The Quantity Flexibility Contract and Supplier-Customer Incentives

Andy A. Tsay
Department of Operations & Management Information Systems, Leavey School of Business, Santa Clara University, Santa Clara, California 95053–0382
atsay@scu.edu

Consider a supply chain consisting of two independent agents, a supplier (e.g., a manufacturer) and its customer (e.g., a retailer), the latter in turn serving an uncertain market demand. To reconcile manufacturing/procurement time lags with a need for timely response to the market, such supply chains often must commit resources to production quantities based on forecasted rather than realized demand.

The customer typically provides a planning forecast of its intended purchase, which does not entail commitment. Benefiting from overproduction while not bearing the immediate costs, the customer has incentive to initially overforecast before eventually purchasing a lesser quantity. The supplier must in turn anticipate such behavior in its production quantity decision. This individually rational behavior results in an inefficient supply chain.

This paper models the incentives of the two parties, identifying causes of inefficiency and suggesting remedies. Particular attention is given to the Quantity Flexibility (QF) contract, which couples the customer's commitment to purchase no less than a certain percentage below the forecast with the supplier's guarantee to deliver up to a certain percentage above. Under certain conditions, this method can allocate the costs of market demand uncertainty so as to lead the individually motivated supplier and customer to the systemwide optimal outcome. We characterize the implications of QF contracts for the behavior and performance of both parties, and the supply chain as a whole.

(*Supply Chain Management*; *Supply Contracts*; *Quantity Flexibility*; *Coordination*; *Forecasting*; *Forecast Revision*; *Materials Planning*)

1. Introduction

1.1. The Setting

Decentralized control is a reality for many supply chains for various reasons. For instance, outsourcing of production to independently held entities, which automatically distributes decision-making authority, is currently a popular business model in many industries (cf. Farlow et al. 1995, Iyer and Bergen 1997). Even for highly vertically integrated firms, today's characteristically global business environments often result in multiple sites worldwide working together to deliver product, while reporting to different organizational functions or units within the corporation. Operational control of these sites may be intentionally decentralized for informational or incentive considerations. In this paper we will consider an external manufacturer (EM) that provides a product to a retailer, which in turn serves an end market. The discussion will also apply to any two consecutive links further upstream in the supply chain that are independently managed, whether they are formally distinct firms or simply behave as such.

To reconcile manufacturing/procurement time lags

1526-5501/99/4510/1339$05.00

with a need for timely response to the market, such supply chains often must commit resources to production quantities based on forecasted rather than realized demand. The immediate impact of this commitment falls on the EM since it perceives the direct costs of installing production capacity and securing raw materials. Because the market demand is uncertain, the retailer prefers to postpone any advance commitment to purchase, ideally pulling finished product forward only as a response to confirmed demand. In fact, the possibility of offloading the risk of overestimating demand to another party may partially motivate some outsourcing decisions. However, this brings a different risk in that some control is relinquished to a party who may not perceive the same concern with serving the end customer.

One way such relationships have traditionally been managed is for the retailer to provide an initial point estimate of its intended purchase to assist the EM's production quantity decision. However, both parties are aware that the retailer's eventual purchase will likely differ from this planning forecast, in part due to the natural resolution of demand uncertainty over time, but also because the estimate may have been colored by the retailer's individual preferences towards overproduction and underproduction. A careful EM will adjust for these incentives, and will also incorporate its own economic prospects. The final outcome will depend largely on how the costs of demand uncertainty are allocated. In some relationships, the retailer has such strategic power that it can expect the EM to fully cover the forecast while reserving the prerogative to completely back out of the purchase. If the balance of power lies at the other extreme, the EM may be able to hold the retailer to the forecast. We will demonstrate that while each of these scenarios might be preferred by one of the players, both lead to an inefficient outcome for the overall system.

Various remedies to these inefficiencies have been attempted, as noted in §2. This paper considers specifically the Quantity Flexibility (QF) contract, which couples the retailer's commitment to purchase no less than a certain percentage below the forecast (a minimum purchase agreement) with the EM's guarantee to deliver up to a certain percentage above. There is also a single procurement or "transfer" price to be charged per unit of product delivered. Under certain conditions, with appropriate choice of these contract parameters, managing the decentralized arrangement via QF contract can achieve systemwide efficiency while the individual decision makers follow their own best interests. Instances of this contract from industrial practice are documented in the next section.

1.2. Instances of Quantity Flexibility Contracts in Industry

The emergence of QF contracts as a response to certain supply chain inefficiencies is described in Lee et al. (1997). Sun Microsystems uses QF contracts in its purchase of various workstation components (cf. Farlow et al. 1995). Nippon Otis, a manufacturer of elevator equipment, implicitly uses such contracts with Tsuchiya, its supplier of parts and switches (cf. Lovejoy 1999). Solectron, a leading contract manufacturer for many electronics firms, has recently installed such agreements with both its customers and its raw materials suppliers (Ng 1997), implying that benefits may accrue to either end of such a contract. QF-type contracts have also been used by Toyota Motor Corporation (Lovejoy 1999), IBM (Connors et al. 1995), Hewlett Packard, and Compaq (Faust 1996). A similar structure, called a "Take-or-Pay" provision, is often embedded in long-term supply contracts for natural resources (cf. Masten and Crocker 1985, Mondschein 1993, National Energy Board 1993). In addition to governing relations between separate companies, QF structures have also appeared at the interface between the manufacturing and marketing/sales functions (taking the role of supplier and buyer, respectively) within single firms (cf. Magee and Boodman 1967).

Little formal documentation exists describing how specific flexibility parameters and transfer prices have been arrived upon by these industrial users. As with many other clauses of supply contracts, many firms are unwilling to detail even seemingly innocuous terms of trade. Any revelation of variance across their multiple suppliers (or customers) can create a perception of favoritism, which may invite ill will or even legal scrutiny. However, interviews by the author with several of the listed companies have elicited

some insights about contract specification. One key consideration is the perceived amount of uncertainty in forecasting end product demand, which certainly drives the order revision that propagates up the supply chain. A related issue is the fungibility of the item. For instance, one computer manufacturer enjoys $\pm 25\%$ per month revision flexibility with its supplier of fairly customized monitors, while settling for $\pm 5\%$ on commodity-like keyboards. However, the existing contracts have apparently thus far been negotiated based on managerial judgment rather than any formal analysis. One of the objectives of this paper is to inform such decisions by describing the impact of the contract parameters on the economic outcomes for each party involved. See Tsay (1995) for more detailed discussion of the industrial implementation of QF contracts.

1.3. Organization of this Paper

Section 2 positions this paper in the literature. Section 3 presents the analytical model, and §4 states the optimal performance benchmark. Section 4 characterizes the decentralized system outcome in absence of some additional control structure, and identifies the causes of inefficiency. The discussion considers both possible informational scenarios concerning market demand: information asymmetry and common beliefs. Section 6 postulates the Quantity Flexibility contract, then derives the behavior that it induces. Such behavior is not always efficient, and the author submits as evidence two extreme (but empirically observed) supply relationships which may in fact be modeled as special cases of the QF contract. Section 7 shows how the choice of flexibility parameters and transfer price can discourage inefficient behavior, even attaining full efficiency under certain conditions. Properties of the efficient parameters and their role in allocating efficiency gains are developed. Section 8 contains numerical exploration which corroborates the preceding results and provides additional insights. The author concludes in §8 with interpretive discussion and some implementation issues. All proofs are omitted due to space limitations, but are available from the author in an Appendix that has gone through the *Management Science* review process.

2. Literature Review

It is not generally the case that a supply chain composed of independent agents acting in their own best interests will achieve systemwide efficiency, often due to some incongruence between incentives faced locally and the global optimization problem. For example, the financial terms of many supply relationships are such that the overstock and understock risks perceived by the supply chain as a whole are visited differently upon the individual parties, a phenomenon known in the economics literature as "double marginalization" (cf. Spengler 1950, Tirole 1988).

One response is to reconsider the nature of the supply contracts along the chain (see Tsay et al. 1999 for a recent review). The general goal is to install rules for materials accountability and/or pricing that will guide autonomous entities towards the globally desirable outcome (cf. Whang 1995). This type of approach recurs in a broad range of settings, including the economic literature on "vertical restraints" (cf. Mathewson and Winter 1984, Katz 1989), the marketing literature of "channel coordination" (e.g., Jeuland and Shugan 1983, Moorthy 1987), and agency theory (cf. Bergen et al. 1992, Van Ackere 1993). Recent examples in the multiechelon inventory literature include Lee and Whang (1997), Chen (1997), and Iyer and Bergen (1997). Of these three, the last is most relevant to our model, as described below.

Iyer and Bergen (1997) model Quick Response (QR) in a manufacturer-retailer supply chain, abstracted as a delay of the supply chain's commitment to quantity. The retailer benefits from procuring under improved information, yet the manufacturer can be made worse off. Since this manufacturer is assumed to produce to order, its payoff is determined once the retailer orders, regardless of how the uncertain market demand eventually resolves. So the manufacturer naturally prefers a large retailer order, even if this includes excess amounts of safety stock that never get sold, and may oppose any process improvement that enables reduction of such safety stock. The authors use this to explain various side-agreements that have been observed to accompany QR efforts, such as requirements for higher service to the end customer, wholesale price increases, or volume commitments across multiple

products. These all force the retailer to buy and/or pay more than it would with QR alone, enough to preserve the manufacturer's original profit. In their emphasis on Pareto improvement, even at the cost of systemwide or local optimality, Iyer and Bergen showcase the importance of individual incentives in implementing supply chain reform, a view sympathetic to the focus of this paper.

In research of more immediate relevance, the quantity ultimately obtained may differ from an estimate made under prior demand information. A class of coordination mechanism that has arisen in response to the reality of order revision is described in the contracting literature of law and economics as "relational contracts" (Masten and Crocker 1985, Crocker and Masten 1991). These formally establish the relationship, but intentionally defer precise decisions about price, quantity, or other aspects of the exchange. The usage of such contracts acknowledges that because information changes over time, a control structure that preserves the ability to act on new information can be advantageous, with respect not only to the usual performance metrics (holding and shortage costs) but also to transactions costs. Examples of papers which explicitly model variants of this type of relationship are described below, segmented into two general categories: those that focus primarily on the buyer's decision problem, followed by those that consider both parties.

Bassok and Anupindi (1995) investigate forecasting and purchasing behavior when the buyer initially forecasts month-by-month demand over an entire year and then may revise each month's purchase once within specified percentage bounds. Bassok and Anupindi (1997a) analyze a contract which specifies that cumulative purchases over a multiperiod horizon must exceed a previously (and exogenously) specified quantity, a form of minimum-purchase agreement. Bassok and Anupindi (1997b) study a rolling-horizon flexibility contract similar to our QF structure, focusing on the retailer's ordering behavior when facing an independent and stationary market demand process. Eppen and Iyer (1997) analyze "backup agreements" in which the buyer is allowed a certain backup quantity in excess of its initial forecast at no premium, but pays a penalty for any of these units not purchased. A similar demand environment is considered by Fisher and Raman (1996), with a retailer affecting supply flexibility by commissioning two-stage production. The initial run covers the early part (~20%) of the selling season, whose sales inform a second run that covers the rest of the season. An exogenous constraint on second-run capacity forces some production to the riskier early commitment.

The preceding papers are primarily single-node models in that they take the perspective of the party exercising the purchase quantity flexibility. While an upstream supplier may be mentioned, its role is passive. This allows analytical consideration of more complex settings, since no attention need be given to how the supplier supports the specified flexibility or the impact on its costs. Furthermore, the issue of how information flows to the upstream party, which is rendered challenging by updating of beliefs about market demand, need not be explicitly addressed. However, a drawback is that the supplier's preferences towards the supply arrangement remain indeterminate. Recent papers that formally model the perspectives of both parties to the flexible supply relationship, which is necessary for evaluating its efficacy in coordinating a decentralized supply chain, are described below. Most are variants of the newsvendor model, and rely at some level on common beliefs about market demand.

Pasternack (1985) analyzes a manufacturer-retailer relationship analogous to the special case considered in our Proposition 6, and determines that coordination can be achieved by allowing the retailer to return all surplus at a partial refund. A key result is that the manufacturer can determine the efficient prices *without knowing the market demand distribution*. However, common information is still requisite for implementation because otherwise the manufacturer cannot properly evaluate each party's expected profits and therefore remains unable to allocate the efficiency gains in a way that will ensure the retailer's participation. Kandel (1996), Ha (1997), and Emmons and Gilbert (1998) obtain similar results when Pasternack's setting is generalized with price-sensitive retail demand. Donohue (1998) studies these contracts with a two-stage decision model similar to ours. First, the buyer commits to a quantity and a wholesale price.

After observing early demand and performing a Bayesian update on the total season's demand, the buyer can place an additional order at a different wholesale price. At the end of the season, the manufacturer takes back any unsold items at a third price. The analysis determines values for the three prices that will lead to system efficiency.

The above examples show that returns policies are efficient only when the buyer enjoys unbounded flexibility. A price premium for exercising that flexibility is then included to deter abusive overordering. In contrast, the QF contract provides flexibility with no explicit penalty for exercise, but uses constraints as a way to motivate appropriate behavior. Only a fraction of the order may be canceled with no financial consequence; the cost of the remainder is what induces more careful ordering. In practice, there may be qualitative reasons why one contractual form is more palatable than the other in particular settings, as discussed in §8 (see also Lariviere 1999). But this subliterature and our work together support a key conclusion, that (under the assumption of common information) efficiency can be restored with either pricing or constraints.

A number of other papers provide more detailed models of production planning within both the supplier and the buyer organizations in which flexibility plays some role. Parlar and Weng (1997) model supply and manufacturing departments that behave independently under a nonlinear internal transfer pricing arrangement. The supply department procures at least as much material as is requested by manufacturing for an initial production run. However, it may purchase more to achieve an economy of scale in procurement, and because of its own beliefs regarding any materials needed for an optional second run, for which it may charge a higher price. Determining that inefficiency will result from decentralized decision making, this work argues for the benefits of coordination via information sharing and collaborative decision making. However, no specific coordination mechanism is proposed or evaluated.

Barnes-Schuster et al. (1998) discuss options for supplier capacity as a means of affecting flexibility for the buyer. After observing demand in the first of two periods, the buyer may exercise some of the options (at an additional fee), or let them expire, thereby losing the original option cost but avoiding any expense that would have been incurred had actual production been commissioned initially. The supplier is obligated to position raw material to the maximum buyer request (firm orders + options), but might not convert it all to finished goods if the demand signal does not warrant it. Tsay and Lovejoy (1999) examine the QF contract in a multiperiod, rolling horizon context. In addition to the customer's ordering policy, they describe an operating policy by which a supplier might economically uphold its end of the supply contract. Both these papers provide frameworks for valuing the parameters of the respective contracts, but neither speaks to the issue of efficiency since no optimal benchmark is currently available in either case.

Weng (1997) considers a manufacturer-distributor supply chain facing price-sensitive stochastic demand, where the decision variables are the distributor's order quantity (which equals the manufacturer's production, since production is make-to-order) and selling price, and the manufacturer's wholesale price. The model assumes that any excess demand must be completely satisfied via a second, more costly production run (much like Donohue 1998, Fisher and Raman 1996, and Parlar and Weng 1997). This turns out to be a key assumption since the total distributor order is then exactly the market demand, and the recourse quantity need not be treated as a distinct decision variable. Coordination can then be achieved by setting the wholesale price for the output of each run exactly equal to the run's production cost, accompanied by a lump-sum payment to the manufacturer. This is intuitive since the distributor then perceives the overage-underage cost structure of the system as a whole, and will make the system-optimal decisions. The side-payment is necessary to ensure the other party's participation. Addressing double-marginalization in a variety of settings by imposing the cost structure of the entire supply chain upon the decision-making party has been proposed by various other authors, including Mathewson and Winter (1984), Moorthy (1987), and Ha (1997).

We summarize the positioning and contribution of this paper as follows. We study the setting of quantity

revision in response to improved demand information. Rather than assuming a passive supplier who simply accommodates the customer's actions, we develop a behavioral model of each party's local incentives. We determine the circumstances under which inefficiency results, and then examine a contractual form that has been observed in industry. We characterize each party's preferences towards the contract parameters, and define when system efficiency can be achieved. As noted above and elaborated upon in §8, there will be settings in which each of the various efficiency-preserving contracts may be more desirable. Our purpose is not necessarily to advocate the QF contract, but to provide insight into the mechanism by which it alters supply chain incentives. This analysis will offer rigorous conclusions about the implications of QF usage and enable comparison to other methods of supply chain coordination.

3. The Model

3.1. Cost Structure

We model the scenario of interest by modifying the newsvendor framework to represent a two-player setting. As such, our cost parameters are the newsvendor parameters augmented with a price per unit of product transferred between the two players:

p = retail price per unit of finished good (collected from the market by the retailer);

c = unit wholesale or "transfer" price (paid by the retailer to the EM);

m = unit production cost (incurred by the EM for raw materials, labor, etc.);

u = salvage value per unit of product not consumed by the retail market; the product has the same salvage value/cost regardless of the ownership;

s = any goodwill loss (in excess of foregone profit) per unit by which market demand exceeds available finished goods;[1] we assume that such demand results in lost sales.

[1] Only the retailer experiences any goodwill losses. This is commonly assumed in multistage inventory models (e.g., Clark and Scarf 1960), and is consistent with the assessment of many managers that end customers tend to blame only the final link in the supply chain for stockouts regardless of where the fault ultimately lies.

The following straightforward assumptions are required of the cost parameters to assure internal consistency: (i) $p > c > m > 0$, (ii) $u < m$, (iii) $s \geq 0$. Parameter values are common knowledge, and all are exogenous to the model except c.

3.2. Decision Structure

The chronology of events, notation, and information structure are as follows:[2]

1. The terms of the supply contract are negotiated.
2. The retailer states q_j, an initial forecast of its purchase quantity.
3. The EM builds the production quantity Q_j, thus determining $\pi_{EM,j}$ and $\pi_{R,j}$, the expected profits of the respective parties. All decisions up to this point are based on the prior distribution of market demand X.
4. μ, a signal about market demand, is observed.
5. The retailer purchases r_j, based on the updated information $X|\mu$. This is the amount of material available to meet market demand. Any EM surplus is salvaged.
6. Market demand X is revealed, and is filled to the extent possible by the retailer's stock. Any retailer surplus is salvaged.

As alluded to in §1, the choice of r_j follows the revelation of μ to signify that the retailer's acceptance of product is based on information different from that used by the EM in planning production. Specifics of the information model are presented in §3.3.

We will obtain the outcome of this multiparty decision problem by backwards induction, assuming that the decision maker at each step acts optimally, given what has already transpired, and anticipating likewise optimal behavior by the decision maker in each subsequent step. In particular, the solution may be obtained by solving the following sequence of optimization problems that considers the decision variables in reverse chronological order:

(I) The retailer's actual purchase is made after q has been stated, Q has been produced, and μ has been observed. So $r_j^*(q_j, Q_j, \mu) \equiv \text{argmax}_r\{G(r|\mu)\}$ *s.t.* r

[2] Subscripting identifies the control scheme being considered: j takes values CC, NC, and QF, to denote control by a central planner (§4), a no-commitment arrangement (§4), and the QF contract (§§5 and 7), respectively.

$\leq Q_j$ *and* {*any other constraint*(*s*) *on r due to the supply arrangement*}, where $G(r|\mu)$ is the retailer's expected profit conditional on the observed μ. (This is not to be confused with the retailer's expected profit from the time 0 perspective, as described in (III).) The impact of q_j, if any, will be through the second constraint, which will vary with the different arrangements we consider.

(II) The EM commits to production prior to the forecast update, according to the rule $Q_j^*(q_j) \equiv \text{argmax}_Q\{\pi_{\text{EM},j}(Q;\ q_j,\ r_j^*(q_j,\ Q_j,\ \mu))\}$ *s.t.* {*any constraint on* Q_j *due to the supply arrangement*}. $\pi_{\text{EM},j}(\ )$ anticipates the possible outcomes of μ, and therefore the retailer's adjustment. Thus, computation of $\pi_{\text{EM},j}(\ )$ requires unconditioning against the distribution of μ.

(III) The retailer states q_j with knowledge of the EM's subsequent production response (from (II)) and its own purchasing policy (from (I)). That is, $q_j^* \equiv \text{argmax}_q\{\pi_{R,j}(r_j^*(q,\ Q_j^*(q),\ \mu))\}$. As in Problem (II), embedded in the objective is an unconditioning of a profit function that is conditional on μ.

Individual Rationality/Participation constraints (cf. Van Ackere 1993) are also relevant, but will be treated explicitly later.

3.3. Demand Structure

We assume random market demand of the form $X = \mu + \epsilon$, where μ represents an unknown location parameter and ϵ is an independent error.[3] μ has mean $\bar{\mu}$, variance σ_μ^2, and distribution function $\Theta(\ )$ which is differentiable and invertible; ϵ is normal with zero mean, variance σ_ϵ^2, and distribution $\Gamma(\ )$. X has mean $\bar{\mu}$ and variance $\sigma_X^2 = \sigma_\mu^2 + \sigma_\epsilon^2$, and distribution $F(\ )$ (the convolution of $\Theta(\ )$ and $\Gamma(\ )$). $\Phi(\ )$ will be used to denote the standard normal distribution. We presume mean and variance values such that μ and X are both almost certainly nonnegative (e.g., $\bar{\mu} \geq 3\sigma_X$). Additional structure will be assumed as necessary. This is similar to the demand model of Iyer and Bergen (1997). However, in their decision model the manufacturer's commitment point is identical to the retailer's, whereas our EM must commit at a particular time, no matter when the retailer's order becomes firm.

[3] Virtually all results in this paper also hold for the multiplicative form $X = \mu \cdot (1 + \epsilon)$. We focus on the additive model for expositional clarity.

As is common in the supply contracting literature, we pursue as simple a model as possible to focus attention on key features of interest. The above model has sufficient richness to capture the central concerns, namely the unfolding of information over time and differences in ability to obtain product as a function of the degree of advance warning.

4. Central Control

To provide an efficiency benchmark, we suppose first that the manufacture and retail are coordinated by a single entity, which delivers the greatest possible expected system profit. This is precisely a standard newsvendor problem with underage and overage costs of $(p + s - m)$ and $(m - u)$, respectively. The appropriate demand distribution is that of X since the production commitment precedes the information update, leading to a unique optimal production of $Q_{CC}^* = F^{-1}((p + s - m)/(p + s - u))$. Any quantity different from this, while potentially preferable to one party or the other, will guarantee system inefficiency. Since there is no intermediate transfer, the constructs q_j and r_j, as well as the transfer price c, play no role here.

5. Decentralized Control with No Commitment

We now postulate a supply chain composed of an independent EM and retailer. Alternatively we may think of two divisions of a single firm, such as operations and marketing, which are managed to selfish, rather than firmwide, objectives. We will consider two cases which differ in their informational assumptions. First, we discuss in §5.1 the more general setting in which the true statistics of market demand are the retailer's private information. Subsequent analysis in this paper will proceed under the assumption of shared beliefs, as outlined in §5.2.

5.1. Asymmetric Information about Market Demand

In this section we suppose the EM does not share the retailer's visibility of the market demand, and is left to form its own beliefs about the retailer's purchasing behavior. The retailer provides a forecast of its intended purchase, which may influence the EM's deci-

sion. As a matter of terminology, we distinguish between the demands encountered by each player. "Market demand" is what the retailer experiences, as characterized in §3.3. Meanwhile, the EM faces a demand from the retailer that will be driven by, but need not be identical to, the market demand.

The retailer's desired purchase quantity after observing μ is invariant to the EM's beliefs. It results from a newsvendor problem in which the relevant demand distribution is that of $X|\mu$ and the transfer price represents the product cost, i.e., $\max_r\{G(r|\mu)\}$ s.t. $r \le Q_{\text{NC}}$ where

$$\begin{aligned} G(r|\mu) &= E_{X|\mu}\{p \cdot \min[X, r] \\ &\qquad - c \cdot r - s[X - r]^+ + u[r - X]^+\} \\ &= (p + s - c)r - sE_{X|\mu}\{X\} \\ &\quad - (p + s - u)E_{X|\mu}\{[r - X]^+\}, \end{aligned} \tag{1}$$

and Q_{NC} is the EM production.[4] As $X|\mu$ is Normal(μ, σ_ϵ^2), the desired purchase absent the constraint is $(\mu + z_\epsilon\sigma_\epsilon)$, where $z_\epsilon \equiv \Phi^{-1}((p + s - c)/(p + s - u))$. The actual purchase will then be $r^*_{\text{NC}}(Q_{\text{NC}}, \mu) = \min[\mu + z_\epsilon\sigma_\epsilon, Q_{\text{NC}}]$.

We represent the EM's beliefs about the retailer's purchase with a distribution function $\Lambda_q(\,)$, subscripted with q to suggest a (likely positive) dependence on the retailer's forecast q_{NC}. The EM's decision problem also has newsvendor structure, except with underage and overage costs of $(c - m)$ and $(m - u)$, respectively. So $Q^*_{\text{NC}} = \Lambda_q^{-1}((c - m)/(c - u))$.

A retailer whose q_{NC} is devoid of economic consequence will state arbitrarily large values in hopes of inducing overproduction, only to purchase more realistically when finally required to place a firm order. Purchasing managers at one electronics company have admitted to the author the regularity of such behavior. Lee et al. (1997) also documents several prominent case studies of this so-called "phantom ordering." A careful EM will anticipate this but must still ascertain how best to deflate the exaggerated numbers so as to avoid overcapacity and inventory. The upshot of this gaming is a production decision made under distorted information, which increases system costs and uncertainties (cf. Lovejoy 1999).

The expression for Q^*_{NC} reveals multiple causes of inefficiency:

(i) The retailer has incentive to bias $\Lambda_q(\,)$ away from reality by exaggerating q_{NC}.

(ii) $\Lambda_q(\,)$ is presumably of lower informational quality to begin with, since the EM is not privy to the retailer's information about market demand.

(iii) The EM is positioning not to the underlying market demand, but to the retailer's purchase behavior. This will derive from the retailer's local overage and underage costs, as opposed to those that the system as a whole perceives.

(iv) The EM's critical fractile is in turn based on its individual overage and underage costs.

The first two factors are due to information asymmetry. An issue similar to the second has been studied by Parlar and Weng (1997) to make a case for information sharing, and Atkinson (1979), who argues that the decision-making authority should rest with the party with the best information. The last two issues are due to the incentives resulting from decentralization and are related to double marginalization (see §2). The retailer's ordering behavior reflects the retailer's cost of product (i.e., the transfer price), instead of the true cost to the system. Likewise, in planning production the EM considers only the revenue it receives per unit (the transfer price), rather than the true revenue for the system. In light of these factors, we conclude that Q^*_{NC} will match the system-optimal quantity only by sheer mathematical coincidence. No further characterization of this setting is attempted here, as virtually any outcome can be produced depending on the assumed form of $\Lambda_q(\,)$ and how q_{NC} is incorporated.

5.2. Common Beliefs about Market Demand

There has recently been much discussion in both industry and academia about the benefits of sharing demand information among supply chain partners, and efforts in this direction are a growing trend (cf. Kumar 1996, Verity 1996, Lee et al. 1997). With this motivation, we now assume that common beliefs about market demand have been achieved (e.g., through collaborative analysis of shared market data), so as to focus on the efficiency implications of the

[4] NC stands for "No Commitment."

existing incentive structures. The key realization is that forecasting and inventory management are distinct activities, and collaboration in one area does not necessarily solve whatever problems might be caused by independent behavior in the other.

While there may be contracts in which the computation of the channel-coordinating parameters does not rely on the distribution of market demand (e.g., the returns policies studied by Pasternack 1985 and others), the assumption of common beliefs underlies the very notion of systemwide efficiency. If an uncertain market demand is viewed using different distributional assumptions, the evaluation of expected profit becomes a matter of opinion. Whether a particular plan is system-optimal at all becomes controversial, and individual rationality/participation constraints are problematic as the party driving the negotiation must still know the other party's beliefs in order to anticipate that party's calculation of expected profit. This modeling challenge is likely why virtually all extant works studying incentives in multiplayer supply chains (including all mentioned in §2) ultimately rely at some level on common beliefs about market demand (cf. Tsay et al. 1999).

If the retailer's forecast implies no formal commitment for either party, then it will not affect the demand encountered by the EM. It would therefore be irrational for the EM to adapt to the retailer's forecast since there is no tangible benefit from doing so. The production decision then rests solely with the EM, which, absent any further incentives will build according to its own economic prospects. As before, the EM solves a newsvendor problem, except that the distribution of the demand faced (i.e., the retailer's purchase) must be imputed from the statistics of μ and ϵ. This is made explicit below.

As established in §5.1, after observing μ the retailer will purchase $r^*_{NC}(Q_{NC}, \mu) = \min[\mu + z_\epsilon\sigma_\epsilon, Q_{NC}]$. This is anticipated by the EM's production decision, now made with respect to the correct distribution of $(\mu + z_\epsilon\sigma_\epsilon)$ instead of $\Lambda_q(\,)$. By newsvendor analysis the production outcome will be $Q^*_{NC} = \Theta^{-1}((c - m)/(c - u)) + z_\epsilon\sigma_\epsilon$. Under these assumptions, we can state unequivocally that underproduction occurs and hence inefficiency results.[5]

Proposition 1. *If the forecast of the retailer's purchase represents no commitment for either party, then for any c, underproduction occurs relative to the optimal centralized solution (i.e., $Q^*_{NC} < Q^*_{CC}$). Therefore, expected total system profit is strictly suboptimal.*

This result is due to the last two factors listed in §5.1: double marginalization and the fact that the EM is positioning not to the true market demand, but to the retailer's ordering policy.

Even with common information about market demand, while adjusting c can determine how profits are allocated, there is no c for which systemwide efficiency can be attained. From this we conclude the potential benefit to both parties of mitigating these types of behaviors, hence the viability of some contractual structure beyond a simple linear transfer price. The QF contract is one such structure whose properties we will examine next.

6. The Quantity Flexibility (QF) Contract

In the QF framework, the supply relationship between the EM and the retailer is parameterized by $\{c, (\alpha, \omega)\}$. c is the unit transfer price, the EM guarantees product availability of up to $q_{QF}(1 + \alpha)$, and the retailer must purchase at least $q_{QF}(1 - \omega)$. $\omega \in [0, 1]$ and $\alpha \in [-\omega, \infty)$.[6]

6.1. The Equilibrium Solution

The equilibrium is obtained by backwards induction, as outlined in §3.2. By analogy to §4, the retailer will ultimately purchase $r^*_{QF}(q_{QF}, Q_{QF}, \mu) = (\mu + z_\epsilon\sigma_\epsilon) \perp [q_{QF}(1 - \omega), Q_{QF}]$.[7] The EM's production is

$$Q^*_{QF}(q_{QF}) \equiv \operatorname{argmax}_{Q \geq q_{QF}(1+\alpha)} \{\pi_{EM,QF}(Q; q_{QF}, r^*_{QF}(q_{QF}, Q, \mu))\},$$

[5] This result does not require that ϵ be normally distributed.

[6] $\alpha \in [-\omega, 0)$ implies that the quantity the EM guarantees is less than the retailer's forecast. This is still feasible as the maximum coverage remains greater than the minimum purchase.

[7] $y \perp [a, b]$ denotes the point closest to y in an interval $[a, b]$.

where

$$\pi_{EM,QF}(Q; q_{QF}, r^*_{QF}(q_{QF}, Q, \mu)) = (c - u)E_\mu\{r^*_{QF}(q_{QF}, Q, \mu)\} - (m - u)Q \quad (2)$$

is the EM's expected profit. While $Q^*_{QF}(q_{QF})$ will be no less than $q_{QF}(1 + \alpha)$ by the terms of the contract, a priori there appears to be no reason the EM might not unilaterally choose to produce $Q^*_{QF}(q_{QF}) > q_{QF}(1 + \alpha)$ to take advantage of the possibility of additional sales. In Proposition 2 we demonstrate that for a QF contract to which both parties would agree, the equilibrium analysis need consider only $Q^*_{QF}(q_{QF}) = q_{QF}(1 + \alpha)$.

Proposition 2. *Under the terms of a QF contract, when $q_{QF} > 0$, the EM produces exactly $q_{QF}(1 + \alpha)$. If $q_{QF} = 0$, for any given c the EM will prefer the NC arrangement.*

An outcome of $q_{QF} > 0$ signifies the retailer's willingness to submit to the conditions of the QF contract. The EM would not offer this contract unless such an outcome could be anticipated, since for a given transfer price this could only make the EM worse off by adding a constraint (relative to the NC arrangement) without compensation. Of course, $q_{QF} > 0$ does not guarantee the EM's preference for the QF contract over the NC arrangement or any other. This issue will be addressed in §8.

Proposition 3 provides properties of the equilibrium that will result when the QF contract is active, excluding the boundary case of $\omega = 1$, which is treated in Proposition 4.

Proposition 3. *Properties of the equilibrium solution (when $\omega < 1$):*

(a) *q^*_{QF} is strictly positive and finite, and may be obtained as the unique solution to*

$$(1 + \alpha)\int_{\mu + z_\epsilon\sigma_\epsilon \geq q(1+\alpha)} G'(q(1 + \alpha)|\mu)d\Theta(\mu) = -(1 - \omega)\int_{\mu + z_\epsilon\sigma_\epsilon \leq q(1-\omega)} G'(q(1 - \omega)|\mu) \times d\Theta(\mu). \quad (3)$$

(b) *While q^*_{QF} depends on α and ω individually, the system inventory $Q^*_{QF}(q^*_{QF}) = q^*_{QF}(1 + \alpha)$ depends on the flexibility parameters only through the ratio $\psi \equiv (1 + \alpha)/(1 - \omega)$.*

(c) *Comparative statics*:

	q^*_{QF}	Q^*_{QF}	$\pi^*_{R,QF}$	$\pi^*_{EM,QF}$
$\uparrow c$	−	−	−	can be + or −
$\uparrow \omega$	+	+	+	can be + or −
$\uparrow \alpha$	can be + or −	+	+	can be + or −

On the left side of Equation (3) in part (a) of Proposition 3, $G'(q(1 + \alpha)|\mu)$ represents the benefit to the retailer of a unit increase in product availability, and the integral takes the expectation over all scenarios in which that additional unit would be desired but not obtained (i.e., the purchase target exceeds the guaranteed amount). The $(1 + \alpha)$ multiplier appears because a unit increase in q affects $(1 + \alpha)$ additional units of product availability. Hence the left side is the retailer's expected marginal benefit from increasing the forecast. By a similar argument, on the right is the expected marginal cost of increasing the forecast, where cost is imposed via the minimum purchase implication. So the stated condition is analogous to the classical newsvendor solution.

Part (b) of Proposition 3 introduces ψ, a measure of the net "amount" of flexibility in a QF contract. This usage is appropriate since system inventory and the allocation of responsibility for it depend on α and ω only through this composite metric, which increases strictly with either parameter.[8]

The comparative statics for q^*_{QF} and Q^*_{QF} illuminate the dynamics of the supply relationship. An increase in the transfer price simultaneously increases the retailer's overage cost and decreases the underage cost, thereby decreasing the retailer's optimal forecast. An increase in ω represents a relaxation of the minimum purchase commitment, so that increasing the forecast makes the retailer no worse off for the cases in which demand turns out low, but better off when demand turns out high since the upside availability of

[8] Lariviere (1999) has observed that $1/\psi$ is the fraction of the total system inventory for which the retailer is ultimately responsible. We use ψ rather than its reciprocal so that a higher value denotes greater flexibility for the buyer.

the product will improve. Since Q^*_{QF} is simply a positive constant multiple of q^*_{QF}, these two changes have the same direction of impact on both expressions. The impact of α on q^*_{QF} is indeterminate due to two countervailing effects. As α is increased, the retailer has some incentive to decrease its forecast since doing so reduces the minimum purchase commitment while maintaining access to product. On the other hand, there is incentive to increase the forecast since its multiplicative impact on the EM's production is now magnified. However, any decrease in $(1 + \alpha)$ is more than offset by the associated increase in q^*_{QF}, so that their product, Q^*_{QF}, increases. With this, we have completely characterized the behavior that will result when the decentralized supply chain submits to a QF contract with parameters $\{c, (\alpha, \omega)\}$.

The properties of the expected profits show the preferences that each party carries into the negotiation of a QF contract. The retailer's preferences are analytically conclusive and intuitive in that the retailer prefers low cost (as in a standard newsvendor problem) and high flexibility. However, sensitivity properties of the EM's profit are not necessarily monotone for the following reason: while the EM has newsvendor cost structure, the demand it faces is sensitive to both price (the transfer price) and flexibility since both of these influence the retailer's purchase. While increasing c may reduce the EM's sales volume, the fatter profit margin may lead to net benefit. Likewise, sometimes the EM can do better by offering more flexibility because this stimulates the retailer's propensity to buy. This will be made explicit in §8.

6.2. Special Cases

In this section we evaluate two ways of governing the supply relationship that might be considered natural alternatives to the NC setting. Both may be represented as special cases of the QF contract. The first is an upside promise with no minimum purchase commitment, i.e., $\omega = 1$. The EM's response to the forecast may be represented as producing some nonzero fraction of the forecast, which can be written as $(1 + \alpha)q_{QF}$ for some unspecified $\alpha > -1$. (Without the upside commitment this would revert to the NC case.) Presumably the retailer would prefer this arrangement. A second possibility, likely to be preferable to the EM, requires the retailer to accept exactly what it forecasts. For example, during the recent worldwide supply shortage in the market for computer memory chips, chip manufacturers had the strategic power to require their customers to lock in firm orders 10 to 12 weeks in advance of delivery. This is represented as $(\alpha, \omega) = (0, 0)$, giving no flexibility to the retailer. As noted in Proposition 4, either arrangement leads to an inefficient outcome. The transfer price allocates profits between the EM and the retailer, but no price will allow the recovery of the efficiency loss.

Proposition 4. *Properties of extreme forms of the QF contract, for any transfer price c:*

(a) *If the retailer is not bound by any minimum purchase commitment ($\omega = 1$, or $\psi = \infty$), overproduction occurs relative to the central-control case.*

(b) *If the retailer must purchase the full amount forecast ($(\alpha, \omega) = (0, 0)$, or $\psi = 1$), underproduction occurs relative to the central-control case.*

In both cases, total system profit is strictly suboptimal.

As with the discussion in §5.1, part (a) of Proposition 4 suggests retailer overforecasting followed by a smaller actual purchase, which offloads some of the cost of demand uncertainty onto the EM. The cause is the retailer's lack of accountability for the initial forecast. Part (b) is simply double marginalization: because the retailer's procurement cost is higher than the product's true cost, his underage and overage costs are, respectively, greater and less than those perceived by the system as a whole. This completes the characterization of double marginalization in the QF setting: if, as in Proposition 1, the EM chooses the quantity, the outcome is underproduction; if the retailer bears full responsibility for this decision, overproduction occurs. These two scenarios are concrete examples that the QF contract is not necessarily efficient. The extent to which QF contracts can achieve system efficiency is explored the following section.

7. QF Contracts and Supply Chain Efficiency

Under the assumptions of this model, system efficiency can be attained by meeting two conditions: (i) the correct quantity (i.e., Q^*_{CC}) must be resident in the

system, and (ii) it must be fully accessible to end customers when market demand occurs. The first condition alone is insufficient, due to scenarios in which the retailer leaves some product with the EM (i.e., $r^*_{QF} < Q^*_{CC}$), then subsequently stocks out. A central planner will do better with the same quantity by positioning it all at the retailer site, ready for market demand. The QF structure enables coordination to the optimal production while both parties pursue individual objectives, which is significant. However, full efficiency can be established only under special conditions.

We now parameterize the equilibrium production as $Q^*_{QF}(c, \psi)$ to explore the role of the contract parameters. Proposition 4 states the existence of a parameter combination under which the QF contract achieves the efficient quantity. Moreover, among these contracts the transfer price moves in the intuitive direction, with the retailer paying more for greater flexibility.

Proposition 5. *For any* $c \in (m, p + s)$:

(a) *There exists a unique* ψ *such that* $Q^*_{QF}(c, \psi) = Q^*_{CC}$.

(b) *Among such combinations, greater flexibility* (ψ) *is associated with a higher transfer price.*

Closed-form characterization of the mapping between the aforementioned c and ψ is unavailable for the general model, but can be derived under a simplifying assumption. This will enable explicit illustration of key insights which are believed to apply to the general case as well. We now consider the special case $\sigma_\epsilon = 0$, meaning that μ is a perfect predictor of market demand.[9] With this demand model, the issue of product mispositioning vanishes. As the retailer's purchase will be based on a perfect signal, the retailer will stock out only when the EM stocks out. Therefore, as long as the production level matches Q^*_{CC}, expected system profits will be maximized. The parameter combinations that achieve this are described in Proposition 6.

Proposition 6. *For the demand model in which* $\sigma_\epsilon = 0$:

(a) *System efficiency will result from a QF contract with total flexibility* $\psi \equiv (1 + \alpha)/(1 - \omega)$ *when the transfer price is*

$$\bar{c}(\psi) \equiv u + \frac{m - u}{\frac{1}{\psi} F\left(\frac{1}{\psi} F^{-1}\left(\frac{p + s - m}{p + s - u}\right)\right) + \frac{m - u}{p + s - u}}. \tag{4}$$

(b) *Any split of expected profits can be achieved with some efficient QF contract.*

(c) *Among these contracts, increasing the flexibility shifts profits to the EM.*[10]

It is straightforward to verify that $d\bar{c}/d\psi > 0$, as in Proposition 5(b). Once $\bar{c}$ is calculated for a given ψ, the corresponding efficient α and ω will be determined only up to a negative linear relationship ($\omega = 1 - (1 + 2)/\psi$). To preserve efficiency, any increase in α, which encourages the retailer to lower its absolute inventory commitment via a lower q^*_{QF}, must be counteracted by reducing ω, which commits the retailer to a larger portion of that q^*_{QF}.

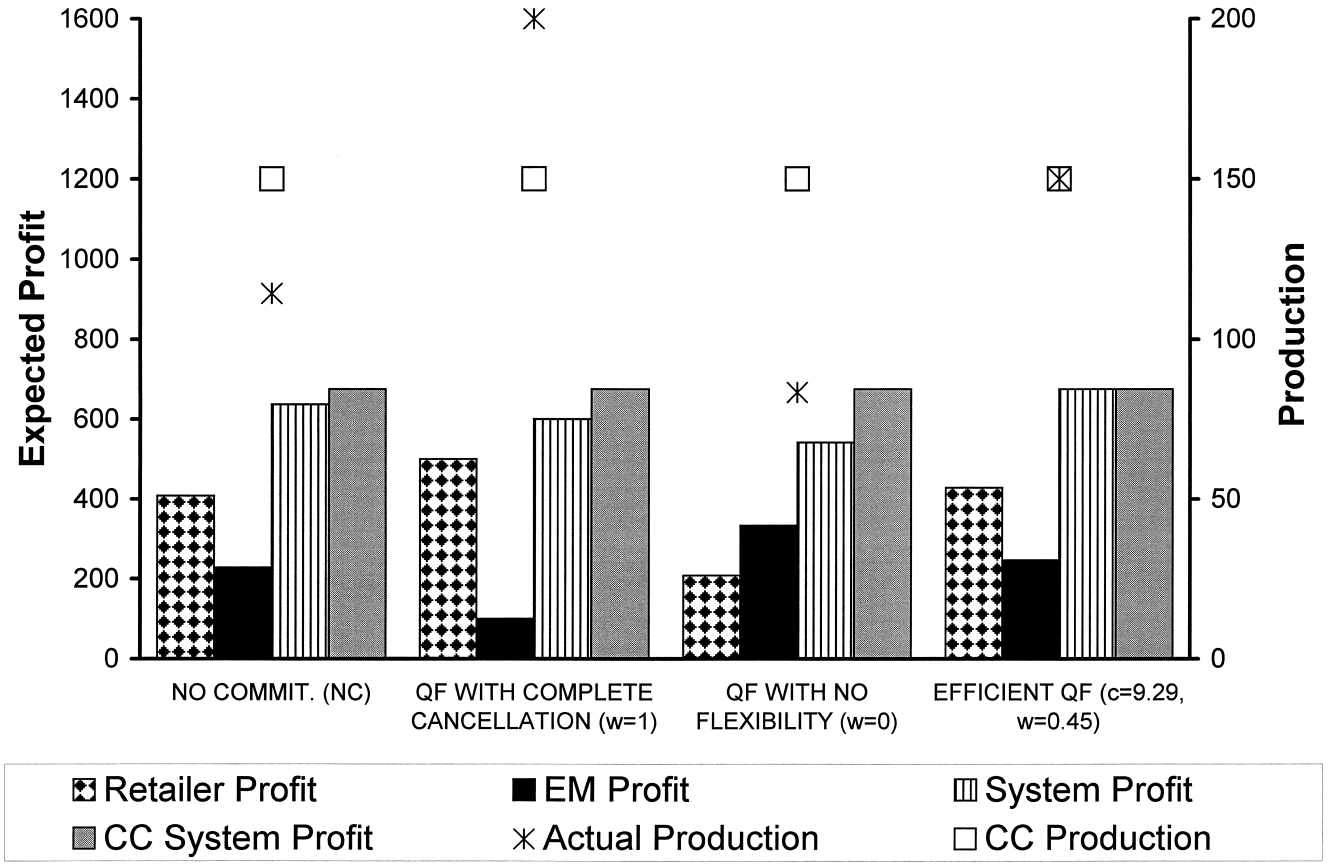
How flexibility affects the profit allocation seems to oppose our intuition that the retailer should prefer access to more flexibility. This was shown to be true *for a fixed transfer price*. However, when the transfer price is simultaneously adjusted to restore efficiency in the system, it turns out that the EM ends up claiming more of the system profits. The idea of a menu of price-flexibility combinations, as implied by part (a) of Proposition 6, is consistent with contracts offered by health and beauty aid distributor McKesson to retailers (cf. Padmanabhan and Png 1995). Parts (b) and (c) provide insight for why McKesson might willingly offer them, as Pareto improvement over any alternative can be achieved with the properly chosen subset of all coordinating QF contracts. This will be illustrated numerically in the next section.

As evidenced by the form of $\bar{c}$, common beliefs about market demand are necessary for installation

[9] Another special case to consider might be $\sigma_\mu = 0$. Since nothing transpires after time zero to motivate the exercise of flexibility, this reduces to the full-commitment scenario of Proposition 4(b), in which double marginalization leads to underproduction for any $c > m$.

[10] The author thanks Marty Lariviere for providing an elegant proof of parts (b) and (c).

Figure 1 Comparison of Control Methods



of an efficient QF contract. We note this here and reiterate the point of view of this paper: even if multiparty supply chains evolve towards collaborative interpretation of shared market data, inefficiency can result due to individual incentives. Thus, mechanisms are needed which can visit upon the decision-making parties the proper share of the costs of demand uncertainty. The QF contract implements this by adding commitments to the supply agreement.

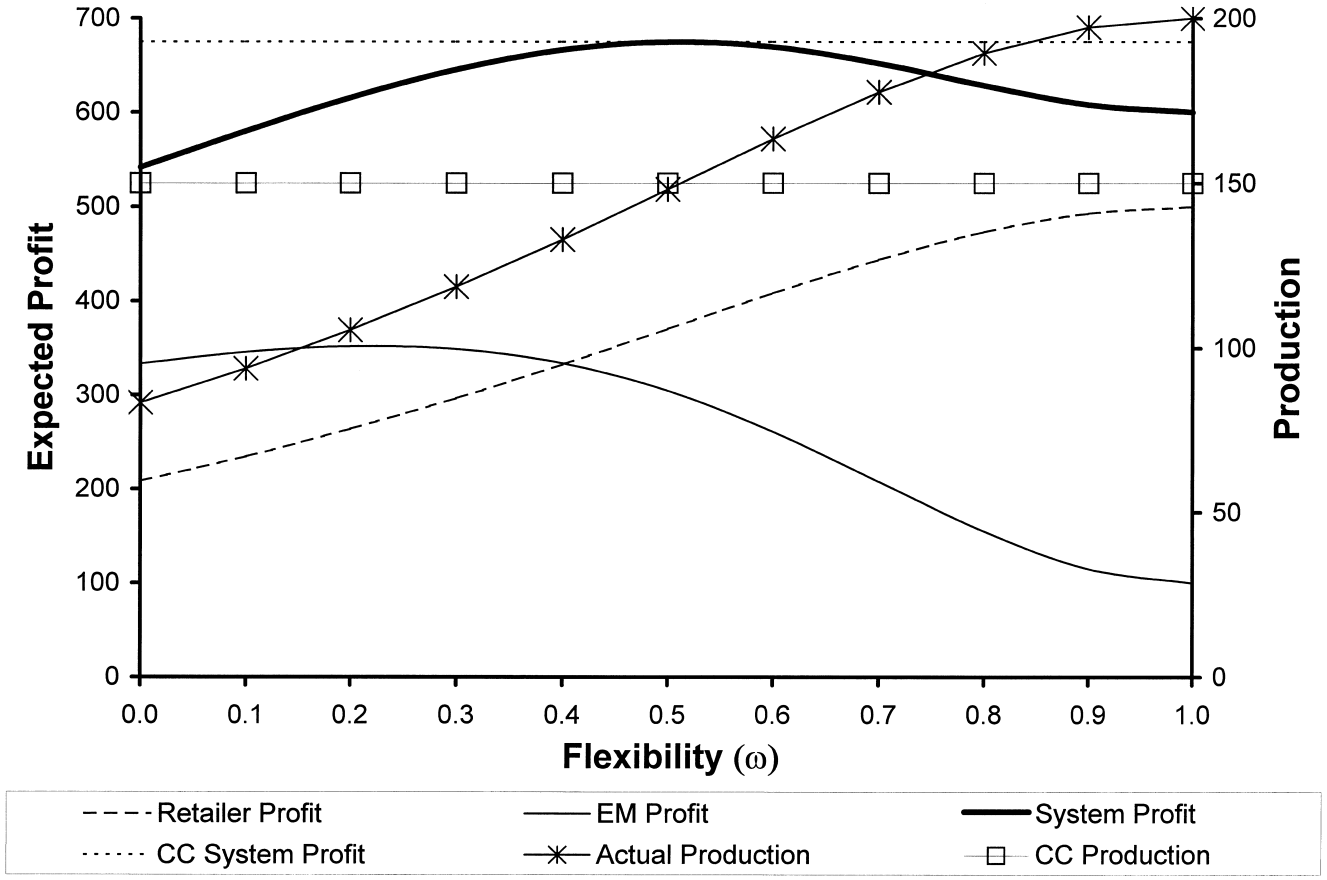
8. Numerical Analysis

In this section we present numerical analysis to corroborate and supplement the previous developments. For analytical convenience we assume $\sigma_\epsilon = 0$, the case in which system efficiency can be achieved with an appropriate QF contract (cf. Proposition 6).[11] Additionally, to enable closed forms for decisions and profits under the various control schemes (omitted for space considerations), we consider market demand X which is uniform over the support $[M - \delta, M + \delta]$, for which $\mu \equiv E[X] = M$ and $\sigma_X = \delta/\sqrt{3}$. Unless otherwise noted, the analysis uses financial parameters $\{p = 15, c = 10, m = 6, u = 3, s = 0\}$, demand parameters $\{M = 100, \delta = 100\}$, and (without loss of generality) $\alpha = 0$ in all QF contracts.

The progression below follows the basic outline of the preceding discussion. After juxtaposing the alternative environments, we demonstrate sensitivity

[11] Analysis of the case of $\sigma_\epsilon > 0$ is computationally formidable, requiring repeated numerical solution to equations containing integrals of complex functions.

Figure 2 **Expected Profit and Production vs. Flexibility** ($c = 10$)



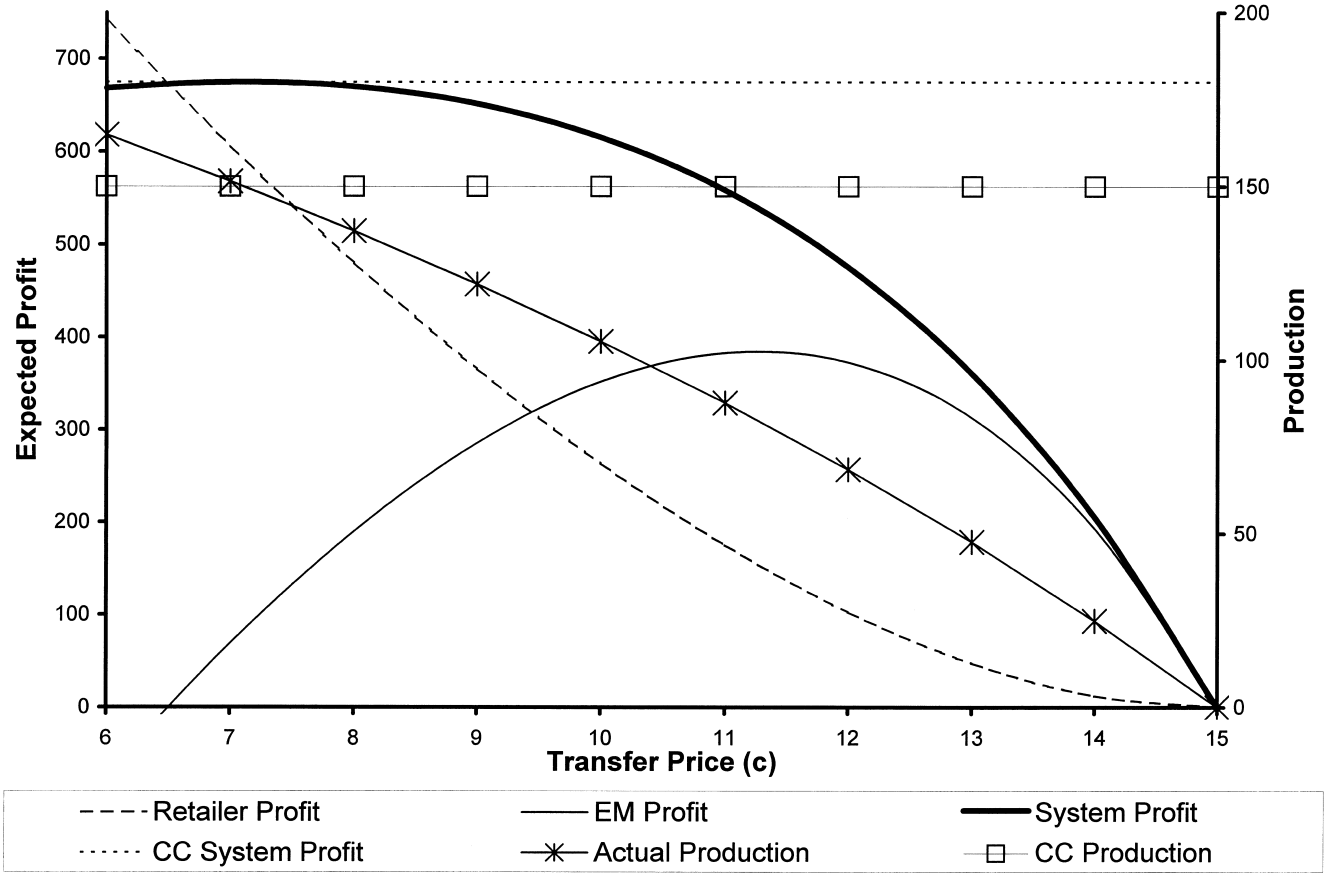
implications of QF control. We then attend to the properties of efficient QF contracts. Many of the following results were proven earlier for the general model. The remainder can serve as motivation for future research.

8.1. QF Contracts and Supply Chain Performance

Figure 1 compares the various modes of control. For each, the bars (associated with the left axis) depict the allocation of expected profit, and how the total compares to what central control could achieve. Stars (associated with the right axis) mark the production relative to the central planner's optimal production, marked with a square. This convention will be used throughout. Figure 1 supports the earlier analyses. The NC setting is inefficient because of underproduction. Allowing the retailer complete cancellation results in overforecasting (the maximum possible demand is initially forecast) and is also abusive to the EM. Taking all flexibility away from the retailer ("QF with No Flexibility") is not the solution, although this certainly improves the lot of the EM. Finally, in the last column we have set $\omega = 0.45$ and then installed the efficient transfer price per Proposition 6(a). The retailer accepts a 55% purchase commitment in exchange for a reduction in unit price from 10 to 9.29, and (by design) both parties are made better off relative to the NC outcome (cf. Figure 5).

We now illustrate sensitivity to each QF contract parameter. Figure 2 considers flexibility (since $\alpha = 0$, this is uniquely specified by ω), while Figure 3 varies the transfer price. These figures depict the preferences of each party, validating Proposition 3(c). Figure 2 shows retailer profit to increase with

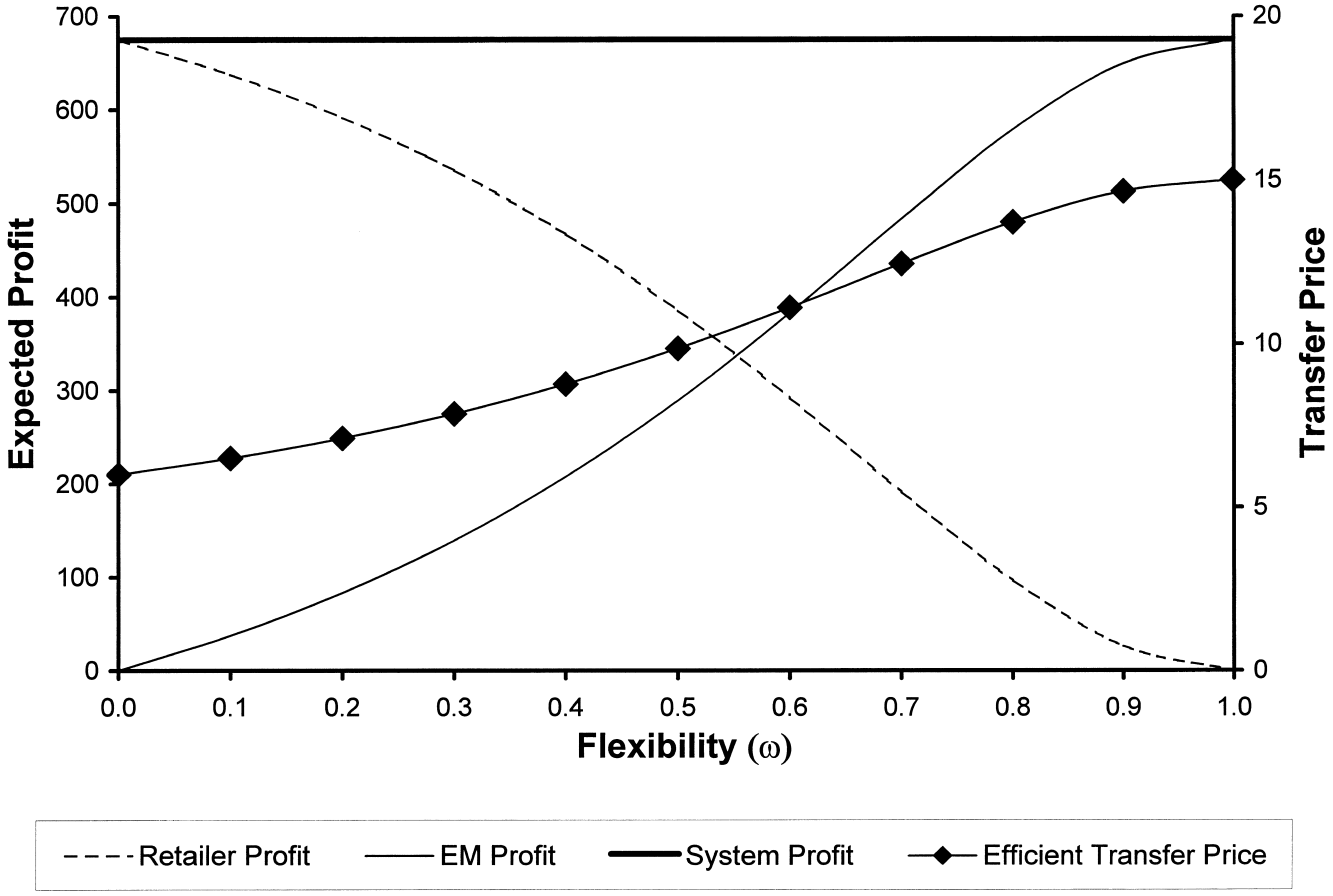
Figure 3 **Expected Profit and Production vs. Transfer Price** ($(\alpha, \omega) = (0, 0.2)$)



flexibility, with the left-to-right progression providing the transition from full purchase commitment to full cancellation privileges. This makes concrete the value of flexibility to a buyer and therefore the willingness to pay for it. Meanwhile, starting from the left of Figure 2, the EM would initially offer some flexibility unilaterally to encourage the retailer to purchase. Efficiency is attained at $\omega = 0.51$, which is consistent with Equation (4) (note the achievement of CC production). Next, Figure 3 confirms that while retail profit is always decreasing in c, the EM's profit first increases, then decreases. The left extreme has the EM pricing at cost ($c = m$). This imparts on the retailer the cost structure of the entire channel, which is first-best in many traditional models (§2 mentions some settings in which double marginalization can be solved in this way). However, the retailer is partially insured against low demand (since it can dump up to 20%), so it will overorder. The retailer does even better than a central planner for this reason, but this benefit comes at the expense of negative profit for the EM, and the net effect is inefficiency. For the given flexibility parameters, efficiency occurs at $c = 7.1$. Of course, in this example the EM individually prefers a transfer price higher than this.

Figures 2 and 3 show the general tension in the contract negotiation process and illustrate the two independent degrees of control that allow the QF contract to coordinate the system without requiring $c = m$, which would leave the EM profitless. The similarity of the general shapes (up to a reflection and

Figure 4 **Expected Profit vs. Flexibility in an Efficient QF Contract**



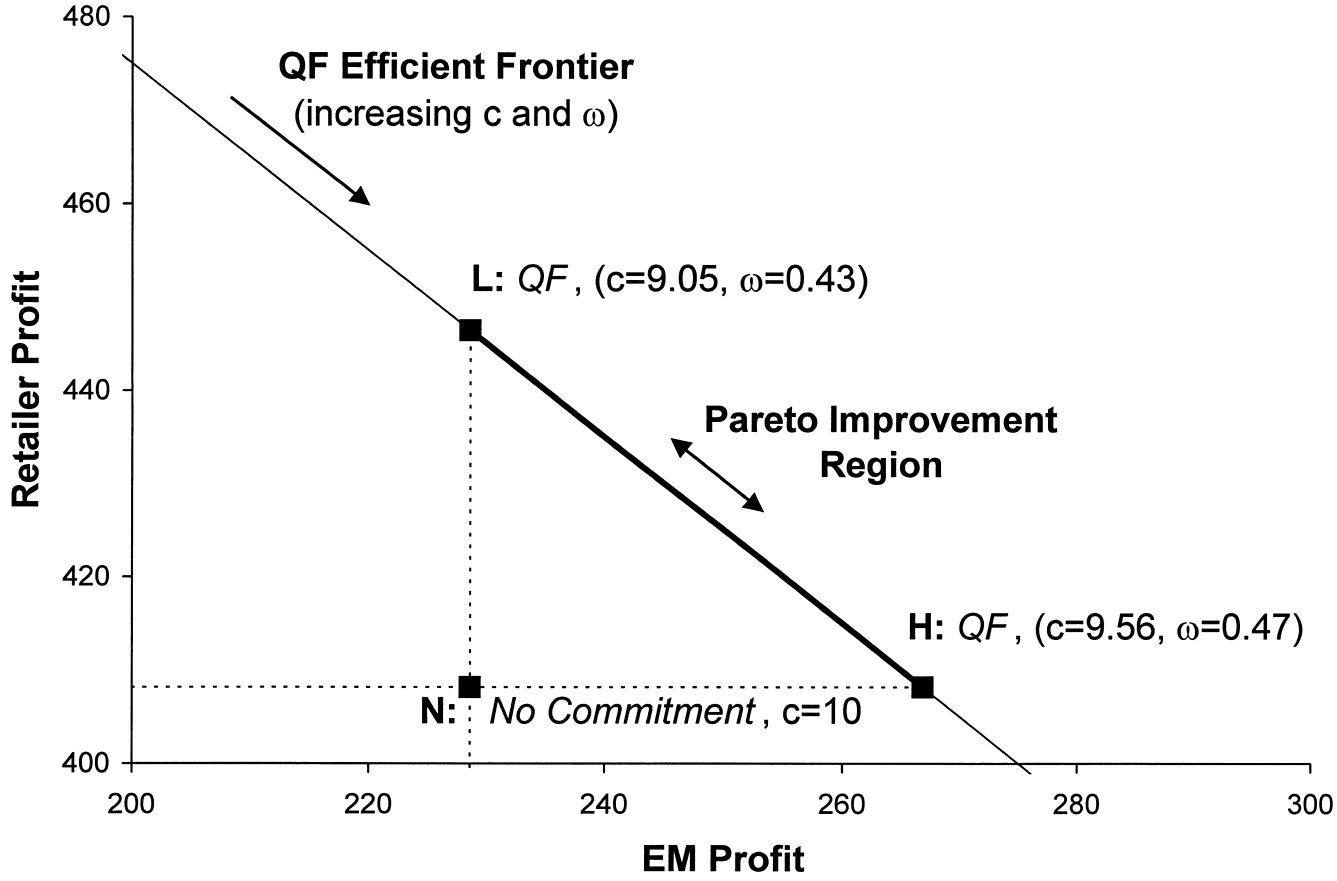
rescaling) underscores the duality between pricing and constraints.

8.2. Efficient QF Contracts

We now examine efficient QF contracts. We no longer report the production since the salient feature of an efficient system is that actual production matches the central control standard.

Figure 4 shows how profit is allocated when the set of efficient contracts is arrayed by ω. As proven for the general model in Proposition 5, whenever the retailer receives more flexibility, the efficient c also increases enough to offset any profit gains. Also, any profit allocation is possible, which attends to the issue of participation constraints. Figure 5 elaborates on obtaining Pareto improvement vis-à-vis any inefficient alternative (this example assumes an NC status quo). Given closed forms for the profit functions, it is straightforward to derive the efficient frontier and a segment ($\overline{\text{LH}}$) on which both parties prefer the QF contract to the NC arrangement (point N). The choice of c within this segment allocates the efficiency gain. At L, which has the lowest efficient c, the retailer reaps all gains and the EM is indifferent; the reverse is true at H. On moving from L to H, c ranges from 9.05 to 9.56, and the purchase commitment $(1 - \omega)$ from 57% down to 53%. All these combinations represent the retailer's willingness to accept greater inventory burden in exchange for a unit price reduction (down from $c = 10$); as one might expect, greater burden requires a larger discount. Note that increasing the transfer

Figure 5 **The QF Efficient Frontier and Pareto Improvement**



price diverts the efficiency gains to the EM, the counterintuitive result stated in Proposition 6(c).

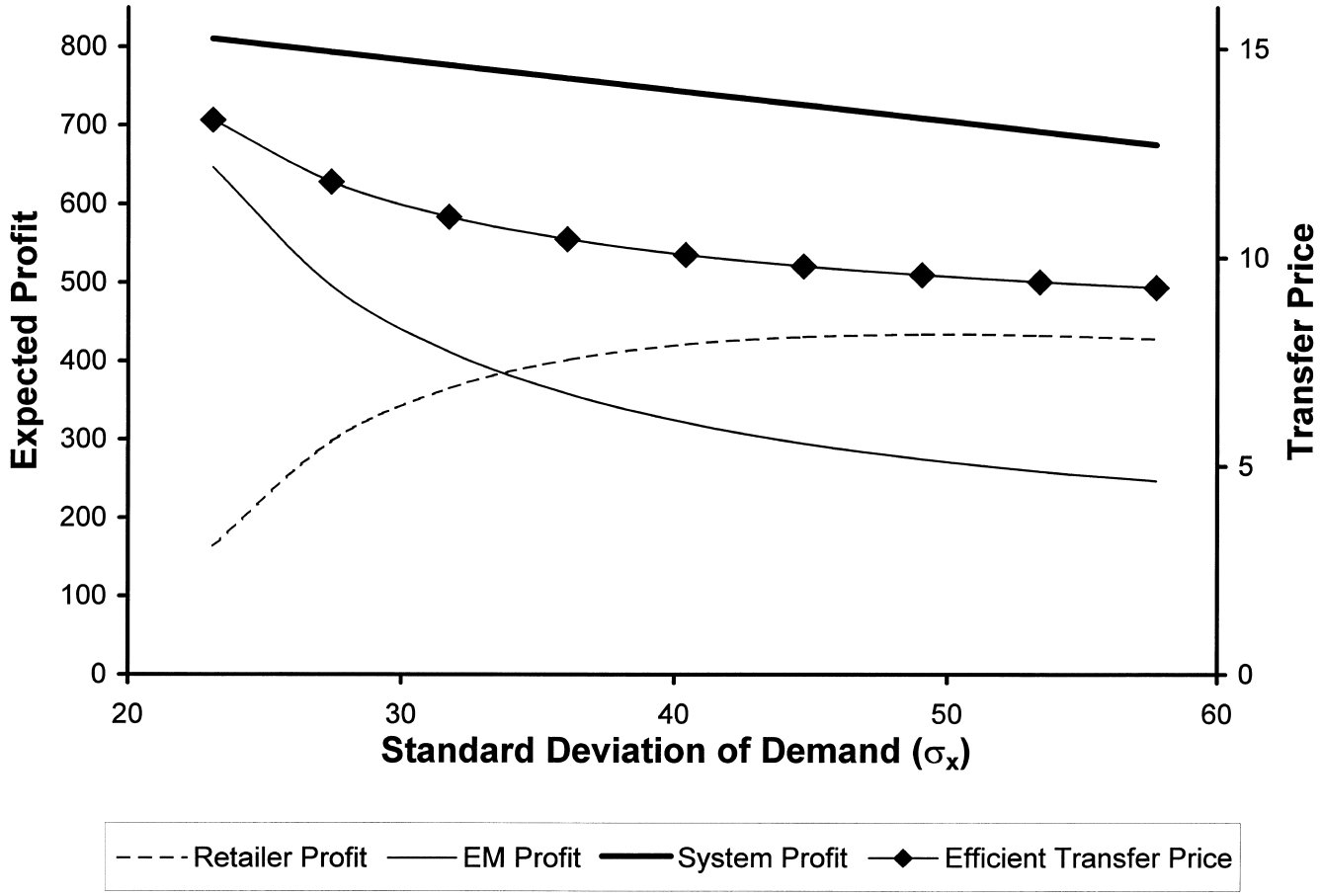
Finally, Figure 6 considers the influence of demand uncertainty on the efficient transfer price and allocation of profits. For the assumed distribution, increasing δ represents a mean-preserving spread that increases σ_X proportionally.

For a fixed flexibility, as σ_X increases the efficient c decreases. This is because the retailer finds the installed flexibility to be decreasingly meaningful in the task of matching the market demand. In other words, this represents a progression towards the dynamics of the full-commitment scenario of Proposition 3(b), in which double marginalization drives down the retailer's purchase. A reduction in transfer price counteracts this effect, while shifting profit to the retailer (since the flexibility is unaltered). Figure 6 underscores the dependence of the efficient contract on the distribution of demand (evident from Equation (4)), and therefore the need to renegotiate the contract whenever beliefs about demand change (if efficiency is desired). Interestingly, though, in the limit the coordinating contract and profit breakdown are approximately insensitive to the demand uncertainty. Again, this is because the installed flexibility plays less and less a role in the retailer's behavior. The dependence of the value of flexibility on demand volatility validates the managerial heuristics noted in §1.3 and corroborated by Tsay and Lovejoy (1999).

9. Summary and Discussion

This paper considers a decentralized supply relationship in which the customer's advance forecast need not imply complete commitment to its subsequent

Figure 6 **Expected Profit and Efficient Transfer Price vs. Demand Variability** ($\omega = 0.2$)



purchase quantity. By examining the incentives on each side of the relationship, we have found that inefficiency will result in the absence of additional structure. We have identified particular forms of behavior, such as overforecasting or simply making decisions based on a local rather than global perspective, that are natural consequences of decentralized control. Double marginalization is one salient issue, and another is that different parties make commitments under different states of information.

We have shown that these problems can be at least partially remedied by the QF contract, in which the retailer commits to a minimum purchase and the EM guarantees a maximum coverage (both stated as a percentage deviation from the retailer's initial forecast). This is conceptually simple and easy to implement, and has a cooperative flavor in that each party accepts some of the inventory and stockout cost burden.

We have illuminated the individual preferences that drive the negotiation of such a contract. Naturally, the retailer always pushes for lower price and greater flexibility. On the other side, basic economics tells us that, in selling to a demand which is sensitive to price, sometimes the EM is rational in unilaterally conceding a more attractive price. But only by explicitly modeling flexibility have we been able to verify that the same holds true for this nonprice attribute.

By itself the QF contract does not guarantee efficiency. However, we have described conditions under which this arrangement will generate efficiency gains that can be shared by the two parties. So the customer commits to a minimum purchase agreement in ex-

change for a unit price reduction. And the supplier is rational in offering that price break in exchange for more predictable sales. There is indeed a tradeoff between flexibility and unit price, with the customer willingly paying more for increased flexibility. We have illustrated how this might unfold, including identifying arrangements that either party could propose with confidence that the other would accept.

The QF contract may thus take its place among the solutions to be considered when the partners in a supply relationship diverge over the ownership of inventory and the commitment implied by a forecast. Virtually all those mentioned in §2 work via differential pricing for the exercise of revision privileges, including returns policies (Pasternack 1985 and others), options (e.g., Barnes-Schuster et al. 1998), and two-tiered pricing (e.g., Weng 1997). Each may be more appropriate than the others for certain settings. Several managers have noted to the author the qualitative appeal of the QF contract, in the sense that the single-price structure is more politically palatable. Whether rational or not, there is a desire to avoid the appearance of having had to purchase additional units at a premium due to poor planning. Instead, the QF contract incorporates the cost of the flexibility into the single unit price. Also, single-rate pricing reflects a philosophical view of many companies in turbulent demand environments (e.g., technology-intensive industries) that order revisions will be "business-as-usual," and the administrative burden is lower if the parties need not worry about additional side-payments (cf. Farlow et al. 1995). The percentage constraints are in place to deter abuse. Indeed, in the examples of Compaq, Solectron, and Sun Microsystems, the observed contracts do not attach any financial consequence to any revisions that remain within the stated bounds.

As established in this analysis, even when the statistics of market demand are common knowledge, there is still a need to properly structure the supply relationship to share the consequences of uncertainty in that demand. Incentives and information are distinct causes of inefficiency and should be managed as such. However, because our results demonstrate efficiency only under shared beliefs, the issue of coordination under information asymmetry remains unresolved. We might conjecture that when the EM relies heavily on the retailer for guidance about market conditions, the QF contract may yet be effective at mitigating the retailer's gaming by attaching economic accountability to the forecasts. For this reason the QF contract seems to extend naturally to supply chains more than two players deep, as each link represents yet another opportunity for information distortion.[12] Certainly, these are issues of import for managers of modern supply chains, and are worthy of substantial additional research.[13]

[12] Solectron is one such example (Ng 1997). Tsay and Lovejoy (1999) characterize material buildups and information dynamics when such supply chains operate over multiple time periods.

[13] The author is grateful to Naren Agrawal, Alex Angelus, Marty Lariviere, Hau Lee, Bill Lovejoy, Steve Nahmias, Rhonda Righter, Steve Smith, and Jin Whang for comments on this paper. Also, two referees and an associate editor have provided insightful and constructive recommendations which have greatly shaped the exposition of this paper.

References

Atkinson, A. A. 1979. Incentives, uncertainty, and risk in the newsboy problem. *Decision Sci.* **10** 341–357.

Barnes-Schuster, D., Y. Bassok, R. Anupindi. 1998. Supply contracts with options: flexibility, information and coordination. Working Paper, University of Chicago, Chicago, IL.

Bassok, Y., R. Anupindi. 1995. Analysis of supply contracts with forecasts and flexibility. Working Paper, Northwestern University, Evanston, IL.

——, —— 1997a. Analysis of supply contracts with total minimum commitment. *IIE Trans.* **29** 373–381.

——, —— 1997b. Analysis of supply contracts with commitments and flexibility. Working Paper, Northwestern University, Evanston, IL.

Bergen, M., S. Dutta, O. C. Walker. 1992. Agency relationships in marketing: a review of the implications and applications of agency and related theories. *J. Marketing* **56**(3) 1–24.

Chen, F. 1997. Decentralized supply chains subject to information delays. Working Paper, Graduate School of Business, Columbia University, New York.

Clark, A. J., H. Scarf. 1960. Optimal policies for a multiechelon inventory problem. *Management Sci.* **6** 475–490.

Connors, D., C. An, S. Buckley, G. Feigin, A. Levas, N. Nayak, R. Petrakian, R. Srinivasan. 1995. Dynamic modeling for re-engineering supply chains. IBM Research Report, T. J. Watson Research Center, Yorktown Heights, NY.

Crocker, K. J., S. E. Masten. 1991. Pretia ex machina? Prices and process in long-term contracts. *J. Law Econom.* **34** 69–99.

Donohue, K. L. 1998. Efficient supply contracts for fashion goods with forecast updating and two production modes. Working

Paper, Department of OPIM, The Wharton School, University of Pennsylvania, Philadelphia, PA.

Emmons, H., S. M. Gilbert. 1998. Note: The role of returns policies in pricing and inventory decisions for catalogue goods. *Management Sci.* **44**(2) 276–283.

Eppen, G. D., A. V. Iyer. 1997. Backup agreements in fashion buying—the value of upstream flexibility. *Management Sci.* **43** 1469–1484.

Farlow, D., G. Schmidt, A. A. Tsay. 1995. Supplier management at Sun Microsystems. Case Study, Graduate School of Business, Stanford University, Stanford, CA.

Faust, M. 1996. Personal communication from a product manager at one of Compaq's suppliers of memory chips, Santa Clara, CA.

Fisher, M., A. Raman. 1996. Reducing the cost of demand uncertainty through accurate response to early sales. *Oper. Res.* **44**(1) 87–99.

Ha, A. Y. 1997. Supply contract for a short-life-cycle product with demand uncertainty and asymmetric cost information. Working Paper, Yale School of Management, New Haven, CT.

Iyer, A., M. E. Bergen. 1997. Quick response in manufacturer-retailer channels. *Management Sci.* **43**(4) 559–570.

Jeuland, A. P., S. M. Shugan. 1983. Managing channel profits. *Marketing Sci.* **2** 239–272.

Kandel, E. 1996. The right to return. *J. Law Econom.* **39** 329–356.

Katz, M. L. 1989. Vertical contractual relations. R. Schmalensee, R. D. Willig, eds. *Handbook of Industrial Organization*, vol. I. Elsevier Science Publishers B.V., New York.

Lariviere, M. A. 1999. Supply chain contracting and coordination with stochastic demand. S. Tayur, R. Ganeshan, M. Magazine, eds. *Quantitative Models for Supply Chain Management*. Kluwer Academic Publishers, Norwell, MA.

Lee, H. L., P. Padmanabhan, S. Whang. 1997. The bullwhip effect in supply chains. *Sloan Management Rev.* **38**(3) 93–102.

——, S. Whang. 1997. Decentralized multi-echelon inventory control systems: incentives and information. Working Paper, Stanford University, Stanford, CA.

Lovejoy, W. S. 1999. *Integrated Operations*. Southwestern College Publishing, Cincinnati, OH.

Magee, J. F., D. M. Boodman. 1967. *Production Planning and Inventory Control*. McGraw-Hill, New York.

Masten, S. E., K. J. Crocker. 1985. Efficient adaptation in long-term contracts: take-or-pay provisions for natural gas. *Amer. Econom. Rev.* **75** 1083–1093.

Mathewson, G. F., R. A. Winter. 1984. An economic theory of vertical restraints. *Rand J. Econom.* **15**(1) 27–38.

Mondschein, M. 1993. Negotiating product supply agreements. *National Petroleum News* **85** 45.

Moorthy, K. S. 1987. Managing channel profits: comment. *Marketing Sci.* **6**(4) 375–379.

National Energy Board. 1993. Natural gas market assessment: long-term Canadian natural gas contracts. *Gas Energy Rev.* **21** 8–11.

Ng, S. 1997. Supply chain management at Solectron. Presentation, Industrial Symposium on Supply Chain Management, Stanford University, Stanford, CA.

Padmanabhan, P., I. P. L. Png. 1995. Returns policies: make money by making good. *Sloan Management Rev.* **37**(1) 65–72.

Parlar, M., Z. K. Weng. 1997. Designing a firm's coordinated manufacturing and supply decisions with short product life cycles. *Management Sci.* **43**(10) 1329–1344.

Pasternack, B. A. 1985. Optimal pricing and returns policies for perishable commodities. *Marketing Sci.* **4** 166–176.

Rockafellar, R. T. 1970. *Convex Analysis*. Princeton University Press, Princeton, NJ.

Ross, K. A. 1980. *Elementary Analysis*. Springer-Verlag, New York.

Spengler, J. J. 1950. Vertical restraints and antitrust policy. *J. Political Econom.* **58** 347–352.

Tirole, J. 1988. *The Theory of Industrial Organization*. MIT Press, Cambridge, MA.

Tsay, A. A. 1995. Supply chain control with quantity flexibility. Ph.D. dissertation, Graduate School of Business, Stanford University, Stanford, CA.

——, W. S. Lovejoy. 1999. Quantity flexibility contracts and supply chain performance. *Manufacturing & Service Oper. Management* **1**(2).

——, S. Nahmias, N. Agrawal. 1999. Modeling supply chain contracts: A review. S. Tayur, R. Ganeshan, M. Magazine, eds. *Quantitative Models for Supply Chain Management*. Kluwer Academic Publishers, Norwell, MA. 299–336.

Van Ackere, A. 1993. The principal/agent paradigm: its relevance to various functional fields. *European J. Oper. Res.* **70** 83–103.

Varian, H. R. 1984. *Microeconomic Analysis*, 2nd ed. W. W. Norton, New York.

Verity, J. W. 1996. Clearing the cobwebs from the stockroom. *Bus. Week* (Oct. 21) 140.

Weng, Z. K. 1997. Pricing and ordering strategies in manufacturing and distribution alliances. *IIE Trans.* **29**(8) 681–692.

Whang, S. 1995. Coordination in operations: a taxonomy. *J. Oper. Management* **12** 413–422.

Accepted by Hau Lee; received July 1, 1996. *This paper has been with the author* $9\frac{1}{2}$ *months for* 3 *revisions.*