymmetric information, either in hidden action or hidden information [27]. To this point, we have incorporated the assumption of asymmetric information only implicitly for the most part. Consider, for example, the assumption, in our analysis of the use of complex contracts to elicit the right inventory decisions, that contracts cannot specify directly the level of inventory. Nothing formal in the models of the literature reviewed prevents firms from contracting on all decisions, including contracting on the amount of inventory. Nothing, that is, except the *assumption* that inventory cannot enter the contract. Inventory, in reality, is very easily observed by an upstream input supplier since it is equal to the quantity purchased. The assumption that inventory cannot be contracted upon, justified implicitly by asymmetric information, speaks to a need for theories in which the asymmetry in information is *explicit*. It is a standard requirement of contract theory that the contracts be optimal within the informational and enforcement constraints of the environment.

We begin with a simple example of the insights that can be garnered by introducing asymmetry of information explicitly. Consider the classic Pasternack theory of buyback policies. In this theory, recall that under a 1–1 market structure, a single upstream firm sets a wholesale price w and an inventory buyback price b. The downstream firm solves $\max_y py - (p-b)E[y-x]^+ - wy$, which yields a choice of y satisfying $F(y) = (p-w)/(p-b)$. If the upstream firm sets b equal to $b^* \equiv [(w-c)/(p-c)]p$ then the optimal y^*, which satisfies $F(y^*) = (p-c)/p$ is elicited.

The inventory problem, under a 1–1 market structure, can be solved easily with a two-part price instead of a buyback policy.[1] The manufacturer under this strategy essentially sells the rights to sourcing its product at cost, in return for a lump sum. The optimal inventory problem is then internalized completely by the downstream agent and the first-best inventory decision is the result. Alternatively, a contract can be written which simply calls for inventory y^* to be purchased. These simple alternatives to the Pasternack contract are not ruled out by the informational structure of the model. *What formal problem does the Pasternack contract solve?*

In the simplest extension to asymmetric information, the Pasternack solution survives as an instrument that can elicit first-best incentives. The other contracts applicable under certainty (two-part pricing, and contracting directly on inventory) do not. To show this, suppose that at the time that the manufacturer offers a contract, the distribution F is itself private information to the retailer. The demand is random as in any inventory model, but now the *distribution* of demand is known to the retailer alone. We can represent this asymmetry in information as a joint distribution over final demand x and a signal θ that the retailer has observed but the manufacturer has not. Let this distribution be $G(\theta, x)$. The first-best choices of inventory that the manufacturer would like to elicit are those $y^*(\theta)$ that satisfy $G(\theta, y^*(\theta)) = (p-c)/p$.

The Pasternack solution elicits this first-best inventory conditional upon *any* realization of θ. This follows from the fundamental result

[1] We are setting aside the other constraints on implementing the residual claimancy contract, including any limits on the buyer's wealth level.

that the optimal fractile solution to the newsvendor problem (from the perspective of the centralized firm, with opportunity cost c per unit, or from the perspective of the decentralized firm, with cost w per unit and returns price b) yields a probability of default that is completely independent of the distribution of demand, i.e., it is distribution-free. The lack of knowledge of θ, i.e., the size of the market, has no impact on the optimal Pasternack contract.

We have demonstrated the robustness of the Pasternack contract in the case of a reservation profit of zero for the agent. This robustness extends to non-zero reservation profits provided that the reservation profit is proportional to the size of the market. If the agent's reservation profit is a fixed dollar amount, then the contract is not robust because the optimal w and b will vary depending on θ.

Another contract that is distribution-free is a revenue-sharing contract. An upstream manufacturer could simply specify a revenue-share, say 30%, and set wholesale price the unit at 30% of marginal production cost. In choosing inventories, the retailer maximizes 30% of profits, which is identical to maximizing the total profits.

Other contracts do not necessarily achieve the robustness of the Pasternack contract. For example, a two-part pricing cannot yield a first-best solution because under this strategy the manufacturer must know the size of the market (θ) to set the appropriate fixed fee in the contract. And a contract specifying the optimal inventory x directly does not work because the first-best level of inventory depends upon θ.

The robustness of Pasternack's solution to a lack of knowledge of θ on the part of the manufacturer exemplifies a desirable property of any solution to a decentralized supply chain problem: the contract eliciting efficient incentives is ideal if it is independent of the size of the market. That is, a doubling of market size should leave the first-best optimal contract unchanged.

11.1 Price and Inventory

Suppose that we introduce price as a decision variable, in addition to inventory. Under what conditions, under our simple asymmetric information world, can the optimal centralized solution continue to be decentralized?

Decentralization is possible when demand is multiplicative in the signal but not in general. To understand this result, recall the contractual solutions to the price and inventory incentive problem without explicit asymmetry in information. In a 1–1 market structure, when demand is uncertain and the downstream firm makes decisions on price and inventory, the following contracts can elicit the first-best decisions from the downstream agent provided that information is symmetric at the time of contracting: (a) a two part price; (b) a contract dictating downstream price and inventory, without a fixed fee; and (c) a price ceiling and a buyback price, without a fixed fee.

The last of these contracts is invariant to the size of the market — provided that the demand is multiplicative in the signal observed by the agent. If demand is $\theta q(p;\xi)$ where θ is a random variable observed by the agent and ξ is a random variable observed by neither agent, then the optimal price in the market $p^*(\theta)$ is independent of θ. This is because multiplicative shocks to demand are "iso-elastic" shifts in demand, leaving demand elasticity unchanged at any price, with the implication that the optimal centralized price is independent of θ. By setting a price ceiling at p^*, setting the necessary buyback price (which again is invariant to the realization of θ), the contract elicits the first-best decisions $(p^*, y^*(\theta))$ on the price and inventory. A mark-up $(w - c)$ collects profits for the upstream firm that are proportional to θ. The firms bear no cost at all from the information asymmetry. Once again, invariance of the first-best contract to the value of the realized signal is the key to the robustness of the optimal contract. Marvel and Wang [132] were the first to observe that knowledge by the manufacturer of the demand distribution is not necessary for implementation of a first-best contract in this context.

11.2 The Mechanism Design Approach to Supply Chain Contracting

The examples of supply chain contracts discussed in this section of the monograph have introduced asymmetric information. But they leave us with a framework in which *first-best profits* are achieved, just as in our models with symmetric information. The modern contracting literature, in contrast, is built on models that necessarily yield second-best equilibria, with incentive compatibility constraints that are binding. The principal–agent model is at the foundation of contract theory. This framework posits a principal (typically an upstream supplier) writing a contract with an agent or agents (typically downstream firms) in order to elicit efficient decisions on the part of the agents. The agent has either hidden action in that his effort cannot be observed or hidden information in that his characteristics, such as costs, cannot be observed. Because of exogenous uncertainty, the principal cannot draw an exact inference from the outcomes that he observes to the hidden action or information on the part of the agent.

If a residual claimancy contract could be written, then incentive distortions would be resolved in the principal–agent problem. In principal–agent contexts, three factors arise that rule out residual claimancy contracts. As discussed in Section 6.1.2, these give rise to three classes of principal–agent models: principal–agent models with a risk averse agent [83, 176]; principal–agent models with a wealth constraint on the part of the agent [166] and principal–agent models with two or more agents.

In the remainder of this section, we apply the principal–agent approach — more generally the mechanism design approach — to supply chain contracting. We start with the simplest model to understand the role of contracts in the resolution of hidden information or adverse selection problems. The contracting solution to problems of this type always involve a trade-off for the principal between eliciting better incentives on the part of the agent and leaving a higher share of rents with the agent. Appropriately designed supply chain contracts can alter this trade-off to the benefit of the principal (the manufacturer).

As always, we begin with the simplest model. This model seeks to characterize the optimal supply chain contract in a setting where a downstream retailer has private information on demand. The incentive issue is setting the right price. A manufacturer produces a product at zero cost with demand equal to $\theta q(p)$ where $q' < 0$. The manufacturer sells through a retailer. The retailer has private information on θ at the time of contracting (and, following the usual simplifying assumption, bears no costs other than the payment to the manufacturer). For expositional purposes, we assume that θ takes on only two values, θ_1 and θ_2, with $\theta_2 > \theta_1$. At the time of contracting, the retailer knows θ. The manufacturer does not. What kind of contract would the manufacturer and retailer write?

The revelation principle [51, 69, 141] tells us that we can, without loss of generality, restrict contracts to offers of *menus* from which the agent is allowed to choose. The simplest class of contracts specifies choices over quantity and total payment, $\{(Q_1, R_1), (Q_2, R_2)\}$. (This is equivalent to a nonlinear pricing contract.) An agent, after choosing (Q_i, R_i) from the menu will set price at the value where the demand is equal to Q_i.[2]

The optimal contract within this class follows a standard setup [134]. The principal maximizes profit subject to individual rationality constraints for both types of agents and incentive compatibility constraints, that each type weakly prefers the choice in the menu intended for him. It is well known that only one of the two IR constraints, and only one of the two IC constraints, will be binding in equilibrium. Including only the constraints that we know will be binding, this problem is

$$\max_{\{(Q_1,R_1),(Q_2,R_2)\}} R_1 + R_2$$

subject to

$$R_1 - q^{-1}\left(\frac{Q_1}{\theta_1}\right) Q_1 \leq 0 \tag{IR}$$

$$q^{-1}\left(\frac{Q_1}{\theta_2}\right) Q_1 - R_1 \leq q^{-1}\left(\frac{Q_2}{\theta_2}\right) Q_2 - R_2 \tag{ICC}$$

The first of these constraints is that the low-demand agent must earn non-negative net returns. The term $q^{-1}(\frac{Q_1}{\theta_1})$ is the price that this agent

[2] The choice from a menu offered in a contract is equivalent to the "announcement" by the agent of his type. A mechanism in which agents with private information announce their types is called a *direct mechanism.* The revelation principle states that the optimal allocation in an asymmetric information setting can be achieved with an incentive compatible direct mechanism.

will set, following the selection of (Q_1, R_1). The second constraint is that the type 2 agent prefers (Q_2, R_2) to (Q_1, R_1).

The first-best outcome (i.e., outcome without the constraints) satisfies $Q_i^* = \theta_i q(p^m)$ where p^m is the (centralized firm's) monopoly price corresponding to the demand curve $q(p)$, with R_i^* fully extracting the profits of agent i: $R_i^* = \theta_i q(p^m)$. It is clear that the first-best outcome $\{(Q_1^*, R_1^*), (Q_2^*, R_2^*)\}$ is not achievable with this simple nonlinear pricing mechanism. At the first-best mechanism, the incentive compatibility condition of the type 2 agent is violated, because instead of earning zero profits by telling the truth, this agent could earn positive profits by accepting (Q_1, R_1), since $q^{-1}(\frac{Q_1}{\theta_2}) * Q_1 - R_1 > q^{-1}(\frac{Q_1}{\theta_1}) * Q_1 - R_1 = 0$. In violating the incentive compatibility constraint, the type 2 agent would set a price for the small quantity Q_1 (at the high demand $\theta_2 q(p)$) that is greater than p^m. The (second-best) optimal mechanism responds to this potential IC constraint violation by shading Q_1 downwards from Q_1^* and reducing R_2 from R_2^*.

Price ceilings: The first-best optimum can be achieved, however, with a price ceiling at p^m. Under this constraint, in addition to the offer of the menu $\{(Q_1^*, R_1^*), (Q_2^*, R_2^*)\}$, while both types of agents earn zero profits neither type has an incentive to deviate. In particular, the type 2 agent cannot gain by selling Q_1^* at a price in excess of p^m.

Buy-backs: This setting allows a price ceiling to elicit the first-best (centralized) outcome, but only because the setting is very special. The first-best optimal price, p^m, is independent of θ because of the multiplicative nature of uncertainty. When, instead, we allow the optimal price to vary with uncertainty, then price ceilings do not elicit the first-best. But a role for *buy-back contracts* emerges, as Arya and Mittendorf [12] have shown. Arya and Mittendorf adopt an assumption of demand uncertainty that captures in an extreme but simple way the dependence of optimal price on the realization of uncertainty. Consider a manufacturer with cost $c(x)$ selling to consumers through a retailer, who bears no cost other than the cost of the product. Consumers each have a unit demand and reservation value made up of the sum of two parameters: $\theta + \alpha$. At the time of contracting, the retailer knows θ. The manufacturer does not. The prior probability densities of α and θ

are $f(\alpha)$ and $g(\theta)$. The inverse hazard function $H(\theta) = [1 - G(\theta)]/g(\theta)$ is non-increasing.

The manufacturer offers the retailer a menu of contracts from which to choose. The contract $[x(\hat{\theta});s(\hat{\theta});b(\hat{\theta})]$ contains quantity $x(\hat{\theta})$, a payment per unit $s(\hat{\theta})$ and a buy-back rate, $b(\hat{\theta})$ which is the right to return any amount purchased at a price. All are functions of the report $\hat{\theta}$ by the agent as to his type.

By the revelation principle, one can restrict attention to incentive compatible contracts, i.e., contracts that leave each agent with the incentive to report the truth about his type. The role of the buy-back policy is in following correlation: the right to resell all units back to the manufacturer at any rate b is *worth more to a low-θ agent than it is to a high-θ agent.* This follows from the fact that the high-θ agent can sell the units at a higher price than the low-θ agent. Anything that makes the menu choice intended for a low type relatively unattractive to a high type relaxes the incentive compatibility constraint (which is binding only in a downward direction). Note that $b(\hat{\theta}) - \theta$ is the critical value of α below which a refund is sought.

The retailer's and manufacturer's expected profit when the retailer's type is θ and the retailer reports $\hat{\theta}$, given the contract $[x(\hat{\theta});\ s(\hat{\theta});\ b(\hat{\theta})]$ are given by:

$$\begin{aligned}\pi_R(\hat{\theta}|\theta) &= \int_0^{b(\hat{\theta})-\theta} b(\hat{\theta})x(\hat{\theta})f(\alpha)d\alpha \\ &\quad + \int_{b(\hat{\theta})-\theta}^{\bar{\alpha}} [\theta + \alpha]x(\hat{\theta})f(\alpha)d\alpha - s(\hat{\theta})x(\hat{\theta}) \\ \pi_M(\hat{\theta}|\theta) &= s(\hat{\theta})x(\hat{\theta}) - c(x(\hat{\theta})) - \int_0^{b(\hat{\theta})-\theta} b(\hat{\theta})x(\hat{\theta})f(\alpha)d\alpha\end{aligned}$$

The optimal contracting solution solves

$$\max_{x(\theta),s(\theta),r(\theta)} \int_0^{\bar{\theta}} \pi_M(\theta|\theta)g(\theta)d\theta$$

subject to

$$\pi_R(\theta|\theta) \geq 0 \qquad \forall\theta \tag{IR}$$

$$\pi_R(\theta|\theta) \geq \pi_R(\hat{\theta}|\theta) \qquad \forall\hat{\theta},\theta \tag{ICC}$$

The optimal buy-back rate, $b^*(\theta)$ takes on a simple form in Arya–Mittendorf. These authors show that the value of the buy-back policy (or returns policy) is measured by the limit that it places on retailer rents. The inverse hazard rate represents these savings, and the optimal buy-back at any θ is set equal to the inverse hazard rate: $b^*(\theta) = (1 - G(\theta))/g(\theta)$.

Other contributions: Many other papers in the supply chain coordination literature incorporate asymmetric information. Corbett et al. [45] consider a manufacturer selling to a downstream retailer with private information on costs, in a model that shares some similarities with the famous Baron and Myerson [16] paper on regulating a monopolist with uncertain costs. These authors derive the optimal nonlinear pricing scheme as well as the optimal contract under additional assumptions on the feasible set of contracts. Blair and Lewis [26] consider the case of private information about demand, rather than costs, by the downstream retailer.

In the Blair and Lewis model, the downstream retailer increases demand through promotional effort. A low level of sales can be due to low promotional effort or a low state of demand; the manufacturer is unable to distinguish between these root causes. Thus Blair and Lewis bring to the supply chain context a canonical principal–agent problem with hidden action and hidden characteristics. The uniqueness of the supply chain setting expands the contract space in Blair and Lewis compared to either a general principal agent model or the setting considered in ref. [45]. Specifically, Blair and Lewis allow for vertical restraints on the downstream price, i.e., resale price maintenance. The optimal contract in Blair and Lewis contains resale price maintenance as well as quantity restrictions.

Asymmetric information is important not just in the 1–1 market structure that we have discussed in this section, but in other market structures as well. The 1–2 setting, e.g., a setting with one manufacturer selling to two retailers, builds on the canonical models of Demski and Sappington [53] and Lazear and Rosen [118]. The key idea from these models is that when hidden information is correlated across agents the report by one agent (in a direct mechanism) conveys information to the principal about other agents' hidden information. The principal can exploit this additional signal to achieve superior outcomes.

The papers discussed adopt the standard assumption that the agent possesses better information in some dimension (action or characteristics) than the principal. In the economics literature, the case where a principal or upstream supplier has private information as well has also been analyzed. An upstream supplier who has private information on the quality or potential success of a new market will signal this quality by retaining a high percentage of residual claim in a contract with a downstream firm that invests in distribution of the product. The contract may, for example, allocate payment to the upstream firm in terms of a downstream royalty or an input price above marginal cost — even if pricing at marginal cost were optimal in terms of eliciting the best incentives on the part of downstream agents. This is the *informed principal's problem* [171]. Desai and Srinivasan [56] analyze nonlinear pricing contracts in a supply chain context in which the manufacturer is informed about demand and the retailer undertakes effort. Gallini and Wright [66] apply the informed principal framework to the issue of technology transfer through licensing contracts.

In the supply chain context, a buyback policy is a natural instrument for an informed principal, since the offer to buy back inventory can signal the principal's private information about high demand. This informed-principal strategy has not, to our knowledge, been explored to date.

We have followed the literature on supply chain control and indeed most of the literature on asymmetric information in economics in taking the state of information as exogenous. Other articles in both economics and management science have taken the important step of studying incentives for management of information flows or exchange. This literature is reviewed in Section 13.

12

Supply Chain Management, Relational Contracting and the Theory of the Firm

We have adopted, to this point in our review, a static contracting approach to the design of supply chain contracts. We have explored how contracts can elicit efficient incentives of firms along a supply chain in decisions such as inventory, pricing and sales effort. Our synthesis is limited thus far in two respects. First, firms in the theory make only short run decisions. We have set aside longer run decisions such as how much investment to undertake in assets specific to particular supply chain partners. Second, firms are assumed to act in their individual short-term interest given the terms of a contract. Each firm decides on effort, inventory or pricing without regard to the externalities imposed on other parties to the contract. Firms have no concerns about protecting the value of long run relationships with their contractual partners.

Supply chain management in reality involves a much richer set of decisions than can be captured within the static contracting framework. Firms are very often engaged in long run relationships with their upstream suppliers of inputs and downstream distributors of their products, and within those relationships invest in specific assets, to the benefit of both themselves and their relationship partners. In any long run relationship, it is in reality not rational to follow the dictates of selfish, short run profit maximization. A firm must take into account the benefits to supply chain partners with which it will be transacting and negotiating contracts in the future. The terms of future contract negotiations, as a matter of common business sense, depend on how beneficial ventures or transactions with the firm have been in the past.

Moreover, firms consider alternatives beyond entering simple contracts. These alternatives include both maintaining long term relationships and vertical integration into input suppliers or downstream distributors. As Coase [42] famously observed, the decision to undertake a transaction within a firm versus in a market is itself a fundamental economic question. The boundaries of the firm are in reality endogenous along a supply chain. Lafontaine and Slade [115] estimate that the value of transactions in U.S. firms is approximately equal to that in U.S. markets. The economics of supply chain management is not simply about contracts among firms whose identities are given exogenously.

This section of our synthesis outlines the most important themes within this broader view of supply chain relationships. Much of the material falls within the modern economic theory of the firm. What are the factors that determine whether a transaction happens within a firm or in a market? And what are the defining characteristics of a firm that distinguish intra-firm transactions from market transactions? This area of economics was relatively quiet after Coase's article, but since the 1970s has been very active.

A related issue is the management of relationships over the long run. Potential incentive distortions may be resolved or mitigated not only through short term contracts or vertical integration but, as noted above, through the establishment of a long term relationship. Long run relationships can share some of the same features as vertical integration in that parties are engaged in repeated transactions with the same set of individuals.

Beyond short term contracts, vertical integration and relational contracting a fourth factor influences incentives along a supply chain: the reputation of a firm even *outside* its current relationships. Any firm in a supply chain will value a reputation as an outlet with a low stock-out rate (a reputation not just among consumers but among upstream manufacturers as well). The firm will value a reputation as one that invests strongly in its relationships and in the quality of its products.

A comprehensive review of this broad area of supply chain management and long run relationships would take a book. The contributions of Coase, Klein, Williamson and others in this area are fundamental and cannot be comprehensively summarized in a short section. Instead of attempting such a summary, we capture the main ideas in this area through a set of simple models, adapting those offered in the literature.

12.1 Specific Investment, Appropriable Quasi-rents and Incomplete Contracts: The Incentive Problems

12.1.1 Introduction

The broader view of supply chain relationships incorporates two new incentive *problems*: distortions in investment in assets specific to a relationship as well as distortions in investment or effort in enhancing potential outside relationships. And the broader view incorporates three *solutions* to incentive problems, or mechanisms to at least mitigate incentive problems, beyond the design of contracts on which we have been focussed: vertical integration; relational contracting or self-enforcing agreements; and reputation. This subsection explores the problems. The following sections explore solutions.

Contracts are incomplete. Firms entering long term contracts cannot specify as part of the contract all future decisions, rights and obligations. Not only are short run decisions such as pricing, inventory and effort missing from contracts, but longer run decisions, such as investment by parties in the relationship, are also non-contractible. The incompleteness of contracts remains the fundamental source of incentive problems when investment is added to the set of decisions.

Incompleteness leads to two kinds of incentive problems. First, incomplete contracts cannot guarantee that investment in assets *specific* to a venture will be adequate. Assets are specific to the extent that they are more valuable within a relationship than outside. For example, an upstream supplier may invest in capital equipment specific to the needs of a downstream buyer. Fisher Body invested in stamping machines and dies to produce bodies of automobiles for General Motors; once constructed the molds had much less value (essentially only the value of the steel) outside the relationship than within.[1] Another example comes from franchising. When a franchisee invests in developing a local market by promotion activities or simply maintaining high product quality, it cannot always be assured (in the absence of specific guarantees) that the franchisor will not take advantage of its efforts by placing a second franchise in the market or even terminating the franchisee and selling its position to a new entity.[2]

The return on relationship-specific investment is referred to *quasi-rents*, and the vulnerability of quasi-rents to appropriation by the non-investing party is referred to as the *hold-up problem* (Klein et al. [104], hereafter KCA). For example, suppose that a coal mine invests one million dollars in development to supply a local energy utility and suppose that a return of 10 per cent would justify the investment. But once the capital is sunk, and prices for the coal are renegotiated with the utility, these prices would not reflect the sunk investment — and therefore might provide less than the required 10 per cent return.

The hold-up problem is not a matter of fairness. Parties to a contract rationally anticipate investment levels as well as the outcomes of future negotiations. The parties could offset any future anticipated hold-up with a transfer of wealth at the time of contracting. The problem is that the incentive to undertake efficient investment is compromised by the anticipation by the investing party that it will not appropriate the full return to its investment. Any business relationship in which an investing party does not capture the full return on its investment is one with inefficiently low investment levels.

The literature on transactions cost economics (KCA; [202, 203, 204, 135]) emphasized a second type of incentive distortion arising from

[1] The reference here is to a classic case of specific investment, the General Motors–Fisher Body relationship [104, 100, 102, 103].

[2] The franchising example raises the issue of a distinction between *relationship-specific* capital and *venture-specific* capital. When the franchise is broken up, the specific capital does not disappear (as it would in the Fisher-Body/GM example). Instead, the brand name capital remains with the other party, the franchisor. Because the capital survives the relationship, it is logically incorrect to refer to it as “relationship-specific” . It remains venture-specific, in the sense that it is valuable only within the specific venture or use.

the combination of incomplete contracts and the specificity of assets. Once specific investment is sunk, parties each have the incentive to undertake effort or investment to capture a higher *share* of quasi-rents. Masten [135] expresses this succinctly:

> Idiosyncratic assets, because of their specialized and durable nature, imply that parties to a transaction face only imperfect exchange alternatives for an extended period. The more specialized those assets, the larger will be the quasi-rents at stake over that period, and hence the greater the incentive for agents to attempt to influence the terms of trade through bargaining or other rent-seeking activities once the investments are in place.

Investment solely to increase the *share* of the pie is a pure inefficiency in terms of maximizing the *size* of the pie. Investment to increase total returns to a relationship is productive; investment to increase one party's share of the returns is unproductive. Williamson emphasizes the cost of constant haggling over the terms of exchange ex post. Another example of inefficient expenditure is investment in outside options, investment undertaken solely to increase the value of the assets in the paths not taken. An oilfield builds a pipeline to a distant set of refineries or an input supplier spends additional resources to render its assets less specific to a particular buyer, all in the interests of increasing the strength of its bargaining position in future negotiations rather than for efficient, productive purposes. An example from one of the author's experiences as a corporate director involves a firm supplying software for integration with a major downstream supplier of hardware, network firewalls. Investment in the compatibility of the software with other suppliers of firewalls (competitors of the main downstream purchaser) is valuable not just because of the prospect that other firms will purchase the software input, but because this investment strengthens the supplier's bargaining position in future negotiations with the main buyer. While valuable from the individual firm's point of view, investment in generality of the input so as to increase outside value is wasteful.

12.1.2 Unilateral Investment in Specific Assets

We capture in simple models the two kinds of incentive problems arising from contractual incompleteness and the need or specialized assets. These two problems are the distortion in incentives for specific investment, and the distortion in investment in outside options. To address the first problem consider the following model.[3] A pair of firms contract in period 1 engage in a project that will yield a return in period 2. The profits from the project, $\pi(I)$, depend entirely on the investment I by firm 1 in period 1, investment undertaken just after the contract. This investment is *productive*. The return $\pi(I)$ is assumed to be concave and differentiable. We suppress all expenditure by the second party: the participation of the second party (i.e., its signature) in both periods is the only required input by this party. The parties cannot contract today over the shares of the return $\pi(I)$. Instead the firms bargain at the beginning of period 2 over the shares of the profit to be distributed in exchange for continued participation by both parties. In the second period, the threat point of each firm in the bargaining is to walk away from the venture, earning 0 dollars.

The firms can transfer wealth freely at the time of contracting. If the firms could contract on investment, I, they would choose investment to maximize $\pi(I) - I$, which is total net return, investing to the point where $\pi'(I^*) = 1$. Since the firms cannot contract on investment or the shares of the return, the contract consists only of the exchange of a lump sum. The bargaining at the beginning of the second term over the share of return and continuation of the project is assumed to follow the Nash solution: equal shares, $\pi(I)/2$ to each firm. Since investment is entirely a sunk cost at the time of bargaining, it is irrelevant to the bargaining game in period 2 (except through the determination of the total gains to be shared).

Since the two firms can engage in wealth transfer at the time of the contract, but cannot commit to anything else, the lump sum transfer is the only element in the contract. Instead of investing at the first-best level, I^*, firm 1 undertakes an investment level in anticipation

[3] For a more elaborate model along these lines, see ref. [76].

of receiving only half of the return in the ex post bargaining. Firm 1 therefore chooses a level of investment, $\hat{I}$, satisfying $\pi'(\hat{I}) = 2$. From the concavity of $\pi(I)$, $\hat{I} < I^*$. The cost to the contracting pair of the non-contractibility of investment is $[\pi(I^*) - I^*] - [\pi(\hat{I}) - \hat{I}]$.

The distortion in the investment decision can, like any incentive distortion, be traced to an externality: the positive externality, $(1/2)\pi'(\hat{I})$, which is the marginal gain that firm 1's investment yields for firm 2. The failure of firm 1 to appropriate the full marginal value of its investment results in too little investment.

Can the distortion in incentives for investment in specific assets be resolved if the shares of profit are contractible (but the investment remains non-contractible)? In the case where only one party is investing in specific assets post-contract, the answer is yes. That party can be allocated the entire stream of returns minus a lump sum paid to the other party. As the holder of the residual claim, the investing party would have first-best incentives.

Bilateral investment in specific assets: It would be wrong to suppose that the asymmetry in investment requirements, with only one party investing, is the source of the incentive problem. In virtually any business relationship, *both* parties invest in capital to enhance the returns from the relationship. If both parties are investing then because there is only one residual claim to allocate, each party must receive less than the full marginal return to its investment. This is the Holmstrom [84] budget constraint that we discussed in Section 6.

In fact, guaranteed returns may yield close to the same result, in the case of two investing parties, as simply Nash ex post bargaining to split the returns. This is because in a bilateral incentive problem, optimal contracted shares are likely to be very close to equal sharing. To see this, consider a venture yielding a profit $F(I_1, I_2)$ as a function of the investment levels of the two parties. No uncertainty is involved, and F is concave and twice-differentiable, with first derivatives F_i and second derivatives F_{ii}, $i \in \{1,2\}$. If the parties can exchange transfers ex post, then without loss of generality the optimal sharing rule can be expressed as shares λ and $(1 - \lambda)$ of returns allocated to the parties 1 and 2. The allocation of shares that maximize the total return, subject to the incentive compatibility constraints is given by the following problem.

$$\max_{\lambda, I_1, I_2} F(I_1, I_2) - I_1 - I_2$$

subject to

$$I_1 = \arg\max_{\hat{I}_1} \lambda F(\hat{I}_1, I_2) - \hat{I}_1$$

$$I_2 = \arg\max_{\hat{I}_2} (1 - \lambda) F(I_1, \hat{I}_2) - \hat{I}_2$$

Replacing the incentive compatibility conditions with the respective first-order conditions and solving the maximization problem yields the following condition for the ratio of optimal shares [144].

$$\frac{\lambda}{(1-\lambda)} = \left(\frac{F_{22}}{F_{11}}\right)^{1/4} \tag{12.1}$$

The result requires no assumptions whatsoever on the form of F.[4] Even with moderate asymmetry in the technology, as reflected by a ratio of F_{22}/F_{11} that is substantially different from 1, the 1/4 power transformation of right-hand side brings the right-hand side close to 1, i.e., the share close to equality.[5] Optimal sharing rules, across a wide range of technology, are close to 50–50.

12.1.3 Incentives for Non-productive Investment

The transactions cost economics literature focusses, as we discuss above, on the inefficient expenditure by a party to increase its share of

[4] The first-order conditions on the two incentive compatibility constraints are

$$\begin{aligned} \lambda F_1 - 1 &= 0 \\ (1-\lambda) F_2 - 1 &= 0 \end{aligned}$$

The Lagrangian for the optimal contract problem becomes $F(I_1, I_2) - I_1 - I_2 + \mu_1(\lambda F_1 - 1) + \mu_2[(1-\lambda)F_1 - 1]$. The first-order conditions for this Lagrangian can be manipulated to eliminate the shadow prices μ_1 and μ_2, leaving Equation (12.1).

[5] For example, even if the ratio is as large as 5.6, the optimal shares are within the 40–60 to 60–40 band. Neary and Winter [144] apply this result to explain a puzzle in sharecropping contracts: why shares in these contracts rarely deviate far from 50–50 over centuries and wide geographical areas.

quasi-rents. Masten [135] and Gibbons [70] refer to this expenditure as *rent-seeking*, because of the parallel with the analysis of rent-seeking in the political economy literature [192]. Rent-seeking is any expenditure by an individual to increase the individual's share of resources, with a zero or negative impact on the total resources to be shared. This includes investment to "keep options open" by rendering assets of higher value in alternative relationships. This type of investment plays the role of strengthening a party's bargaining position in future negotiations. As one would expect it can be excessive because it imposes a negative externality on the contractual partner.

To see this, consider the following model. An upstream and downstream agent enter a contract that will yield a gross payoff π in period 2. The parties need only participate in both periods 1 and 2 to earn π in period 2. With no investment, the threat point of each party is to walk away, earning 0. The first party, however, can invest an amount a that will raise the value of its threat point, or alternative option, to $u(a)$. Given investment a, negotiations over sharing the profit π take place. The Nash bargaining solution equal shares of quasi-rents will split the gains $\pi - u(a)$ between the parties. (The investment a is not deducted from the returns being bargained over since it is sunk at the time of bargaining.) This results in shares $[u(a) + \frac{1}{2}[\pi - u(a), \frac{1}{2}[\pi - u(a)]]$. Anticipating the outcome of the negotiation, the first party chooses a to maximize $u(a) + \frac{1}{2}[\pi - u(a)] - a = \frac{1}{2}[\pi + u(a)] - a$. This yields an equilibrium investment $\hat{a}$ solving $u'(\hat{a}) = 2$. The total profits shared by the parties are reduced by the nonproductive investment $\hat{a}$. The source of the distortion is a negative externality: with a marginal increase in a, the non-investing party's return is reduced by $\frac{1}{2}u'(a)$.

The effort or investment that we have labeled as a can refer to any action on the part of an investing party that yields private gain but is collectively a deadweight loss. Investment in outside options is just one example. As noted above, Williamson [201, 203, 204] and Klein et al. [104] identify another: the costs of haggling over rights and obligations within the relationship. As our simple model demonstrates haggling or other "inefficient" efforts to extract private gain in a relationship can be second-best efficient if these efforts encourage greater investment in specific assets.

12.1.4 Investment in Both Productive and Non-productive Assets: The Efficiency Role of Rent-seeking

When both productive and non-productive investment is undertaken, we suggest the seemingly wasteful "non-productive" investment in outside options can have a type of second-best efficiency role. The investment can actually be beneficial because it induces greater investment in specific assets. Consider the following modification of the model. Only the first party invests, but now this party undertakes two kinds of investment. *Specific investment* I determines the payoff to be shared ex post via $\pi(I)$. And the party can invest an amount a per unit I that increases the value of the investment to outside parties. The function $G(a) \cdot \pi(I)$ represents the value of the investment to outside parties.[6] G is smooth, $G(0) = 0$ and $G(a)$ is bounded above by 1. The investment aI is rent-seeking, in the sense that it increases the investor's share of the pie but reduces the total size of the pie. Let the outside parties be competitive, so that all potential returns from using the investment with outside parties accrue to the firm undertaking the investment.

In this model, given investments I and a, the ex post bargaining takes place with a profit of $\pi(I)$ available internally should both parties continue to participate in the project. A profit $G(a) \cdot \pi(I)$ would be available to the investing party should the relationship break down. The outcome of bargaining is that the investing party captures a value of equal to the outside value of its investment (its threat point in bargaining), $G(a) \cdot \pi(I)$, plus half of the surplus from internal use of the assets, which is $\pi(I) - G(a) \cdot \pi(I)$. Subtracting the investment costs yields the payoff to the investing party[7]:

$$\frac{1}{2}[\pi(I) + G(a) \cdot \pi(I)] - I(1 + a)$$

[6] The multiplicative form of $G(a)$ and $\pi(I)$ is essential. The investment in outside options increases the outside value *per unit* of the specific investment.

[7] Note our assumption that there is a type of economies of scale in the investment a: it is of the same cost to raise the outside value of the asset by a given percentage when $I = 1000$ as when $I = 10$. For capital such as software, this is a sensible assumption. For physical capital such as machines, a more natural formulation would be that a profit of $G(a)\pi(I)$ requires an investment of $I + aI$. The essential points are unchanged when this alternative formulation is adopted.

The first order conditions in I and a are

$$\begin{aligned} \frac{1}{2}\pi'(I)[1 + G(a)] - (1 + a) = 0 \\ \frac{1}{2}G'(a)\pi(I) - 1 = 0 \end{aligned} \tag{12.2}$$

As a benchmark, note that the first-best efficient levels are $a^* = 0$ and I^* satisfy $\pi'(I^*) = 1$. The first-order conditions (12.2), compared to the first-best, show that the equilibrium investment $\hat{I}$ in specific investment is too low. And the investment in the outside option — being pure waste from the point of view of efficiency — is excessive.

If $G(a)$ is close to zero even for high levels of a, then the model represents essentially the case where a is unavailable and specific investment is too low. Suppose, however, that $G(a)$ allows the creation of an efficient outside option at very low-cost. That is, $G(a)$ is close to 1 even for small a. Then *the opportunity to invest in outside options improves efficiency.* This is because the investment a allows the investing party to largely recoup the return on its investment. When $G(a)$ is close to 1 at even small levels of a, the first-order condition determining π in (12.2) is approximately the first-best condition. The amount a is an investment in the *generality* of the asset in this example. When $G(a) = 1$, the investment I is perfectly general in the sense that the marginal value created for the asset does not depend on whether the asset is used internally or externally. General investment carries no incentive problems; as we have emphasized it is only specific investment that gives rise to the hold-up problem. When a renders $G(a) = 1$ for small a, then the investment a has, at low cost, transformed the decision on I from one of specific investment to one of general investment. As Gibbons [70] expresses the point using a similar example, which he captures nicely in the expression "hold-up may be your friend." Segal and Whinston [170] offer a comprehensive, high-level review of the property rights theory.

To summarize, investment by an individual in a contractual relationship can vary among three types: specific investment, investment in outside alternatives and general investment which adds value equally inside and outside the relationship. With incompleteness in contracts, specific investment is too low; investment in outside alternatives is excessive; and general investment, like Goldilocks' porridge, is just right.

12.2 The First Solution: Vertical Integration

12.2.1 Introduction

One potential resolution to the two distortions resulting from the specificity of assets is vertical integration. Throughout this monograph we have used as a benchmark a mythical centralized firm with no incentive distortions. All decisions in this centralized firm are taken in the collective interest. The neoclassical economic literature on the incentives for vertical integration [199] has adopted the same ideal, as a means of focussing on the benefits of vertical integration rather than the costs. In our search for the simplest contractual resolutions of incentive problems along supply chains, we have recognized the costs of vertical integration as well as its complete infeasibility in some cases, but this recognition has been only implicit.

If vertical integration led in reality to the centralized firm that we have used as a benchmark, then vertical integration would indeed be an easy solution to the hold-up problem. Integration would in fact solve any of the incentive problems that we have examined in this monograph.

In reality, of course, vertical integration is not a panacea. Internalizing transactions within a firm presents its own incentive and transaction problems. The vertical integration or make-or-buy decision, as Coase [42] famously pointed out in 1937, must trade off the costs of transacting in the market with the costs of transacting inside a firm. The costs involved in market transactions, including the hold-up problem, have been discussed above. Choosing transactions within a firm involves exposure to a different, if similar, set of incentive problems. Much of the contracting literature can be interpreted as the design of optimal contracting within a firm, between an employer and an employee or between owners of a firm and the CEO. Incentives are far from perfect within a firm. Contracts and other incentive mechanisms mitigate intra-firm inefficiencies but not perfectly.

While this monograph has focussed on decentralized solutions to incentive problems, the make-or-buy decision is such a critical part of supply chain management, it is well worth taking a closer look at the costs and benefits of complete internalization. To undertake this exercise, let us suppose that a task, such as the production of an input, is transferred from the market to within the firm. Consider the two possible impacts on efficiency of this integration: incentives to undertake the task itself, and the impact of internalization on the management of the other, existing assets.

Regarding the first impact, taking a task such as the production of an input inside the firm means that the market incentives to undertake the task must be replaced by the incentives within the firm. The manager of an input-providing firm that becomes the division of a downstream firm is no longer incented by market prices but by the threat of termination upon poor performance, the potential for promotion and higher wages upon strong performance. These incentives involve costly monitoring and, as basic incentive theory demonstrates, do not necessarily lead to first-best outcomes. The incentive distortions and monitoring costs, which together Jensen and Meckling [93] refer to as *agency costs*, may well be greater than the transactions costs including costs of incentive distortions associated with the market.

Consider next the second component of integration efficiencies, the impact of the internalization of the additional task on the general agency costs of managing the entire firm. Suppose that the additional return on assets resulting from the internalization of production is uncertain — and, for simplicity, that the additional uncertainty is independent of the uncertain return on existing assets. The issue reduces to the impact of the addition of white noise on the total agency costs within a firm.

We collapse the agency problems within a firm to a single principal–agent problem, with the CEO as the agent. (This has been the standard approach since Jensen and Meckling.) Consider the addition of white noise to the returns in this principal–agent problem. A CEO is disciplined to exert effort by four mechanisms, apart from any inherent desire to do a good job:

- A manager is incented by *performance-based pay*, as in the basic principal–agent model (e.g., [83, 176]).
- A manager is *monitored* by shareholders, through the board of directors, and is vulnerable to being fired upon underperformance. Shareholders, at the top of the corporate hierarchy, are the residual claimants and are motivated as such to be the ultimate monitors [4].
- Managers are disciplined by the *market for corporate control*: a poorly managed firm may be bought by a firm that values the firm's assets because it will apply more highly skilled management talent to the assets. The poorly performing manager will then be out of a job. As Manne [129] states "The lower the stock price, relative to what it could be with more efficient management, the more attractive the take-over becomes to those who believe that they can manage the company more efficiently. And the potential return from the successful takeover and revitalization of a poorly run company can be enormous." See also ref. [167].
- A manager always faces the likelihood of being back on the labor market in the future and therefore has the incentive to maintain a strong reputation. Reputation is enhanced by strong performance of the firm being managed. These *career concerns* enhance the manager's incentive to perform well [85].

12.2.2 An Offsetting Cost of Integration: An Additional Task Clutters Incentives

Assume that the new task requires no additional oversight by the manager. (We thus set aside any possible costs of "stretching" managerial input to a larger set of tasks.) The impact of the addition of the task on the four sources of discipline on the manager therefore reduces to the impact of the addition of white noise to each of these sources. It is straightforward to demonstrate, with simple models, that each of the four disciplining factors are compromised by the addition of the noise. The basic principal–agent model solves to give a lower value for the principal. The ability of shareholders to monitor the manager is compromised by the addition of uncertainty, as is the inference by the market for corporate control the CEO ability from the company's performance. The power of reputation to discipline management, which operates through the ability of future labor markets to infer the characteristics of management will also be reduced by the noise added to firm returns by the internalization of the new tasks. All four mechanisms for strengthening managerial incentives to perform are reduced in power by the addition of noise to the output of the firm, because all four mechanisms involve inference of the performance of the manager from the observed output of the firm. Bringing more tasks inside the firm adds to the noise and renders the output less powerful as a signal of managerial talent. This is a fundamental cost of integration.

In terms of the formal development of the principle, the effect of noise in the principal–agent model is well known. The effect on monitoring by shareholders or inference by the market for capital control has not been developed formally, but is clear. We outline here the effect on reputation in a simple version of the Holmstrom [85] career concerns model. This theory will also be useful in our discussion, in a later section, of a *firm's* reputation in a supply chain.

Formal development of the dilution effect for reputational discipline: The simple version of the Holmstrom model (see ref. [57]) considers a manager earning profits on behalf of shareholders in each of two periods. In each period, the profits y earned by the manager depend additively upon three factors: the managers talent or type, θ, the manager's current effort, a, and noise: $y_t = \theta_t + a_t + \varepsilon_t$, $t = 1, 2$. The noise terms ε_t are independent and identically distributed each period according to a normal distribution with mean 0 and variance σ^2. The distribution of talent across the potential pool of managers is normally distributed with mean θ_T and variance σ_T^2. The manager participates in a competitive labor market in each of two periods and discounts the second period wage at a discount rate δ. Talent is unknown to everyone, including the manager, and effort is known only to the manager. The cost of effort, $c(a)$ is convex. Profits are observed at the end of each period but cannot be described (or contracted upon) ex ante. As a result, the contract for the manager is simply a wage, w_t in each period. In the second period, the market rationally anticipates an effort of zero, $a_2 = 0$, and therefore pays the manager a wage equal to her expected talent, as inferred from the observation of profit at the end of the first period. It is this inference that gives the manager the incentive to exert effort in the first period: a higher first-period output increases the market's expectation of the manager's talent (i.e., the manager's reputation) at the beginning of the second period. This expectation is the manager's wage in the second period, w_2. An equilibrium level of effort, a^*, in the first period is a level of effort such that if the market anticipates a^* from the manager it is in fact in the manager's interest to exert a^*. An equilibrium wage in the second period is $w_2^* = E(\theta|(y_1, a^*))$. The manager chooses a in equilibrium to maximize:

$$w_1 - c(a) + \delta E[E(\theta|(y_1, a^*))] \tag{12.3}$$

In this expression, the inner expectation is with respect to talent θ and the outer expectation is with respect to performance, y_1. The first-order condition characterizing a^* is [57]:

$$[c'(a^*) = \delta\sigma_T^2/(\sigma_T^2 + \sigma^2)] \tag{12.4}$$

Thus incentives depend upon the signal-to-noise ratio σ_T^2/σ^2.[8]

Simple comparative statics on Equation (12.4) shows that a^* is decreasing in σ^2. Thus, the addition of noise (in our context, through

[8] The following is a derivation of this equilibrium condition within the simplified framework of our model (adapted from ref. [57]). Let $f(\theta, y|a)$ denote the joint distribution of θ and y given effort level a and $\hat{f}(y|a) = \int f(\theta, y|a)d\theta$ be the marginal distribution of y. In equilibrium,

$$w_2(y_1) = E_\theta(\theta|y_1, a^*) = \int \theta \frac{f(\theta, y|a^*)}{\hat{f}(y|a^*)} d\theta \tag{12.5}$$

The agent's objective is $\max w_1 + \delta E_{y_1} w_2(y_1) - c(a)$. Omitting w_1 and substituting in (12.5) yields as the agent's objective

$$\max \delta E_{y_1}\{E_\theta(\theta|y_1, a^*)\} - c(a) = \int \left(\int \theta \frac{f(\theta, y|a^*)}{\hat{f}(y|a^*)} d\theta \right) \hat{f}(y|a)dy - c(a) \tag{12.6}$$

The first-order condition for this maximization is

$$\int\int \theta f(\theta, y|a^*) \frac{\hat{f}_a(y|a)}{\hat{f}(y|a^*)} dy d\theta - c'(a) = 0$$

the internalization of a task inside the firm) will diminish the power of reputational forces to elicit high effort on the part of the manager. Prior to internalization of the task within a large integrated firm, the task is managed within a smaller, more focussed firm supplying the input. It is reasonable to assume that the non-integrated organization, in a more focussed production arrangement, involves more accurate inference of managerial effort and therefore higher incentives for the task. Overall, the integration of a task within a firm therefore reduces the efficiency of managerial effort.

In summary, integrating a transaction in response to distortions in market allocation does not magically allow a first-best outcome. The impact on agency costs of internalization must be balanced against the costs of transacting in the market.

In considering the incentives of only a single agent, this sketch of the costs of vertical integration is but a starting point. The transactions-cost literature in economics (see especially Williamson [202, 204]) discusses more fully the costs and benefits of vertical integration.

12.2.3 The Property Rights Theory

The property rights theory (PRT) of Grossman and Hart [75] is the first formal model to incorporate both the costs and benefits of integration. This theory is the most prominent recent contribution to the theory of the firm. The focus of the PRT is on integration as an allocation

We can use the fact that $E(\hat{f}_a/\hat{f}) = 0$, to rewrite this as

$$\operatorname{cov}\left(\theta, \frac{\hat{f}_a}{\hat{f}}\right) - c'(a) = 0 \tag{12.7}$$

with $y_t = \theta_t + a_t + \varepsilon_t$, the conditional distribution of y_1 given θ and a_1 is normal with mean $\theta + a_1$ and variance $\sigma_\varepsilon^2 + \sigma_\theta^2$. Therefore, letting $\bar{\theta}$ be the mean of θ we have that $\hat{f}(y|a)$ is proportional to

$$\exp - \left[\frac{(y - \bar{\theta} - a)^2}{2(\sigma_\varepsilon^2 + \sigma_\theta^2)}\right]$$

Substituting in $y_t = \theta_t + a_t + \varepsilon_t$ and differentiating w.r.t. a yields $\frac{\hat{f}_a}{\hat{f}} = \frac{(\theta - \bar{\theta}) + \varepsilon}{\sigma_\varepsilon^2 + \sigma_\theta^2}$ from which $\operatorname{cov}\big(\theta, \frac{\hat{f}_a}{\hat{f}}\big) = \frac{\sigma_\theta^2}{\sigma_\varepsilon^2 + \sigma_\theta^2}$. Substituting this into (4) yields the equilibrium condition.

of ownership rights. Hart [79] offers the following summary in a recent review:

> Our approach yields a theory of asset ownership and firm boundaries. In our formal model, parties renegotiate their incomplete contract once an uncontracted-for contingency arises: this renegotiation occurs under symmetric information (both observe the contingency). The parties therefore reach an ex post efficient outcome. However, the division of surplus depends on the allocation of asset ownership. Suppose you and I can both undertake non-contractible investments that make us more productive as long as we have access to asset A. If I own A, then I can use my residual control rights to deny you access ("hold you up"), thereby reducing your share of the ex post surplus. Under reasonable assumptions, this will reduce your incentive to invest. On the other hand, if you own asset A, then you can use your residual control rights to hold me up, and knowing this I will invest less. Under plausible assumptions, it can be shown that ownership of asset A should be allocated to the party whose investment is more important.

We discuss the PRT in a setting where at the outset only one party, an upstream supplier, owns an asset that can be used in a venture with a downstream buyer. Both the buyer and the seller can undertake investment to increase value of the asset within the relationship and investment to increase the potential value of the asset outside the relationship. We consider these two types of investment separately (as in our previous discussion in this section), and then consider the interaction between the two types of investment. The PRT game starts with the allocation of *control rights* to the assets. Control or ownership rights here refer to the right to use the asset in an alternative venture, with a different partner. Once control rights are allocated, investment decisions are undertaken. These investment decisions cannot be specified in the contract between the two parties. Some time after the investment decisions have been made, the two parties negotiate on the use of the assets and the split of the returns generated by the assets. This ex post negotiation takes place under full information by both parties and therefore leads to the efficient decision, i.e., the decision that maximizes the combined profits of the firms. While realized profits are observable, these profits are not contractible, because the products, prices and profits cannot be described in an ex ante contract.

In short, the investment levels and profit shares are not contractable but the allocation of control rights is contractible under the PRT. The shares to the two parties are determined by the allocation of control rights: the party allocated the control rights in the contract is protected against hold-up by the other party and will therefore, according to Hart's summary above, will be guaranteed a larger share of the return from investing in the asset than if the other party has control.[9]

12.2.3.1 Specific Investment by the Supplier

Suppose that the supplier initially owns an asset essential to the profitability of the relationship. The supplier invests i_S with the effect of increasing the surplus generated from the asset in the relationship. The supplier is uniquely capable of investing to improve returns from the asset. The asset has *potential* value in uses beyond the relationship, such as the use in exchange with other buyers. But the supplier's investment is entirely specific to the relationship in that it has no impact on the outside value of the asset. The parties can contract on the *ownership* of the asset. If the supplier retains ownership of the asset — the right to dictate in the future any decisions on use of the asset — we say that

[9] The model is most easily interpretable in terms of investment in human capital, as Hart [82] notes. But the model applies potentially to incentives for any type of non-contractable investment. Consider, for example, an individual supplying software development skills and investment who partners with another individual supplying skills and effort in marketing a software product. Both types of investment are essential for success in the software industry. Yet neither type of investment can, realistically, be written into a contract between the two parties in their venture. The exact nature of a software product and the lines of code necessary for development of a successful product are impossible to specify exactly. And the brand name capital, awareness and marketing channels that are built up within a software firm are equally impossible to write down. The shares of profits may be somewhat easier to contract upon in reality, via shares of equity in a joint firm, but often both parties are involved in many other closely related activities in a way that makes it impossible to attribute the precise costs and revenues to the product, and hence to write down shares of profits in a way that would be enforceable by courts or another third party.

the buyer and seller are non-integrated. If the buyer is allocated ownership by the contract, then upstream vertical integration has taken place. Thus, integration in the PRT refers entirely to the allocation of control rights. We denote the ownership as $A = \mathbf{0}$ for non-integration and $A = \mathbf{1}$ for upstream integration.

The integration decision is taken at the outset of the game, then investment decisions are taken. Once the investment is made, a set of opportunities opens up as to the possible decisions that can be taken to deploy the assets. The assets could be used in the relationship (to produce an input that will be used by the downstream firm). Alternatively, the assets could be deployed in another use, such as in production elsewhere. Investment by the seller determines the internal surplus generated via $\pi(i_S)$.

The alternative returns to the buyer and the seller each depend upon the ownership decision. We let $w_B(A)$ and $w_S(A)$ represent the parties' returns from alternative relationships, and assume that $w_B(1) > w_B(0)$ and $w_S(0) > w_S(1)$. For example, if the outside option is simply to sell the assets to an alternative user at a price w, then $[w_B(0);\, w_S(0)] = (0, w)$ and $[w_B(1);\, w_S(1)] = (w, 0)$. The alternative returns are invariant to investment.

Once investments have been undertaken, the parties negotiate the terms of exchange for the input, which divides up the surplus $\pi(i_S)$. Negotiations, we assume, will allocate shares of surplus to the two parties according to the Nash bargaining solution. That is, the total surplus over outside alternatives is divided equally. Letting the returns to the buyer and the seller be $\pi_B(i_S; A)$ and $\pi_S(i_S; A)$ we have $\pi_B(i_S; A) = w_B(A) + \frac{1}{2}\{\pi(i_S) - [w_B(A) + w_S(A)]\}$ and $\pi_S(i_S; A) = w_S(A) + \frac{1}{2}\{\pi(i_S) - [w_B(A) + w_S(A)]\}$.

The supplier's return on investment is given by $\pi_S(i_S; a)$ and the *increase* in this return, from having control over the assets as opposed to releasing control to the buyer via vertical integration, is (after simplification):

$$\Delta = \pi_S(i_S; 0) - \pi_S(i_S; 1) = \frac{1}{2}\{[w_S(0) - w_B(0)] - [w_S(1) - w_B(1)]\}$$

Note that Δ *is completely independent of* i_S. The integration decision has no impact at all on the investment decision. The change in control of the assets gives the seller greater rents, but does not change the marginal impact of the seller's investment on rents. In this simplest case, involving only specific investment and investment only by the supplier, the vertical integration decision is irrelevant.[10]

12.2.3.2 Rent-seeking Investment by the Supplier

We move now to the case where investment (still by the supplier only) affects the value of the use in outside alternatives. This type of investment we referred to above as rent-seeking investment because it is non-productive in terms of increasing surplus in the use to which the asset will be put; the investment is undertaken only to attract a higher share to the investing party in future negotiations.[11]

In the case of rent-seeking investment by the supplier, vertical integration becomes relevant. To demonstrate this, suppose that the supplier's investment, a_S, has *only* the effect of increasing the supplier's outside alternative:

The rent-seeking investment has value to the supplier only if the supplier controls the asset since only then can the supplier threaten to use the asset outside. We can thus write

$$w_S(a_s; A) \begin{cases} \bar{w}_S \colon & A = 1, \\ \tilde{w}_S(a_S) \colon & A = 0 \end{cases}$$

with $\partial \tilde{w}_S / \partial a_S > 0$. In the case of rent-seeking investment, the impact of vertical integration on investment incentives is captured by the difference in payoffs to the supplier under the two integration alternatives. Retaining the same notation as above this difference is,

$$\Delta = \pi_S(a_S; 1) - \pi_S(a_S; 0) = -\frac{1}{2}\{[\tilde{w}_S(a_S) - w_B(0)] + [\bar{w}_S - w_B(1)]\} \tag{12.8}$$

The equilibrium supplier investment is 0 in the case of integration and is positive, given by the solution to $\frac{1}{2}\partial \tilde{w}_S / \partial a_S - 1 = 0$, in the case

[10] See Whinston (2003) for a discussion of parameter values for which the integration decision is irrelevant in a related model.

[11] Note that in a more complex model, with uncertainty, investment in outside alternatives could have a (productive) *option value* reflecting the chance that outside options turn out to be more valuable. We set aside any productive value of outside investment.

of non-integration. Since the efficient level of non-productive investment is 0, integration is optimal. Upstream integration eliminates the supplier's incentive to enhance the value of the asset in outside alternatives since the supplier cannot take the asset upon leaving the relationship. The decision to vertically integrate, in this case of non-productive investment in outside alternatives, is optimal because it *minimizes* the supplier's investment incentives.

12.2.3.3 Both Productive and Rent-seeking Investment by the Supplier

We move now to consider, within the PRT, the feature that we have introduced into the discussion of vertical integration and investment incentives: the *interaction* of incentives for specific investment and rent-seeking investment. Firms in relationships pursue investments that both add to the value of assets in the relationship and add to the value of assets in outside alternatives. The investment in flexibility, generality or non-specificity of the assets in reality serves two roles. As in our models above, the investment serves to enhance the firm's bargaining power in negotiations within the relationship. But investment in flexibility or non-specificity also adds to the value of the firm should the current relationship fail. In this sense, investment in generality of assets can be a kind of insurance. We continue to focus entirely on the first of these sources of incentives in considering the case where a firm invests in both productive and rent-seeking investment.

If the impacts of productive investment, i_S, and rent-seeking investment, a_S, are *independent*, then these impacts of the investments are captured by $\pi(i_S)$ and $\tilde{w}_S(a_S)$ as in the two cases above. The impact of vertical integration on investment incentives, as captured by the difference, Δ, in payoffs to the supplier in the case of vertical integration and non-integration is unchanged from (12.8). Productive investment is unaffected by vertical integration and rent-seeking investment is eliminated by vertical integration. Integration is optimal because its only effect on investment decisions is to eliminate wasteful investment.

It is realistic, however, to allow the two kinds of investments to interact in the sense that a_S serves to add *proportionate* value to the asset when the asset is transferred to an outside use. This is captured in an assumption that the outside value of the asset can be expressed as $\tilde{w}_S(i_S, a_S) = G(a_S)\pi(i_S)$; here a_S determines the proportion of the inside value that can be captured in an outside use.[12] $G(a_S) \leq 1$ is a measure of the generality of the asset. In this case, the payoffs to the supplier to the supplier under non-integration and vertical integration are the following

$$\begin{aligned}\pi_S(i_S, a_S; 0) &= \tilde{w}_S(i_S, a_S) + \frac{1}{2}\{\pi(i_S) - [w_B(0) + \tilde{w}_S(i_S, a_S)]\}\\ &= \frac{1}{2}[\pi(i_S) - w_B(0) + G(a_S)\pi(i_S)]\\ \pi_S(i_S, a_S; 1) &= \frac{1}{2}[\pi(i_S) - w_B(1) + \bar{w}_S]\end{aligned}$$

and the impact of integration on the supplier's payoff is now given by

$$\Delta = \frac{1}{2}[-w_B(1) + \bar{w}_S + w_B(0) - G(a_S)\pi(i_S)]$$

The changes in incentives are reflected in $\partial\Delta/\partial i_S$ and $\partial\Delta/\partial a_S$, which are given by

$$\begin{aligned}\frac{\partial\Delta}{\partial i_S} &= -\frac{1}{2}G(a_S)\pi'(i_S)\\ \frac{\partial\Delta}{\partial a_S} &= -\frac{1}{2}G'(a_S)\pi(i_S)\end{aligned}$$

The impact of integration is now completely dependent on the cost of achieving asset generality. If this cost is low, i.e., if $G(a_S)$ is very near 1 for even small levels of a_S, then the choice of productive investment I_S in the non-integration case determined essentially by $\frac{1}{2}[\pi'(i_S) + G(a_S)\pi'(i_S)] = 0$ or $\frac{1}{2}[\pi'(i_S) + 1 \cdot \pi'(i_S)] = 0$ which is equivalent to $\pi'(i_S) = 1$, yielding the first-best incentives. And little of the "wasteful" investment a_S need be expended to achieve this effect. The optimal organizational form is to remain non-integrated. This principle is consistent with the rent-seeking literature: the message is that vertical integration is unnecessary when assets can easily be rendered non-specific. On the other hand if $G(a_S)$ remains very near 0 for even large

[12] In other words, under this formulation investment in a_S allows investment in I_S to add value to both inside and outside ventures. Thus a_S is an investment in asset generality.

levels of a_S, then the opportunity to invest in a_S plays only a small role in enhancing the incentives for productive investment. The optimal organizational form is vertical integration, with the buyer retaining control over the asset and preventing the supplier from capturing any gains from rent-seeking.

12.2.3.4 Summary of the Property Rights Theory of Vertical Integration

Our brief review of the PRT has established several points. With only specific or productive investment at issue, the assignment of control rights is irrelevant. The incentives for investment are unaffected by the integration decision. Only when we allow for rent-seeking investment, or investment in outside options, does the allocation of control rights matter. In this case, the integration decision hinges on the *interaction* of productive investment and investment in outside options. (This interaction is captured in our function G.) Where the return on investment in outside options is independent of the internal value of the asset, then we are in a pure rent-seeking world. The allocation of control rights is directed towards minimizing the investment (rent-seeking) incentives. When the non-productive investment increases value of outside ventures in proportion to inside value, and at low cost, then we are in a world of low-cost investment in asset generality. With any asymmetry between the buyer and the seller in the extent to which non-productive investment adds to asset generality at low cost, the integration decision is one that allocates control rights to the party for which non-productive investment is more in the nature of an investment in asset generality.

We began this section of the monograph by discussing two potential incentive problems that arise within the longer run perspective of supply chain management: the incentive to under-invest in specific assets and the incentive to over-invest in rent-seeking. These two incentive problems with market transactions are at the heart of the explanation of the benefits of vertical integration within the Coase–Williamson–Klein literature. In that literature, vertical integration is profitable if it enhances the incentives for specific investment and dampens the incentives for rent-seeking.

What is the relationship between the Coase–Williamson–Klein literature and the PRT? As Gibbons [70] discusses, this question elicits a wide variety of responses — from the view that the earlier literature has finally been formalized by Grossman and Hart [75] to the view that these are alternative perspectives. Our set of simple models suggests three points. First, with respect to rent-seeking behavior, the PRT indeed formalizes the idea that this wasteful investment can be reduced by vertical integration. Second, with respect to specific investment, the earlier literature (especially Grout [76]) has a simple theory of why vertical integration solves the inadequacy of incentives. Once a firm is integrated, the firm itself captures the full return from specific investment (instead of 1/2 the return in the simplest model). The problem of appropriability of quasi-rents largely disappears. In the simplest PRT model, in contrast, vertical integration has no impact at all on incentives for purely specific investment. The seller, in our example, earns higher profits when he retains control of the asset (non-integration) but his *marginal* return from additional investment is unaffected. Third, the analysis by Williamson in particular allows specific investment to be contractible. The problem of excessive rent-seeking exists even if specific investment is contracted. The Williamson analysis applies, for example, to the case where specific investment is in easily-measurable physical assets. The PRT model, in contrast, assumes that all investment is non-contractible and in particular does not apply to contractible investment in physical assets.[13]

[13] The standard PRT model has both parties owning assets at the outset, but has only one type of investment (e.g., [27, p. 498–508]). The failure to distinguish productive (inside) investment from nonproductive (outside) investment in our view, leads to some confusion. For example, Bolton and Dewatripont, [27, p. 506], note that in their framework, allocation of all assets to the upstream supplier increases the supplier's incentive to invest to the extent (and only to the extent) that the investment increases the value of the assets to the supplier in an outside relationship — to the extent, that is, that the supplier's threat point in the negotiation with the buyer is increased. But this investment is then referred to as "specific investment." And the allocation of all assets to the supplier is described as one way to enhance the supplier's incentive to invest. We suggest that if the supplier's investment mainly increases the value of the assets in another relationship, the ownership structure will be chosen to minimize, not maximize, incentives to invest.

12.3 Relational Contracting

Most supply chain contracting models assume that contractual obligations can be enforced by a court. Litigation is expensive and disruptive, however, and disputes can be resolved in other ways. Macaulay [126] and others (see ref. [88]) have noted that trading partners are often reluctant, or unable, to use courts to resolve disputes. Litigation about contract terms usually marks the end of the trading relationship, and trading partners do not use the courts to help resolve disputes that arise in any normal trading relationship. Baker et al. [14] note that supply chains "often involve long-run, hand-in-glove supplier relationships through which the parties reach accommodations when unforeseen or uncontracted-for events occur."

A recent and growing literature on *relational contracts* seeks to better understand implicit agreements between firms.[14] A relational contract is an implicit agreement sustained by the value of future returns to the relationship. "Implicit" here means that the agreement need not be written down as an explicit, court-enforced agreements. A relational contract must be *self-enforcing*, and the trading partners adhere to the agreed upon obligations because it is in their best interest to do so. Deviations from agreements are discouraged by the threat of receiving less beneficial terms of trade in the future. This self-interested adherence to the agreement is achieved because players value the future revenue stream from continuing the relationship.

[14] While the word "contract" is usually defined to mean a "legally enforceable promise," Eisenberg [60] notes that in everyday usage, it is often stripped of its legal connotation and can refer to any bargain struck by two (or more) parties. In Eisenberg's definition, a "contract" (or a "bargain") is "an exchange in which each party views his performance as the price of the other party's performance." Much of the existing supply chain literature on contracts uses the legal interpretation of the word. The relational contracting literature generally follows Eisenberg's definition and uses the word "contract" to refer to an agreed upon exchange between parties which need not be legally binding. Another way to interpret this is to assume the compliance with the terms of a relational contract are voluntary. Cachon and Lariviere [37] discuss the notion of contract compliance, where they either require a firm to supply the contracted quantity (forced compliance) or they allow the firm to supply any quantity upto the contracted quantity (voluntary compliance). Cachon [33] expands on this idea in his survey paper.

Relational contracts have three main advantages over explicit contracts. First, the potential scope for relational contracts, or self-enforcing agreements, is very broad. A manufacturer and a retail distributor may agree that the performance of the retailer be adequate in dimensions of sales effort, including enthusiasm, that would be impossible to write down in an explicit contract, but which are nonetheless within the scope of implicit contracts. Second, relational contracts are flexible. The parties to a relational contract may agree that should unforeseen circumstances arise, a fair revision of the contract be undertaken even where writing down ex ante the revisions conditional upon all possible future circumstances would be impossible. Third, relational contracts avoid the costs of writing and court enforcement. Because of these advantages, many business relationships are governed more by relational contracting than by explicit contracts.

Because of these advantages, efficient relational contracts can emerge and be enforced even when the terms of the contracts conflict with existing law. The mix of relational and explicit contracts varies substantially from place to place in general and among supply chain participants. Obviously, developing countries rely heavily on relational contracts because of the weaker tools for court enforcement. But even among developed countries, relational contracts are relied upon much more in some countries than others. It is well known that business relationships in Japan are dominated by an infrastructure or network of relational contracts, as compared to the U.S. Reflecting this difference in reliance on implicit versus explicit contracts, Japan has only one-twentieth of lawyers per capita as the U.S.[15]

At the theoretical foundation of relational contracting is a simple principle from the theory of supergames. Consider two players, interacting for an infinite number of periods in an implicit contract. Suppose that the benefit to either player from cheating on the contract in any period is x and that the benefits from sustaining cooperation are π per period, for each agent. The two players can sustain the cooperation as an equilibrium outcome using grim trigger strategies ("cooperate unless there is a history involving cheating, then cheat forever") providing x

[15] Magee [128]: Table 1.

is not greater than the present value of π for each period. That is, an agreement can be sustained if $x < \pi \cdot \delta/(1 - \delta)$ where δ is the players' common discount rate. In more complex environments, where the detection of cheating is not immediate or where players do not interact in a way that yields a constant stream of benefits, the condition must, of course, be adjusted.

Relational contracts are imperfect even when they are feasible. The cooperative equilibrium is of course only one equilibrium among a continuum in the relational contracting supergame. No economic argument can support economic cooperation as the *unique* outcome of interaction even when the necessary condition for cooperation does hold. Cheating every period by both players is always a subgame perfect Nash equilibrium. This element of informal contracting thus introduces uncertainty (*intrinsic* uncertainty in the game) that could be avoided by explicit contracts. The game theoretic foundations for relational contracting assume perfect information, in the simplest model, but in reality imperfect information about contracting partners, especially at the beginning of a relationship, adds to the uncertainty. And the chief disadvantage of relational contracting is simply that the necessary condition for efficiency may fail. Where transactions between parties are large, discrete with large gaps of time between them, the gains from cheating will exceed the present value of the profit from the relationship into the future. The relational contract will fail.

Plambeck and Taylor [156] offer an analysis of a canonical setting for relational contracting in supply chains. The setting is the one we have discussed: a buyer and seller facing uncertainty about future requirements write an incomplete contract. Relational contracting plays a role because these firms anticipate the potential for future business through repeated interactions. With asymmetric information and imperfect monitoring of investments, optimal relational contracts are complex. Taylor and Plambeck characterize the optimal relational contract and also study the performance of simpler contracts.

Relational contracting is not a strategy that is used in isolation. Klein et al. [101] makes the point that explicit contracts can play the role of reference points in an ongoing relationship based on implicit contracts. (See also ref. [81].) A theme in refs. [81] and [14] is that asset ownership, explicit contracts and relational contracting are used together. Explicit contracts are designed to maximize the range over which relational contracts can be sustained. MacLeod [127] offers a superb discussion of the intersection of explicit contracts and relational contracting.

12.4 Reputation

Incentives of firms along a supply chain are influenced not just by the terms of the contracts that they write with their suppliers and buyers, and not just by the value of future transactions *within* their current relational contract partners, but also by reputation. A firm that gains a reputation for low quality or frequent stock-outs, for example, will suffer relative to a firm that sustains a reputation as a reliable supplier. Earlier in this section of the monograph we discussed reputation of managers within firms, in the context of illustrating a cost of vertical integration in reducing the incentives for effort into managing a task. Here, the focus is the reputational forces *for the firm* as a substitute for explicit contracts, vertical integration or relational contracts in coordinating incentives along a supply chain.

In the economic research on reputational dynamics from which one could draw upon in applying this idea to supply chain management, the classic articles are Shapiro [173], Klein and Leffler [105] and the Holmstrom [85] "career concerns" model. We discussed Holmstrom's model, in our subsection on vertical integration in the context of reputational incentive forces on managerial incentives. Recall that in this model, there is *both* hidden action and hidden information. An agent's reputation is the mean of the market's rational posterior distribution as to his type at any time. The agent exerts effort to increase observed output so as to raise his reputation, which is conditioned on output. In applying this model to investigate the power of reputational forces to coordinate supply chain incentives, we consider one particular decision that is subject to incentive distortions: the choice of adequate inventories.

In the inventory context, in the case of a single firm, the natural hidden information assumption is private information about firms' unit costs of production. (These are the costs of producing inventory.) A stock-out and low fill-rate in one period would lead buyers to increase their (Bayesian-updated) probability that the firm is high cost and therefore less likely to produce adequate inventory in the future. The firm's "reputation" at any time would be perceived probability that a firm was a high-cost type, conditional upon its history of stock-outs; and the firm would invest in inventory in the current period partly to avoid a drop in reputation resulting from a stock-out.

The full development of reputational dynamics would be of most value initially in a single-firm, optimal inventory model.[16] But the integration of reputational dynamics, while complex, would also add insight the theory of supply chain coordination as developed here. A new issue that would emerge in a decentralized equilibrium would be reputational spillovers: when costs are correlated among outlets, the inference by consumers from a stock-out at one outlet would be that the probability of a future stock-out higher at all outlets. The integration of reputational dynamics into inventory "availability" models would also be complicated by the fact that rational consumers would infer product availability not only from the history of fill rates but also from prices, as in ref. [49].

[16] Somewhat related ideas are explored in ref. [147] and references therein, but they do not represent reputation as a state variable.

13

Conclusion and Additional Issues in Supply Chain Management

Under the ideal conditions of frictionless, perfectly competitive markets, the management of firms along a vertical production chain would be simple. The firms would choose the right quantities of inputs to buy and outputs to produce. No incentive *coordination* via complex contracts would be necessary. The price mechanism itself would, through Adam Smith's invisible hand theorem, achieve coordination perfectly. Firms, in this ideal world, would react anonymously to input and output prices and not establish any kind of contractual or other business relationships.

In reality, we observe complex contracts all along a supply chain. This monograph reviews the dual questions at the economic foundations of supply chain management. The positive economic question: *why* do we observe complex contracts, taking the form that they do? And the normative, management question: how do we *design* an optimal contract in a given set of market conditions? Our methodology in tackling these questions involves a careful delineation of the sources of coordination failure of the price system under realistic market conditions. We identify for each market environment the externalities among firms that would distort their decisions if the firms relied solely on the simplest price contracts or mechanisms. These decisions can be pricing, quantity orders (inventories), sales effort and any other decision affecting the supply chain. Then we identify contracts that are minimally sufficient to resolve the incentive distortions. These contracts internalize or balance the externalities in some decisions and constrain the decisions of firms dimensions for which externalities are not internalized.

The main departure of market conditions from the competitive ideal is market power. We adopt in our analysis a series of different market structures incorporating market power. We identify in each setting the incentive distortions that arise in the market structure, as well as the contracts that optimally respond to the distortions. To elicit efficient decisions on prices, inventories and sales effort, contracts must incorporate nonlinear pricing, vertical price restraints (resale price maintenance), inventory buyback prices, or exclusivity restrictions. For example, resale price maintenance is in many settings essential to elicit efficient decisions on inventories.

In most of our analysis, we follow the operations management literature on supply chain coordination and the neoclassical literature on vertical control in searching for the simplest contract that can duplicate the outcome that maximizes the combined profits of the contracting firms. Equivalently, we search for the simplest contracts that can duplicate the decisions of an ideal, centralized firm. This framework recognizes the costs of writing complete contracts only implicitly, in asking how the complete contract can be avoided. The neoclassical economic approach to the incentives for vertical integration [199] similarly focusses on the benefits of integration, setting aside the real-world costs of integration by supposing that the new entity implements efficient decisions costlessly.

We then synthesize, in Section 12, recent approaches to the full costs and benefits of vertical integration. Viewed from the perspective of a firm looking upstream, this is the make-or-buy decision, but a firm faces a parallel decision in the choice of vertical integration versus decentralization (e.g., vertical integration versus franchising) in downstream operations. Following the modern literature on integration, we delineate the impact of moving a transaction from the market to the firm. Vertical integration does not magically eliminate the costs of transacting. Instead, following Coase's principle, one must compare the various sources of transactions costs within a firm to the costs of transacting in a market to arrive at an optimal make-or-buy decision. "Transactions costs" include the costs of mitigating incentive distortions as well as the consequences of remaining distortions. Vertically integrating to eliminate the hold-up problem associated with specific assets and vertical integration as a decision on the assignment of the control rights of assets, are two examples of prominent themes in this area.

In synthesizing the foundational issues in supply chain management, we have left open many important topics. We conclude with a brief discussion of six of these.

The 2–2 market structure: Within the set of market structures, we left open the case of the "2–2" structure. In this setting, two upstream manufacturers compete, both offering products through two downstream retailers. While not dealt with substantially in the literature, this is an obviously realistic structure. Any retailer in this setting may have an *influence* over which product a consumer will buy. Producers then compete for the sales effort or influence of retailers in selling their products. A more attractive contract by a manufacturer, including a higher markup of price over cost, leaves the retailer with an incentive to push the manufacturer's product. At the same time, retailers compete in a retail market that rewards them for gaining reputations as respecting consumers' preferences or well-being in recommending products. A prominent role of multiproduct retailers such as department stores is to screen products on behalf of consumers, offering sets of products of high utility for any cost level [130]; two extremes of this model are well-known. If differentiation in the retail market is so strong that retailers are effectively local monopolists, then the model collapses to the Bernheim–Whinston [17] common agency model. If retail differentiation approaches zero, and competitive retailer pressure increases to promote the highest utility product package (conditional upon costs), then the model collapses to two manufacturers competing through competitive intermediaries. We conjecture that contract equilibria in the intermediate structure in which retailers compete imperfectly involves manufacturer competition for retailer influence (through, perhaps, contracts such as resale price maintenance or slotting allowances), constrained in their ability to capture retailer influence by the discipline of competition across retailers.[1]

Market allocation versus contracts: the portfolio approach: We have to this point analyzed the optimal single contract or organizational arrangement for a single transaction. One can pose the problem as whether a particular input should be supplied by the market with: (a) a simple contract; or (b) through a more complex contract; or (c) as through vertical integration of the transaction within the firm. In reality, firms often use a portfolio approach. A firm may, for example, obtain some amount of an input from the market and produce an additional amount internally. Or, looking downstream, a firm may employ the strategy known as *dual distribution*, in distributing some of its product through owned outlets and the rest of its output through independent (but contractually constrained) outlets. The portfolio approach extends the Coasian question of markets *versus* intra-firm allocation to the question of the *mix* of each alternative to adopt.

An example of this approach would be the need for a firm to integrate into supply — to some extent — to assure optimally critical supply of an input. A lumber mill may purchase most of its electricity from the market, but retain a diesel generating system for use in extreme circumstances. The provision of the input, electricity, is neither entirely internal to the firm nor purchased entirely in the market. The optimum is a mix of the two. Dual distribution, to take another example, is used extensively in the restaurant chain business with some outlets being owned centrally and others operated as franchises. Wu and Kleindorfer [209] characterize the optimum mix of spot market

[1] Adida and DeMiguel [2] adopt the market structure of multiple manufacturers selling through multiple retailers, in analyzing supply chain competition. These authors show that with sufficiently many firms at each level, the equilibrium (which can be computed using numerical optimization techniques) approaches first-best efficiency. Asymmetries or concentration in either the manufacturer or retailer level introduce distortions. On the economics side of the literature, a well-known article, ref. [149] in the context of exclusionary strategies or strategies that raise profits by raising the costs of production for a rival. For a recent economic theory of vertical restraints in a 2–2 market structure see ref. [162].

purchases of an input and contracted procurement of the input, in the context of electronic B2B exchanges. Tunca and Zenios [193] consider auctions and relational contracts, and characterize conditions where both procurement approaches co-exist and where only one of the two approaches exists in equilibrium. Gallini and Lutz [67] offer a theory of dual distribution involving signalling of high quality by the upstream firm. Lafontaine [112] and Lafontaine and Shaw [113] offer evidence related to the optimal choices of dual distribution.

Contractual assignment of decision rights along a supply chain: The contractual assignment of decision rights (decisions on inventory, pricing and asset disposition) has entered our survey in only one place: the property rights theory of vertical integration. In our analysis of the coordination of optimal inventory, for example, we have taken as predetermined the decision of which firms should take *responsibility* for the inventory decision. The bulk of the literature follows this approach. In reality, however, the decision of whether a downstream firm should invest in inventory or leave the decision to upstream suppliers, is important and sometimes highly contentious. After the 2000 bust in the technology sector, Cisco was criticized for having indemnified the risks of inventory investment by upstream suppliers.[2] It seems likely to us that it was in fact efficient to allocate inventory risk to the downstream firm, Cisco, since it was in a better position to remain informed about demands for final products. With asymmetric information arising over time in any organization or set of contracts, decisions should in general be allocated to those who are best informed about the costs and benefits of the decisions. On the issue of downstream versus upstream inventory decisions, Mishra and Raghunathan [139] analyze vendor-managed inventory as an alternative to retailer inventory decisions. These authors show that vendor-managed inventory can intensify upstream competition to the benefit of the downstream firm. (This result is established under the assumption of non-transferability of profits, unlike our main approach in this paper.) Cachon [34] compares the impact of downstream inventory decisions with upstream (vendor-managed) inventory but in a model that does not involve the

[2] See ref. [142].

choice of the optimal allocation of the responsibility to take inventory decisions.[3]

Flexibility in production decisions: Increased operational flexibility and responsiveness can significantly improve supply chain performance, and firms often invest in technologies (e.g., flexible manufacturing systems) and processes (e.g., delayed differentiation) that enhance flexibility and responsiveness. In a supply chain context, however, the strategic impact of such investments need to be considered. An early paper to study the strategic effects of investing in responsiveness within a supply chain was ref. [92]. The authors make the point that if an upstream firm reduces its lead time to supply a downstream firm, the upstream firm may end up decreasing its sales because the downstream firm has less need for high inventories. This argument works only under the assumption of non-transferable profits: if the investment in lead time reduction would increase joint profits, then it would be undertaken if the firms could share the increased profits. Krishnan et al. [109] show, in a common agency setting, that lead time reduction by one upstream agent can reduce the downstream retailer's incentive to promote that agent's product. Simple distribution contracts such as minimum-take contracts, advance-purchase discounts, and exclusive dealing, when adopted in conjunction with investments in lead time reduction, can remedy the incentive problem. Anand and Girotra [6] also identify another potential pitfall of operational improvement. These authors argue that delayed differentiation, which is beneficial in a centralized setting, may not be an equilibrium under competition. In their model, delayed differentiation allows a firm to exit a competitive market and instead sell the product in other markets. By not delaying differentiation, firms make a credible commitment to compete aggressively. This commitment carries a strategic advantage.

Coalitional incentives within a supply chain: Competing firms in a supply chain often have an interest in collaborating on specific aspects of their operations. For example, competing retailers that stock inventories in anticipation of demand may find it mutually beneficial to

[3] The order of moves in setting the wholesale price, w, drives the results of the Cachon model.

pool their inventories. Or they may transship inventory once demand uncertainty is resolved, by moving unsold inventory from one location to another location where there is excess demand. Anupindi and Bassok [9] analyze the impact of inventory pooling (centralization) by downstream retailers on the manufacturer's profits. The manufacturer benefits only if consumers facing a stock-out at a particular are unlikely to search for the product at other outlets. Anupindi et al. [10] and Granot and Sosic [73] analyze the sequential inventory ordering and allocation decisions and characterize the impact of different (ex post) allocation rules on the incentives of retailers to share residual inventories. Depending on the allocation rule, retailers may not share residual inventories. Sosic [179] uses the notion of "farsighted" stability where retailers consider not just immediate payoffs but also reactions of other retailers to their decisions. Bernstein and Nagarajan [24] summarize related issues.

Information management in supply chains: We have synthesized incentive management issues within supply chains and organizations taking the information structure of the environment or organization as predetermined. In reality, the management of information is itself a critical area and the distribution of information is endogenous. A literature in operations management has focussed on the challenges, benefits and incentives associated with information sharing in supply chains. Information distortion in supply chains, resulting in the "bullwhip effect," has been discussed in ref. [120]. In a vertical supply chain, Gavirneni et al. [68] and Lee et al. [121] have demonstrated the value of information sharing. Ren et al. [160] show that long-term relationships (i.e., repeated interactions) can provide incentives for firms to share forecast information truthfully.

In a setting with one manufacturer and multiple retailers, Cachon and Fisher [35] demonstrate the value of sharing information. The supply chain structure analyzed by Cachon and Fisher [35] also allows the study of "information leakage" in the supply chain. In ref. [123] retailers compete as Cournot firms and procure an input from a common supplier. The price set by the input supplier in Li's model depends on information shared by retailers with the supplier. Retailers in Li's model who choose not to share information with the supplier can infer the information of those who do from the input price. Li shows that this potential for "information leakage" inhibits incentives to share information, and in fact in his model retailers do not share at all with the upstream supplier. Li and Zhang [124] and Gal-Or et al. [65] investigate similar issues. Anand and Goyal [7] analyze a model in which product or material flows and information flows are jointly determined. Information acquisition is a decision variable in their model, rather than an exogenous endowment. The decision to share information is made ex post to the realization of signals rather than committed to ex ante as in other papers in the literature. Firms may decide not to share information, so as to avoid distortions in product market flow that may result from unavoidable leakage. A related topic is the incentive for voluntary sharing of information by competitors, which is studied by Vives [197], Li [122], Gal-Or [63, 64] and Shapiro [174]. The integration of information management, product flow management and longer run investment management remains a critical and open area in supply chain management.

We have merely sketched additional topics within the foundations of supply chain management. And this monograph is merely an outline of the main issues in the economic foundations of supply chain management. As Spence [180] suggested in his Nobel Prize lecture:

> the range of economic activity that can be effectively coordinated across a complex multifirm supply chain is just starting to be explored by companies and those who do research on these subjects. The boundaries of the firm, transaction costs, supply-chain architecture and coordination, and outsourcing are all facets of a large mosaic in which incentives, communication and coordination, and the boundaries of the firm are worked out. . . . these issues are far from being settled in the world of practice and in the world of economic research.

References

[1] J. M. Abito and J. Wright, “Exclusive dealing with imperfect downstream competition,” *International Journal of Industrial Organization*, vol. 26, no. 1, pp. 227–246, 2008.

[2] E. Adida and V. DeMiguel, “Supply chain competition with multiple manufacturers and retailers,” *Operations Research*, vol. 59, no. 1, pp. 156–172, 2011.

[3] P. Aghion and P. Bolton, “Contracts as a barrier to entry,” *American Economic Review*, vol. 77, no. 3, pp. 388–401, 1987.

[4] A. A. Alchian and H. Demsetz, “Production, information costs, and economic organization,” *The American Economic Review*, vol. 62, no. 5, pp. 777–795, 1972.

[5] K. S. Anand, R. Anupindi, and Y. Bassok, “Strategic inventories in vertical contracts,” *Management Science*, vol. 54, no. 10, pp. 1792–1804, 2008.

[6] K. S. Anand and K. Girotra, “The strategic perils of delayed differentiation,” *Management Science*, vol. 53, no. 5, pp. 697–712, 2007.

[7] K. S. Anand and M. Goyal, “Strategic information management under leakage in a supply chain,” *Management Science*, vol. 55, no. 3, pp. 438–452, 2009.

[8] J. J. Anton and G. D. Varma, “Storability, market structure, and demand-shift incentives,” *The RAND Journal of Economics*, vol. 36, no. 3, pp. 520–543, 2005.

[9] R. Anupindi and Y. Bassok, “Centralization of stocks: Retailers vs. manufacturer,” *Management Science*, vol. 45, no. 2, pp. 178–191, 1999.

[10] R. Anupindi, Y. Bassok, and E. Zemel, “A general framework for the study of decentralized distribution systems,” *Manufacturing & Service Operations Management*, vol. 3, no. 4, pp. 349–368, 2001.

[11] K. J. Arrow, “The role of securities in the optimal allocation of risk-bearing,” *The Review of Economic Studies*, vol. 31, no. 2, pp. 91–96, 1964.

[12] A. Arya and B. Mittendorf, “Using return policies to elicit retailer information,” *The RAND Journal of Economics*, vol. 35, no. 3, pp. 617–630, 2004.

[13] A. Arya and B. Mittendorf, “Benefits of channel discord in the sale of durable goods,” *Marketing Science*, vol. 25, no. 1, pp. 91–96, 2006.

[14] G. Baker, R. Gibbons, and K. J. Murphy, “Relational contracts and the theory of the firm,” *The Quarterly Journal of Economics*, vol. 117, no. 1, pp. 39–84, 2002.

[15] A. Balakrishnan, M. S. Pangburn, and E. Stavrulaki, ““Stack them high, let ‘em fly”: Lot-sizing policies when inventories stimulate demand,” *Management Science*, vol. 50, no. 5, pp. 630–644, 2004.

[16] D. P. Baron and R. B. Myerson, “Regulating a monopolist with unknown costs,” *Econometrica*, vol. 50, no. 4, pp. 911–930, 1982.

[17] B. D. Bernheim and M. D. Whinston, “Common marketing agency as a device for facilitating collusion,” *The RAND Journal of Economics*, vol. 16, no. 2, pp. 269–281, 1985.

[18] B. D. Bernheim and M. D. Whinston, “Common agency,” *Econometrica*, vol. 54, no. 4, pp. 923–942, 1986.

[19] F. Bernstein and G. A. DeCroix, “Decentralized pricing and capacity decisions in a multitier system with modular assembly,” *Management Science*, vol. 50, no. 9, pp. 1293–1308, 2004.

[20] F. Bernstein, G. A. DeCroix, and Y. Wang, “Incentives and commonality in a decentralized multiproduct assembly system,” *Operations Research*, vol. 55, no. 4, pp. 630–646, 2007.

[21] F. Bernstein and A. Federgruen, “A general equilibrium model for industries with price and service competition,” *Operations Research*, vol. 52, no. 6, pp. 868–886, 2004.

[22] F. Bernstein and A. Federgruen, “Decentralized supply chains with competing retailers under demand uncertainty,” *Management Science*, vol. 51, no. 1, pp. 18–29, 2005.

[23] F. Bernstein and A. Federgruen, “Coordination mechanisms for supply chains under price and service competition,” *Manufacturing & Service Operations Management*, vol. 9, no. 3, pp. 242–262, 2007.

[24] F. Bernstein and M. Nagarajan, “Competition and cooperative game theory models is supply chains,” *Foundations and Trends in Technology, Information and Operations Management*, Forthcoming.

[25] S. Bhattacharyya and F. Lafontaine, “Double-sided moral hazard and the nature of share contracts,” *The RAND Journal of Economics*, vol. 26, no. 4, pp. 761–781, 1995.

[26] B. F. Blair and T. R. Lewis, “Optimal retail contracts with asymmetric information and moral hazard,” *The RAND Journal of Economics*, vol. 25, no. 2, pp. 284–296, 1994.

[27] P. Bolton and M. Dewatripont, *Contract Theory.* MIT Press, 2005.

[28] G. Bonanno and J. Vickers, “Vertical separation,” *The Journal of Industrial Economics*, vol. 36, no. 3, pp. 257–265, 1988.

[29] J. I. Bulow, "Durable-goods monopolists," *Journal of Political Economy*, vol. 90, no. 2, pp. 314–332, 1982.

[30] J. I. Bulow, J. D. Geanakoplos, and P. D. Klemperer, "Multimarket oligopoly: Strategic substitutes and complements," *Journal of Political Economy*, vol. 93, no. 3, pp. 488–511, 1985.

[31] M. L. Burstein, "The economics of tie-in sales," *The Review of Economics and Statistics*, vol. 42, no. 1, pp. 68–73, 1960.

[32] D. A. Butz, "Vertical price controls with uncertain demand," *Journal of Law and Economics*, vol. 40, no. 2, pp. 433–460, 1997.

[33] G. P. Cachon, "Supply chain coordination with contracts," in *Handbooks in Operations Research and Management Science*, vol. 11*, Supply Chain Management: Design, Coodination and Operation*, (A. G. de Kok and S. C. Graves, eds.), pp. 229–339, North Holland, Amsterdam: Elsevier, 2003.

[34] G. P. Cachon, "The allocation of inventory risk in a supply chain: Push, pull, and advance-purchase discount contracts," *Management Science*, vol. 50, no. 2, pp. 222–238, 2004.

[35] G. P. Cachon and M. Fisher, "Supply chain inventory management and the value of shared information," *Management Science*, vol. 46, no. 8, pp. 1032–1048, 2000.

[36] G. P. Cachon and A. G. Kok, "Competing manufacturers in a retail supply chain: On contractual form and coordination," *Management Science*, vol. 56, no. 3, pp. 571–589, 2010.

[37] G. P. Cachon and M. A. Lariviere, "Contracting to assure supply: How to share demand forecasts in a supply chain," *Management Science*, vol. 47, no. 5, pp. 629–646, 2001.

[38] G. P. Cachon and M. A. Lariviere, "Supply chain coordination with revenue-sharing contracts: Strengths and limitations," *Management Science*, vol. 51, no. 1, pp. 30–44, 2005.

[39] G. P. Cachon and P. H. Zipkin, "Competitive and cooperative inventory policies in a two-stage supply chain," *Management Science*, vol. 45, no. 7, pp. 936–953, 1999.

[40] F. Chen, "Decentralized supply chains subject to information delays," *Management Science*, vol. 45, no. 8, pp. 1076–1090, 1999.

[41] A. J. Clark and H. Scarf, "Optimal policies for a multi-echelon inventory problem," *Management Science*, vol. 6, no. 4, pp. 475–490, 1960.

[42] R. H. Coase, "The nature of the firm," *Economica*, vol. 4, no. 16, pp. 386–405, 1937.

[43] R. H. Coase, "The problem of social cost," *Journal of Law and Economics*, vol. 3, pp. 1–44, 1960.

[44] R. H. Coase, "Durability and monopoly," *Journal of Law and Economics*, vol. 15, no. 1, pp. 143–149, 1972.

[45] C. J. Corbett, D. Zhou, , and C. S. Tang, "Designing supply contracts: Contract type and information asymmetry," *Management Science*, vol. 50, no. 4, pp. 550–559, 2004.

[46] A. T. Coughlan, "Competition and cooperation in marketing channel choice: Theory and application," *Marketing Science*, vol. 4, no. 2, pp. 110–129, 1985.

[47] A. T. Coughlan and B. Wernerfelt, “On credible delegation by oligopolists: A discussion of distribution channel management,” *Management Science*, vol. 35, no. 2, pp. 226–239, 1989.

[48] K. J. Crocker and S. E. Masten, “Pretia ex Machina? Prices and process in long term contracts,” *Journal of Law and Economics*, vol. 34, no. 1, pp. 69–99, 1991.

[49] J. D. J. Dana, “Monopoly price dispersion under demand uncertainty,” *International Economic Review*, vol. 42, no. 3, pp. 649–670, 2001.

[50] J. D. J. Dana and N. C. Petruzzi, “Note: The newsvendor model with endogenous demand,” *Management Science*, vol. 47, no. 11, pp. 1488–1497, 2001.

[51] P. Dasgupta, P. Hammond, and E. Maskin, “The implementation of social choice rules: Some general results on incentive compatibility,” *The Review of Economic Studies*, vol. 46, no. 2, pp. 185–216, 1979.

[52] G. Debreu, *Theory of Value: An Axiomatic Analysis of Economic Equilibrium.* 1959.

[53] J. S. Demski and D. Sappington, “Optimal incentive contracts with multiple agents,” *Journal of Economic Theory*, vol. 33, no. 1, pp. 152–171, 1984.

[54] R. Deneckere, H. P. Marvel, and J. Peck, “Demand uncertainty, inventories, and resale price maintenance,” *The Quarterly Journal of Economics*, vol. 111, no. 3, pp. 885–913, 1996.

[55] R. Deneckere, H. P. Marvel, and J. Peck, “Demand uncertainty and price maintenance: Markdowns as destructive competition,” *The American Economic Review*, vol. 87, no. 4, pp. 619–641, 1997.

[56] P. S. Desai and K. Srinivasan, “Demand signalling under unobservable effort in franchising: Linear and nonlinear price contracts,” *Management Science*, vol. 41, no. 10, pp. 1608–1623, 1995.

[57] M. Dewatripont, I. Jewitt, and J. Tirole, “The economics of career concerns, Part I: Comparing information structures,” *Review of Economic Studies*, vol. 66, no. 1, pp. 183–198, 1999.

[58] R. Dorfman and P. Steiner, “Optimal advertising and optimal quality,” *American Economic Review*, vol. 44, no. 5, pp. 826–836, 1954.

[59] P. Dudine, I. Hendel, and A. Lizzeri, “Storable good monopoly: The role of commitment,” *American Economic Review*, vol. 96, no. 5, pp. 1706–1719, 2006.

[60] M. Eisenberg, “Relational contracts,” in *Good Faith and Fault in Contract Law*, (J. Beatson and D. Friedmann, eds.), pp. 291–304, Clarendon Press, 1995.

[61] D. Fudenburg and J. Tirole, *Game Theory.* Massachusetts Institute of Technology, 1991.

[62] C. Fumagalli and M. Motta, “Exclusive dealing and entry, when buyers compete,” *American Economic Review*, vol. 96, no. 3, pp. 785–795, 2006.

[63] E. Gal-Or, “Information sharing in oligopoly,” *Econometrica*, vol. 53, no. 2, pp. 329–343, 1985.

[64] E. Gal-Or, “Information transmission — Cournot and Bertrand equilibria,” *Review of Economic Studies*, vol. 53, no. 1, pp. 85–92, 1986.

[65] E. Gal-Or, T. Geylani, and A. J. Dukes, “Information sharing in a channel with partially informed retailers,” *Marketing Science*, vol. 27, no. 4, pp. 642–658, 2008.

[66] N. Gallini and B. Wright, “Technology transfer under asymmetric information,” *The RAND Journal of Economics*, vol. 21, no. 1, pp. 147–160, 1990.

[67] N. T. Gallini and N. A. Lutz, “Dual distribution and royalty fees in franchising,” *Journal of Law, Economics, & Organization*, vol. 8, no. 3, pp. 471–501, 1992.

[68] S. Gavirneni, R. Kapuscinski, and S. Tayur, “Value of information in capacitated supply chains,” *Management Science*, vol. 45, no. 1, pp. 16–24, 1999.

[69] A. Gibbard, “Manipulation of voting schemes: A general result,” *Econometrica*, vol. 41, no. 4, pp. 587–601, 1973.

[70] R. Gibbons, “Four formal(izable) theories of the firm?,” *Journal of Economic Behavior & Organization*, vol. 58, no. 2, pp. 200–245, 2005.

[71] V. P. Goldberg and J. R. Erickson, “Quantity and price adjustment in long-term contracts: A case study of petroleum coke,” *Journal of Law and Economics*, vol. 30, no. 2, pp. 369–398, 1987.

[72] J. R. Gould and L. E. Preston, “Resale price maintenance and retail outlets,” *Economica*, vol. 32, no. 127, pp. 302–312, 1965.

[73] D. Granot and G. Sosic, “A three-stage model for a decentralized distribution system of retailers,” *Operations Research*, vol. 51, no. 5, pp. 771–784, 2003.

[74] S. J. Grossman and O. D. Hart, “An analysis of the principal-agent problem,” *Econometrica*, vol. 51, no. 1, pp. 7–45, 1983.

[75] S. J. Grossman and O. D. Hart, “The costs and benefits of ownership: A theory of vertical and lateral integration,” *Journal of Political Economy*, vol. 94, no. 4, pp. 691–719, 1986.

[76] P. A. Grout, “Investment and wages in the absence of binding contracts: A nash bargaining approach,” *Econometrica*, vol. 52, no. 2, pp. 449–460, 1984.

[77] S. Gupta and R. Loulou, “Process innovation, product differentiation, and channel structure: Strategic incentives in a duopoly,” *Marketing Science*, vol. 17, no. 4, pp. 301–316, 1998.

[78] G. K. Hadfield, “Problematic relations: Franchising the law of incomplete contracts,” *Stanford Law Review*, vol. 42, pp. 927–992, 1989-1990.

[79] O. Hart, “Thinking about the firm: A review of daniel spulber’s the theory of the firm,” *Journal of Economic Literature*, vol. 94, no. 1, pp. 101–113, 2011.

[80] O. Hart and J. Moore, “Property rights and the nature of the firm,” *Journal of Political Economy*, vol. 98, no. 6, pp. 1119–1158, 1990.

[81] O. Hart and J. Moore, “Contracts as reference points,” *The Quarterly Journal of Economics*, vol. 123, no. 1, pp. 1–48, 2008.

[82] O. D. Hart, *Firms, Contracts, and Financial Structure.* Clarendon Press, 1995.

[83] B. Holmstrom, “Moral hazard and observability,” *The Bell Journal of Economics*, vol. 10, no. 1, pp. 74–91, 1979.

[84] B. Holmstrom, “Moral hazard in teams,” *The Bell Journal of Economics*, vol. 13, no. 2, pp. 324–340, 1982.

[85] B. Holmstrom, "Managerial incentive problems: A dynamic perspective," *Review of Economic Studies*, vol. 66, no. 1, pp. 169–182, 1999.

[86] B. Holmstrom and P. Milgrom, "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design," *Journal of Law, Economics, & Organization*, vol. 7, pp. 24–52, 1991.

[87] A. P. Hourihan and J. W. Markham, *The Effects of Fair Trade Repeal: The Case of Rhode Island*. Cambridge, MA: Marketing Science Institute and Center for Economic Studies, 1974.

[88] M. Hviid, "Long-term contracts and relational contracts," in *Encyclopedia of Law and Economics*, (B. Bouckaert and G. D. Geest, eds.), pp. 46–72, Edward Elgar: Cheltenham, 2000.

[89] E. Iacobucci and R. A. Winter, "Vertical restraints across jurisdictions," in forthcoming *in Oxford Handbook on International Antitrust Economics*, (R. D. Blair and D. D. Sokol, eds.), Oxford University Press, 2013.

[90] P. M. Ippolito, "Resale price maintenance: Empirical evidence from litigation," *Journal of Law and Economics*, vol. 34, no. 2, pp. 263–294, 1991.

[91] P. M. Ippolito and T. R. Overstreet, "Resale price maintenance: An economic assessment of the federal trade commission's case against the corning glass works," *Journal of Law and Economics*, vol. 39, no. 1, pp. 285–328, 1996.

[92] A. V. Iyer and M. E. Bergen, "Quick response in manufacturer-retailer channels," *Management Science*, vol. 43, no. 4, pp. 559–570, 1997.

[93] M. C. Jensen and W. H. Meckling, "Theory of the firm: Managerial behavior, agency costs and ownership structure," *Journal of Financial Economics*, vol. 3, no. 4, pp. 305–360, 1976.

[94] A. P. Jeuland and S. M. Shugan, "Managing channel profits," *Marketing Science*, vol. 2, no. 3, pp. 239–272, 1983.

[95] L. Jiang and Y. Wang, "Supplier competition in decentralized assembly systems with price-sensitive and uncertain demand," *Manufacturing & Service Operations Management*, vol. 12, no. 1, pp. 93–101, 2010.

[96] R. Jing and R. A. Winter, "Exclusionary contracts," University of British Columbia Working Paper, 2011.

[97] M. L. Katz, "Vertical contractual relations," in *Handbook of Industrial Organization*, vol. 1, (R. Schmalensee and R. D. Willig, eds.), pp. 655–721, North Holland, Amsterdam: Elsevier, 1989.

[98] M. L. Katz, "Game-playing agents: Unobservable contracts as precommitments," *The RAND Journal of Economics*, vol. 22, no. 3, pp. 307–328, 1991.

[99] B. Klein, "Transaction cost determinants of "Unfair" contractual arrangements," *The American Economic Review*, vol. 70, no. 2, pp. 356–362, 1980.

[100] B. Klein, "Vertical integration as organizational ownership: The fisher body-general motors relationship revisited," *Journal of Law, Economics, & Organization*, vol. 4, no. 1, pp. 199–213, 1988.

[101] B. Klein, "Why would hold-ups occur: The self-enforcing range of contractual relationships," *Economic Inquiry*, vol. 34, no. 3, pp. 444–463, 1996.

[102] B. Klein, "Fisher general motors and the nature of the firm," *Journal of Law and Economics*, vol. 43, no. 1, pp. 105–142, 2000.

[103] B. Klein, "The economic lessons of fisher body and general motors," *International Journal of the Economics of Business*, vol. 14, no. 1, pp. 1–36, 2007.

[104] B. Klein, R. G. Crawford, and A. A. Alchian, "Vertical integration, appropriable rents, and the competitive contracting process," *Journal of Law and Economics*, vol. 21, no. 2, pp. 297–326, 1978.

[105] B. Klein and K. Leffler, "The role of market forces in assuring contractual performance," *Journal of Political Economy*, vol. 89, no. 4, pp. 615–641, 1981.

[106] B. Klein and K. M. Murphy, "Vertical restraints as contract enforcement mechanisms," *Journal of Law and Economics*, vol. 31, no. 2, pp. 265–297, 1988.

[107] H. Krishnan, "The Monopolist and the Newsvendor," University of British Columbia Working Paper, 2011.

[108] H. Krishnan, R. Kapuscinski, and D. A. Butz, "Coordinating contracts for decentralized supply chains with retailer promotional effort," *Management Science*, vol. 50, no. 1, pp. 48–63, 2004.

[109] H. Krishnan, R. Kapuscinski, and D. A. Butz, "Quick Response and Retailer Effort," *Management Science*, vol. 56, no. 6, pp. 962–977, 2010.

[110] H. Krishnan and R. A. Winter, "Vertical control of price and inventory," *American Economic Review*, vol. 97, no. 5, pp. 1840–1857, 2007.

[111] H. Krishnan and R. A. Winter, "Inventory dynamics and supply chain coordination," *Managment Science*, vol. 56, no. 1, pp. 141–147, 2010.

[112] F. Lafontaine, "Agency theory and franchising: some empirical results," *The RAND Journal of Economics*, vol. 23, no. 2, pp. 263–283, 1992.

[113] F. Lafontaine and K. L. Shaw, "The dynamics of franchise contracting: Evidence from panel data," *Journal of Political Economy*, vol. 107, no. 5, pp. 1041–1080, 1999.

[114] F. Lafontaine and M. Slade, "Exclusive contracts and vertical restaints: Empirical evidence and public policy," 2005.

[115] F. Lafontaine and M. Slade, "Vertical integration and firm boundaries: The evidence," *Journal of Economic Literature*, vol. 45, no. 3, pp. 629–685, 2007.

[116] F. Lafontaine and M. Slade, "Inter-Firm Contracts: Evidence," in *Handbook of Organizational Economics*, (R. Gibbons and J. Roberts, eds.), Princeton University Press, 2012. Forthcoming.

[117] M. A. Lariviere and E. L. Porteus, "Selling to the newsvendor: An analysis of price-only contracts," *Manufacturing & Service Operations Management*, vol. 3, no. 4, pp. 293–305, 2001.

[118] E. P. Lazear and S. Rosen, "Rank-order tournaments as optimum labor contracts," *Journal of Political Economy*, vol. 89, no. 5, pp. 841–864, 1981.

[119] H. Lee and S. Whang, "Decentralized multi-echelon supply chains: Incentives and information," *Management Science*, vol. 45, no. 5, pp. 633–640, 1999.

[120] H. L. Lee, V. Padmanabhan, and S. Whang, "Information distortion in a supply chain: The bullwhip effect," *Management Science*, vol. 43, no. 4, pp. 546–558, 1997.

[121] H. L. Lee, K. C. So, and C. S. Tang, "The value of information sharing in a two-level supply chain," *Management Science*, vol. 46, no. 5, pp. 626–643, 2000.

[122] L. Li, "Cournot oligopoly with information sharing," *The RAND Journal of Economics*, vol. 16, no. 4, pp. 521–536, 1985.

[123] L. Li, "Information sharing in a supply chain with horizontal competition," *Management Science*, vol. 48, no. 9, pp. 1196–1212, 2002.

[124] L. Li and H. Zhang, "Confidentiality and information sharing in supply chain coordination," *Management Science*, vol. 54, no. 8, pp. 1467–1481, 2008.

[125] S. A. Lippman and K. F. McCardle, "The competitive newsboy," *Operations Research*, vol. 45, no. 1, pp. 54–65, 1997.

[126] S. Macaulay, "Non-contractual relations in business: A preliminary study," *American Sociological Review*, vol. 28, no. 1, pp. 55–67, 1963.

[127] W. B. MacLeod, "Reputations, relationships, and contract enforcement," *Journal of Economic Literature*, vol. 45, no. 3, pp. 595–628, 2007.

[128] S. P. Magee, "The optimum number of lawyers and a radical proposal for legal change," 2010.

[129] H. G. Manne, "Mergers and the market for corporate control," *Journal of Political Economy*, vol. 73, no. 2, pp. 110–120, 1965.

[130] H. P. Marvel and S. McCafferty, "Resale price maintenance and quality certification," *The RAND Journal of Economics*, vol. 15, no. 3, pp. 346–359, 1984.

[131] H. P. Marvel and J. Peck, "Demand uncertainty and returns policies," *International Economic Review*, vol. 36, no. 3, pp. 691–714, 1995.

[132] H. P. Marvel and H. Wang, "Inventories, manufactuer returns policies, and equilibrium price dispersion under demand uncertainty," *Journal of Economics and Management Strategy*, vol. 16, no. 4, pp. 1031–1051, 2007.

[133] H. P. Marvel and H. Wang, "Distribution contracts to support optimal inventory holdings under demand uncertainty," *International Journal of Industrial Organization*, vol. 27, pp. 625–631, 2009.

[134] A. Mas-Colell, M. D. Whinston, and J. R. Green, *Microeconomic Theory*. New York: Oxford University Press, 1995.

[135] S. E. Masten, "The organization of production: Evidence from the aerospace industry," *Journal of Law and Economics*, vol. 27, no. 2, pp. 403–417, 1984.

[136] F. Mathewson and R. Winter, "Territorial restrictions in franchise contracts," *Economic Inquiry*, vol. 32, no. 2, pp. 181–192, 1994.

[137] G. F. Mathewson and R. Winter, "An economic theory of vertical restraints," *The RAND Journal of Economics*, vol. 15, no. 1, pp. 27–38, 1984.

[138] T. W. McGuire and R. Staelin, "An industry equilibrium analysis of downstream vertical integration," *Marketing Science*, vol. 2, no. 2, pp. 161–191, 1983.

[139] B. K. Mishra and S. Raghunathan, "Retailer- vs. vendor-managed inventory and brand competition," *Management Science*, vol. 50, no. 4, pp. 445–457, 2004.

[140] J. H. Mortimer, "Vertical contracts in the video rental industry," *The Review of Economic Studies*, vol. 75, no. 1, pp. 165–199, 2008.

[141] R. B. Myerson, "Incentive compatibility and the bargaining problem," *Econometrica*, vol. 47, no. 1, pp. 61–73, 1979.

[142] V. Narayanan and A. Raman, "Aligning incentives in supply chains," *Harvard Business Review*, vol. 82, no. 11, pp. 94–102, 2004.

[143] V. G. Narayanan, A. Raman, and J. Singh, "Agency costs in a supply chain with demand uncertainty and price competition," *Management Science*, vol. 51, no. 1, pp. 120–132, 2005.

[144] H. M. Neary and R. A. Winter, "Output shares in bilateral agency contracts," *Journal of Economic Theory*, vol. 66, no. 2, pp. 609–614, 1995.

[145] S. Netessine and F. Zhang.

[146] W. Y. Oi, "A disneyland dilemma: Two-part tariffs for a mickey mouse monopoly," *The Quarterly Journal of Economics*, vol. 85, no. 1, pp. 77–96, 1971.

[147] T. Olsen and R. Parker, "Inventory management under market size dynamics," *Management Science*, vol. 54, no. 10, pp. 1805–1821, 2008.

[148] J. A. Ordover and J. C. Panzar, "On the nonlinear pricing of inputs," *International Economic Review*, vol. 23, no. 3, pp. 659–675, 1982.

[149] J. A. Ordover, G. Saloner, and S. C. Salop, "Equilibrium vertical foreclosure," *The American Economic Review*, vol. 80, no. 1, pp. 127–142, 1990.

[150] T. R. Overstreet, *Resale Price Maintenance: Economic Theories and Empirical Evidence.* Washington, DC: Federal Trade Commission, 1983.

[151] M. Packalen, "Market share exclusion," University of Waterloo Working Paper, 2011.

[152] V. Padmanabhan and I. Png, "Returns policies: Make money by making good," *Sloan Management Review*, vol. 37, no. 1, pp. 65–72, 1995.

[153] B. A. Pasternack, "Optimal pricing and return policies for perishable commodities," *Marketing Science*, vol. 4, no. 2, pp. 166–176, 1985.

[154] G. Perakis and G. Roels, "The price of anarchy in supply chains: Quantifying the efficiency of price-only contracts," *Management Science*, vol. 53, no. 8, pp. 1249–1268, 2007.

[155] N. C. Petruzzi and M. Dada, "Pricing and the newsvendor problem: A review with extensions," *Operations Research*, vol. 47, no. 2, pp. 183–194, 1999.

[156] E. L. Plambeck and T. A. Taylor, "Implications of renegotiation for optimal contract flexibility and investment," *Management Science*, vol. 53, no. 12, pp. 1872–1886, 2007.

[157] E. L. Porteus, "Responsibility tokens in supply chain management," *Manufacturing & Service Operations Management*, vol. 2, no. 2, pp. 203–219, 2000.

[158] E. B. Rasmusen, J. M. Ramseyer, and J. S. Wiley, "Naked Exclusion," *American Economic Review*, vol. 81, no. 5, pp. 1137–1145, 1991.

[159] L. E. Read, "I, pencil: My family tree as told to Leonard E. Read," http://www.econlib.org/library/Essays/rdPncl1.html, 1999.

[160] Z. J. Ren, M. A. Cohen, T. H. Ho, and C. Terwiesch, "Information sharing in a long-term supply chain relationship: The role of customer review strategy," *Operations Research*, vol. 58, no. 1, pp. 81–93, 2010.

[161] P. Rey and J. Stiglitz, "The role of exclusive territories in producers' competition," *The RAND Journal of Economics*, vol. 26, no. 3, pp. 431–451, 1995.

[162] P. Rey and T. Verge, "Resale price maintenance and interlocking relationships," *The Journal of Industrial Economics*, vol. 58, no. 4, pp. 928–961, 2010.

[163] W. P. Rogerson, "The first-order approach to principal-agent problems," *Econometrica*, vol. 53, no. 6, pp. 1357–1367, 1985.

[164] R. E. Romano, "Double moral hazard and resale price maintenance," *The RAND Journal of Economics*, vol. 25, no. 3, pp. 455–466, 1994.

[165] M. Salinger and M. Ampudia, "The simple economics of the price-setting newsvendor problem," *Management Science*, 2011.

[166] D. Sappington, "Limited liability contracts between principal and agent," *Journal of Economic Theory*, vol. 29, no. 1, pp. 1–21, 1983.

[167] D. Scharfstein, "Product-market competition and managerial slack," *The RAND Journal of Economics*, vol. 19, no. 1, pp. 147–155, 1988.

[168] F. M. Scherer and D. Ross, *Industrial Market Structure and Economic Performance.* Chicago: Rand McNally, 1990.

[169] I. R. Segal and M. D. Whinston, "Naked exclusion: Comment," *American Economic Review*, vol. 90, no. 1, pp. 296–309, 2000.

[170] I. R. Segal and M. D. Whinston, "Property rights," in *Handbook of Organizational Economics*, (R. Gibbons and J. Roberts, eds.), Princeton University Press, 2012. Forthcoming.

[171] S. Severinov, "An efficient solution to the informed principal problem," *Journal of Economic Theory*, vol. 141, no. 1, pp. 114–133, 2008.

[172] G. Shaffer, "Slotting allowances and resale price maintenance: A comparison of facilitating practices," *The RAND Journal of Economics*, vol. 22, no. 1, pp. 120–135, 1991.

[173] C. Shapiro, "Premiums for high quality products as returns to reputations," *The Quarterly Journal of Economics*, vol. 98, no. 4, pp. 659–680, 1983.

[174] C. Shapiro, "Exchange of cost information in oligopoly," *Review of Economic Studies*, vol. 53, no. 3, pp. 433–446, 1986.

[175] C. Shapiro and J. E. Stiglitz, "Equilibrium unemployment as a worker discipline device," *The American Economic Review*, vol. 74, no. 3, pp. 433–444, 1984.

[176] S. Shavell, "On moral hazard and insurance," *Quarterly Journal of Economics*, vol. 93, no. 4, pp. 541–562, 1979.

[177] J. Simpson and A. L. Wickelgren, "Naked exclusion, efficient breach, and downstream competition," *American Economic Review*, vol. 97, no. 4, pp. 1305–1320, 2007.

[178] A. Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations.* 1776.

[179] G. Sosic, "Transshipment of inventories among retailers: Myopic vs. farsighted stability," *Management Science*, vol. 52, no. 10, pp. 1493–1508, 2006.

[180] M. Spence, "Signaling in retrospect and the informational structure of markets," *The American Economic Review*, vol. 92, no. 3, pp. 434–459, 2002.

[181] J. Spengler, "Vertical integration and anti-trust policy," *Journal of Political Economy*, vol. 58, no. 4, pp. 347–352, 1950.

[182] X. Su, "Intertemporal pricing with strategic customer behavior," *Management Science*, vol. 53, no. 5, pp. 726–741, 2007.

[183] X. Su, "Intertemporal pricing and consumer stockpiling," *Operations Research*, vol. 58, no. 4, Part 2 of 2, pp. 1133–1147, 2010.

[184] X. Su and F. Zhang, "Strategic customer behavior, commitment, and supply chain performance," *Management Science*, vol. 54, no. 10, pp. 1759–1773, 2008.

[185] R. Swinney, "Selling to strategic consumers when product value is uncertain: The value of matching supply and demand," *Management Science*, 2011.

[186] R. Swinney and G. P. Cachon, "Purchasing, pricing, and quick response in the presence of strategic consumers," *Management Science*, vol. 55, no. 3, pp. 497–511, 2009.

[187] R. Swinney and G. P. Cachon, "The value of fast fashion: Quick response, enhanced design, and strategic consumer behavior," *Management Science*, vol. 57, no. 4, pp. 778–795, 2011.

[188] T. A. Taylor, "Supply chain coordination under channel rebates with sales effort effects," *Management Science*, vol. 48, no. 8, pp. 992–1007, 2002.

[189] L. G. Telser, "Why should manufacturers want fair trade?," *Journal of Law and Economics*, no. 3, no. 3, pp. 86–105, 1960.

[190] J. Tirole, *The Theory of Industrial Organization*. Cambridge, MA: MIT Press, 1988.

[191] J. Tirole, "Incomplete contracts: Where do we stand?," *Econometrica*, vol. 67, no. 4, pp. 741–781, 1999.

[192] G. Tullock, "Efficient rent seeking," in *Toward a Theory of the Rent-Seeking Society*, (R. T. J. Buchanan and G. Tullock, eds.), pp. 97–112, College Station: Texas A&M University Press, 1980.

[193] T. I. Tunca and S. A. Zenios, "Supply auctions and relational contracts for procurement," *Manufacturing & Service Operations Management*, vol. 8, no. 1, pp. 43–67, 2006.

[194] US Department of Commerce, "Franchising in the Economy: 1986–88," 1988.

[195] H. Varian, *Microeconomic Analysis*. 1992.

[196] H. R. Varian, "Price discrimination," in *Handbook of Industrial Organization*, vol. 1, (R. Schmalensee and R. Willig, eds.), pp. 597–654, Elsevier, 1989.

[197] X. Vives, "Trade association disclosure rules, incentives to share information, and welfare," *The RAND Journal of Economics*, vol. 21, no. 3, pp. 409–430, 1990.

[198] Y. Wang, "Joint pricing-production decisions in supply chains of complementary products with uncertain demand," *Operations Research*, vol. 54, no. 6, pp. 1110–1127, 2006.

[199] F. R. Warren-Boulton, "Vertical control with variable proportions," *Journal of Political Economy*, vol. 82, no. 4, pp. 783–802, 1974.

[200] O. Williamson, "Outsourcing: Transaction cost economics and supply chain management," *Journal of Supply Chain Management*, vol. 44, no. 2, pp. 5–16, 2008.

[201] O. E. Williamson, "The vertical integration of production: Market failure considerations," *The American Economic Review*, vol. 61, no. 2, pp. 112–123, 1971.

[202] O. E. Williamson, *Markets and Hierarchies: Analysis and Antitrust Implications*. 1975.

[203] O. E. Williamson, “Transaction-cost economics: The governance of contractual relations,” *Journal of Law and Economics*, vol. 22, no. 2, pp. 233–261, 1979.

[204] O. E. Williamson, *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. 1985.

[205] R. B. Wilson, *Nonlinear Pricing*. US: Oxford University Press, 1997.

[206] R. A. Winter, “Vertical control and price versus nonprice competition,” *The Quarterly Journal of Economics*, vol. 108, no. 1, pp. 61–76, 1993.

[207] R. A. Winter, “Presidential address: Antitrust restrictions on single-firm strategies,” *Canadian Journal of Economics*, vol. 42, no. 4, pp. 1207–1239, 2009.

[208] J. Wright, “Exclusive dealing and entry, when buyers compete: Comment,” *The American Economic Review*, vol. 99, no. 3, pp. 1070–1081, 2009.

[209] D. J. Wu and P. R. Kleindorfer, “Competitive options, supply contracting, and electronic markets,” *Management Science*, vol. 51, no. 3, pp. 452–466, 2005.