2011 MSOM Special Interest Group Meetings
Ann Arbor, Michigan, June 26, 2011

PROCEEDINGS

Supply Chain Management SIG

Supply Chain Management SIG Chair:
Felipe Caro
Paper Title: The Relational Advantages of Intermediation

[Previous Title: An Advantage of Sourcing through Intermediaries]

Author 1:
Elena Belavina,
Phd Candidate, INSEAD
Boulevard De Constance, Fontainebleau, INSEAD. 77300.
elena.belavina@insead.edu.

Author 2:
Karan Girotra,
Assistant Professor, INSEAD
Boulevard De Constance, Fontainebleau, INSEAD. 77300.
karan.girotra@insead.edu
Presenting Author
THE RELATIONAL ADVANTAGES OF INTERMEDIATION

ABSTRACT. This paper provides a novel explanation for the use of supply chain intermediaries such as Li & Fung Ltd. We find that even in the absence of the well-known transactional and informational advantages of mediation, intermediaries improve supply chain performance. In particular, intermediaries facilitate responsive adaptation of the buyers’ supplier base to their changing needs while simultaneously ensuring that suppliers behave as if they had long-term sourcing commitments from buying firms. In the face of changing buyer needs, an intermediary that sources on behalf of multiple buyers can responsively change the composition of future business committed to a supplier such that a sufficient level of business comes from the buyer(s) that most prefer this supplier. On the other hand, direct buyers that source only for themselves must provide all their committed business to a supplier from their own sourcing needs, even if they no longer prefer this supplier. Unlike existing theories of intermediation, our theory better explains the observed phenomenon that while transactional barriers and information asymmetries have steadily decreased, the use of intermediaries has increased rapidly, even among large companies such as Walmart.

1. INTRODUCTION

This paper is inspired by the phenomenal growth of supply chain intermediaries that source products or services on behalf of other firms. These often completely take over the sourcing function—they select, verify and approve suppliers, they allocate business between different suppliers, and they manage the relationship with each supplier, including providing incentives for investments, performance and compliance.

A notable sourcing intermediary is Li & Fung Ltd., which provides sourcing services to major brands and retailers worldwide, including Walmart, Target, Zara, Marks & Spencer PLC, Levis, and Philip Morris. Li & Fung has grown at a compounded annual rate of 23% for the last 14 years to achieve annual sales of over HK$ 120 Billion. While best known for sourcing apparel and toys from the low-cost economies of Asia, the group today operates in an expanding range of categories. It is present in over 40 economies across North America, Europe and Asia, with a global sourcing network of nearly 15,000 international suppliers, as well as thousands of buyers. It has abilities to provide both low-cost and quick, responsive sourcing. Yet, Li & Fung does not own any means of production or transport, nor is it in the business of directly retailing the vast majority of the products it


Key words and phrases. Global Sourcing, Intermediaries, Relational Contracts, Supply Chain Relationships, Flexibility, Contracting, Li & Fung, Supply Chain Management, Repeated Games.
sources. Essentially, it provides only an interface between multiple buyers and suppliers (McFarlan et al. (2007); Magretta (1998); Cheng (2010)). Olam International is another fast-growing sourcing intermediary with a similar business model. It is best known for sourcing agricultural products from Asia, Africa and Latin America. It provides sourcing services to some of the world’s biggest fast food chains and food companies such as McDonald’s, Nestlé, etc. Firm’s annual revenues for 2009 were S$10.4 Billion (Bell and Shelman (2009)).

The benefits and costs that intermediaries bring to supply chains have long been studied by scholars in Finance, Economics and Supply Chain Management (cf. Wu (2004) for a comprehensive summary of past research on intermediation). Two main benefits are identified to justify the existence of intermediaries: transactional benefits and informational benefits. Transactional benefits include the ability of intermediaries to provide immediacy by holding inventory or reserving capacity, and the benefits that arise out of the reduced costs of trade. Intermediaries that aggregate demand can use their scale for better utilization of facilities, amortization of fixed costs, reduction in the costs of searching and matching, and ease of price discovery. These transactional benefits are most salient for smaller firms that do not individually possess the scale to justify fixed investments, and when the institutional barriers to trade are high. A second class of benefits arises from the informational role that intermediaries play. An intermediary’s exposure to and better ability to synthesize dispersed information allows it to reduce information asymmetries, ensure better price discovery, and provide superior administration of contractual coordination mechanisms. Both these gains increase the efficiency of a supply chain, and the intermediary can appropriate some of these gains while sharing the rest with its supply chain partners. On the other hand, an additional tier in a supply chain is known to increase incentive misalignment, which can lead to insufficient stocking levels, poor information sharing and insufficient investments (Cachon and Lariviere (2005)).

Interestingly, with advancements in communication technologies and reductions in barriers to trade, many scholars have predicted a "flat world", in which global economic integration and democratizing technologies would render both the informational and transactional roles of intermediaries irrelevant. In particular, scholars have long hypothesized that one of the major business impacts of the internet would be the dis-intermediation of traditional entities (Wigand and Benjamin (1995); Friedman (2007)). Online platforms such as Alibaba.com have indeed rendered the traditional price discovery and matching roles of intermediaries irrelevant. The growth of intermediaries in the face of changes brought about by the internet and economic integration suggests that the conventional view on the advantages of intermediation may be incomplete.
Further, it is instructive to examine the firms that have decided to move away from direct sourcing to mediated sourcing. In January 2010, Walmart Inc. decided to enter into an open-ended sourcing arrangement with Li & Fung Ltd. (Cheng (2010)). The agreement delegated the sourcing of certain Walmart products to Li & Fung, which was expected to bring revenues in excess of US$2 Billion to Li & Fung. Many of Li & Fung’s clients are similar large firms, such as Target, Gap, Benetton, etc. Existing theory on the role of intermediaries based on scale and informational advantages seem less credible in explaining the move of big firms to adopt mediated sourcing. In particular, firms like Walmart arguably have more scale, similar market access, and local information than the intermediaries that they hire.\footnote{In 2011, Walmart’s annual revenues were US$421.85 billion, compared to Li & Fung’s US$15.96 Billion. Walmart also operates over 189 super-centers in China and employs over 50,000 local employees, making it one of the larger organized hypermarket chains in China. Source: Li & Fung Annual Report, 2010. Walmart 10-K filing, Q1 2011.} An anecdotal analysis of the reasons provided by firms for employing sourcing intermediaries highlights two key themes. First, the ability of firms like Li & Fung to ensure better supplier collaboration, investments and compliance with quality, social and environmental norms is highlighted. Supplier investments in capacity and in ensuring compliance are cited as major business risks that are alleviated by intermediation. Second, it is argued that mediated sourcing allows firms to be more responsive in adapting their supplier base in the face of changes in the business environment such as supply chain disruptions brought about by adverse natural events, political upheaval, and volatility in the trade environment (energy costs, exchange rates, tariffs, etc.) (Fung et al. (2008); Loveman and O’Connell (1995); McFarlan et al. (2007)).

This paper provides a new, previously unidentified advantage of sourcing through intermediaries. We develop a stylized model to compare direct and mediated sourcing. Our model captures two key features of the sourcing environment: the fact that buyer’s preferences over suppliers change over time as the business environment changes, and the presence of incomplete contracts due to non-verifiability/non-contractability of supplier investments in capacity, quality or compliance with social, environmental norms, limited legal liabilities, etc. (Grossman and Hart (1986); Hart and Moore (1988); Aghion and Holden (2011)).

Our analysis illustrates that an intermediary that pools the sourcing needs of different buyers is better than individual direct buyers at incentivizing beneficial supplier behavior and at responsively adjusting the buyers’ supplier base. With incomplete contracts that typify the sourcing of all but the simplest commodities, suppliers are typically incentivized by committing to provide a future business contingent on supplier performance. However, with changing preferences over suppliers, meeting these commitments may require sourcing from less-preferred suppliers. An intermediary that sources on behalf of multiple buyers breaks this trade-off by exploiting differences between...
different buyers’ preferences over suppliers. Essentially, an intermediary can responsively change the composition of the committed business such that the level of business required to ensure desired supplier behavior comes as much as possible from the buyer(s) that most prefer this supplier. On the other hand, direct buyers, which source only for themselves, must provision all the committed business from their own sourcing needs, irrespective of what their preferences over suppliers may be. Essentially, sourcing for multiple buyers provides intermediaries with a certain flexibility in meeting the commitment to provide future business to a supplier— the flexibility of choosing which buyer to match to which supplier.

We rigorously demonstrate the existence and the operation of this effect in a model with two buyers, two suppliers and an intermediary that allows for any generic game-theoretic interactions between buyers, suppliers and intermediaries that contribute to contractual incompleteness. We also allow buyer preferences over suppliers to vary in an arbitrary, stochastic, non-stationary, heterogeneous fashion. Our analysis illustrates that the key to the existence of the highlighted advantage is a difference in buyer preferences over suppliers, at any given time. This difference could arise out of stochastic preferences over suppliers of ex-ante identical buyers or with deterministic but non-stationary preferences of heterogeneous buyers.

Our analysis of mediated sourcing makes three key contributions: First, we provide a new explanation for the existence of intermediaries and for their rapid growth. Second, to the best of our knowledge, this is the first paper in the supply chain literature that provides a generic, rigorous and highly adaptable foundation for analyzing incomplete contracts in a three-tier, multi-buyer, multi-supplier repeated-sourcing setting. Third, our analysis contributes to the sourcing and procurement literature by bringing together the largely parallel literatures on operational flexibility (cf. Goyal and Netessine (2011)) and relational contracts (cf. Taylor and Plambeck (2007a)). Our analysis captures the changing preferences over suppliers, central to the operational flexibility literature and the incomplete contractability that drives results in the relational contracting literature. Our analysis illustrates the trade-off between the opposite sourcing strategies prescribed in the two streams and demonstrates how mediated sourcing breaks the trade-off.

2. Literature Review

Strategies for sourcing have been a central focus of recent research in supply chain management. Work on flexible sourcing to manage changing sourcing needs and relational contracts to deal with contractual incompleteness are most relevant to our study.
Studies on Flexible Sourcing. Flexible sourcing or responsively sourcing from multiple suppliers has been suggested as a strategy to deal with the changing business environment. Kouvelis et al. (2004) demonstrates the pervasive exposure of global sourcing firms to risks arising out of subsidized financing, tariffs, regional trade rules and taxation. Kouvelis et al. (2001) and Ding et al. (2007) develop strategies for a firm that faces uncertainty in currency exchange rates. Allon and Van Mieghem (2010), Lu and Van Mieghem (2009) study the choice between sole and dual sourcing strategies and consider the influence of changing logistics costs and foreign trade barriers on configuring global networks. Finally, Tomlin (2006) and Chod et al. (2010) examine the value of these flexible sourcing strategies under different contingencies. In line with this literature, our model allows for buyers to have changing preferences over suppliers and is agnostic to the source of these changing preferences, thus allowing us to address each of the reasons highlighted above.

Studies on Relational Contracting. This literature addresses the inefficiencies that arise due to the profit-relevant non-contractible actions of sourcing partners. This has been a central focus of microeconomics research for over three decades (cf. Aghion and Holden (2011) for a summary), and there is a growing body of operations literature that highlights the use of informal agreements (relational contracts) as a remedy to these inefficiencies. Taylor and Plambeck (2007a,b) study settings where price and capacity are non-contractible and are the primary sources of inefficiency. Debo and Sun (2004) study a setup where inventory levels are non-contractible and where building long-term relationships can mitigate the traditional double marginalization losses. Plambeck and Taylor (2006) study joint production with unobservable utility-relevant actions and a non-contractible output yield. Ren et al. (2010) consider forecast sharing by a buyer in a setup where he has an incentive to inflate the forecasts. In each of these studies, building long-term relationships is presented as a mechanism for providing inter-temporal incentives that mitigate myopic opportunistic behavior. In line with this literature, the transaction step game of our model (introduced in Section 3.1) captures these non-contractible aspects of sourcing interactions. As in our treatment of changing buyer preferences over suppliers, rather than model any of the specific non-contractible actions studied in this literature, we consider a generic game that captures the key elements of each of the above settings.

Trade-off between Flexible Sourcing and Relational Contracting. Flexible sourcing and relational contracting are competing strategies. Tunca and Zenios (2006) consider the trade-off between relational contracts and flexible procurement auctions in a setting with multiple buyers and sellers. Swinney and Netessine (2009) look at the same trade-off when there is a possibility of supplier
bankruptcy or default. Li and Debo (2005, 2009) illustrate the long-term shortcomings and benefits of committing to source from a single supplier when future sourcing options may change.

Our study continues in the tradition of examining the trade-off between relational contracts and flexible sourcing, and we demonstrate the utility of mediated sourcing in relieving this trade-off. While sourcing intermediaries have never before been studied in the context of this trade-off, intermediaries and the reasons for their existence have been extensively studied, a comprehensive summary can be found in Wu (2004). We add to this literature by providing a novel reason for the existence of intermediaries. Finally, while most of the cited literature pertains to supply chains for physical goods, the sourcing issues described above arise equally in the outsourcing of services (Ren and Zhang (2009); Roels et al. (2010)).

3. Model Setup, Direct and Mediated Sourcing

3.1. Model Preliminaries. Consider two buyers, $b_1$ and $b_2$, that repeatedly source products or services available from two potential suppliers, $s_1$ and $s_2$. Each supplier has ample capacity and the capability to meet the sourcing needs of one or both buyers. Buyers and suppliers discount future profits at a discount rate $\delta$, which captures the time value of money and the probability of exit of the buyers and suppliers from the market.

We model the repeated trade between these buyers and suppliers as an infinitely repeated game— in each stage game, both buyers source the product. The sourcing exercise itself proceeds in three steps (Figure 3.1). First is the Information Gathering step, where the differences in the costs of sourcing from the two suppliers are revealed. Second is the Supplier Selection step, where each buyer’s business is distributed amongst the two suppliers. Finally, the product or service is actually sourced in the Transaction step. These three steps constitute the stage game that is repeated in every period $t \in \{0, 1, 2, \ldots\}$. We describe the three steps in detail below.

**The Information Gathering Step.** In this step, buyers acquire information about the prices, capabilities and performance of different suppliers to ascertain the advantage of one supplier over the other. This advantage could arise out of a match between the buyers’ product specifications
and the suppliers’ idiosyncratic capabilities, or differences in exchange rates, transportation or telecommunication costs, cross-border tariffs, pass-through input costs, etc. To capture the dynamic business environment and the evolution of the buyers’ business, we allow this relative advantage to change stochastically from one sourcing period to another. In particular, at time $t$, the profits of buyer $i$, $i \in \{1, 2\}$ if he sources from supplier 1, include an additive component, $X^t_i$, the relative advantage of supplier 1 in supplying buyer $i$, that is publicly drawn from a probability distribution function that has both positive and negative support and can be asymmetric. $F^t \left(X^t_1, X^t_2\right)$ denotes this joint bi-variate distribution of the relative advantage that supplier 1 has in supplying buyer 1 and 2. $F^t_1$ and $F^t_2$ are the partial densities. All else being equal, if the realization of $X^t_i$ is positive, buyer $i$’s profits will be higher if he sources from supplier 1 than from supplier 2, and supplier 1 is the current preferred or, taking a total cost of ownership view, the “lower-cost” supplier. Note here that we make no assumptions on the stationarity of the buyers’ preferences over suppliers, nor do we assume that the buyers are symmetric. Our setup allows heterogeneous buyers’ preferences over suppliers to randomly and systematically vary over time, in both their direction and intensity, in an arbitrary fashion.

**The Supplier Selection Step.** In this stage, the sourcing business is allocated between the two suppliers. To facilitate clear illustration, we assume that the two buyer’s sourcing needs are comparable in dollar value, and without loss of generality, we normalize that value to one unit.\(^2\)

**The Transaction Step.** The actual sourcing of the product or service takes place in this step. Both the buyer and supplier can now undertake some actions that influence the profits of their sourcing partner. On the supplier side, these could include operational actions such as efforts in ensuring quality, ensuring timely delivery, conforming to technical and labor standards, following environmental and social norms, maintaining confidentiality of proprietary information, providing prompt after-sales service and support, etc. On the buyer’s side, these could include accurate sharing of demand information, timely payments, access to new business opportunities, cross-investments, access to capital, training, technology transfer, recommendations, rewards, sanctions, etc.

We model all buyer-supplier interactions in the transaction step as a completely general finite two-player game that can capture any economic interactions during the sourcing stage between the buyer and supplier, including those mentioned above. We denote the extensive form of this generic game by $\Gamma$. In game $\Gamma$, the set of buyer and supplier feasible actions are denoted as $A_b, A_s \subset \mathbb{R}^n$. The set of feasible action profiles is then given by $A \equiv A_s \times A_b$. Each element of set $A$, $a$, describes the

\(^2\)A simple extension allows us to consider buyers with different sourcing budgets. All effects presented below continue to hold.
actions undertaken by the two players in this game. On completion of game \( \Gamma \), the action profile \( a \) is perfectly and publicly observable. Buyer and supplier profits are given by arbitrary, general profit functions \( u_b, u_s \colon A \to \mathbb{R} \). We denote the Nash equilibrium of game \( \Gamma \) as \( a^N \in A \), associated with actions corresponding to “opportunistic behavior”, and we assume that it is unique and the payoff associated with it is inefficient. In particular, there exists an efficient outcome \( a^C \in A \), associated with “cooperative behavior”, that makes each player better off.\(^3\)

The above setup allows any number of sequential or simultaneous buyer or supplier actions, and the profits can be any arbitrary function of these actions. We consider situations where self-interested behavior and the consequent Nash equilibrium outcome are inefficient. The classic prisoner’s dilemma type game is a simple example of the game, \( \Gamma \). In the sourcing context, game \( \Gamma \) captures situations where incomplete contracts and incentive misalignment lead to a departure from first-best behavior. This departure could arise on account of poor performance on unobservable quality dimensions and the accompanying low prices (Tunca and Zenios (2006)), insufficient investments in unverifiable capacity (Taylor and Plambeck (2007b)), inefficiencies due to limited information sharing (Ren et al. (2010)), etc. Additionally, our setup also captures some key decentralization issues from the service outsourcing literature related to service quality, capacity building, utilization, etc. (Ren and Zhang (2009); Roels et al. (2010)).

**Alternate Supply Chain Structures.** We model and compare two alternate sourcing structures:

1. **Direct Sourcing:** Each buyer sources directly from the suppliers. The buyers allocate business between suppliers, and each buyer acts in the transaction step.

2. **Mediated Sourcing:** Both buyers source through a third party, the Intermediary. The intermediary allocates business between the suppliers for each buyer and the intermediary acts for both buyers in the transaction step.

Finally, note that in our setup there are no fixed investments, fixed order costs or other scale advantages, nor are there any information asymmetries or any benefits from information aggregation. Thus, the previously documented transactional and informational advantages of mediation do not exist in our setup. Based on existing theory, mediated sourcing should offer no advantage over direct sourcing. In fact, in the presence of incentive misalignment, one would apriori expect vertical integration and reduction of tiers to be superior due to limited incentive misalignment. In the next sections, we describe the game in each of the two sourcing setups, and we compare the equilibrium

\(^3\)Note here that the payoffs are a function of the actions of only the buyer and the supplier engaged in the transaction and are not affected by transactions that the buyer/supplier undertakes with other players.
outcomes in Section 4. A formal, technical description of the games and the equilibria is provided in the Appendix.

3.2. Direct Sourcing. In direct sourcing, the buyers act independently, and their choices can be analyzed in two identical but distinct games. We analyze buyer $i$’s game next.

The Stage Game at Time $t$. Figure 3.2 illustrates the stage game played between buyer $i$, and suppliers 1 and 2.

First, the random cost advantage of supplier 1 in supplying buyer $i$, $X^t_i \sim F^t_i(x)$, is drawn. Next, buyer $i$ sources a fraction $\theta^t_i\{X^t_i\} \rightarrow [0,1]$, from supplier 1 and $1-\theta^t_i\{X^t_i\}$ from supplier 2. Finally, buyer $i$ and each of the suppliers play the transaction step subgame $\Gamma$. We denote the game involving buyer $i$ and supplier $j$ as $\Gamma^{ij}$, and the actions in this game are denoted as $a^{t}_{ij} \in A$. The stage game payoffs are:

$$
\begin{align*}
\alpha_{b}^* &= \theta^t_i \cdot u(b(a^t_{i1}) + X^t_i) + (1 - \theta^t_i) \cdot u(b(a^t_{i2})) , \\
\alpha_{s1} &= \theta^t_i \cdot u_s(a^t_{i1}) , \\
\alpha_{s2} &= (1 - \theta^t_i) \cdot u_s(a^t_{i2}) .
\end{align*}
$$

The action profile $\alpha^*_{si}$, which prescribes setting $\theta^t_i$ as $\theta^t_i \equiv I(X^t_i \geq 0)$ followed by actions $a^t_{i1} = a^t_{i2} = a^N$, is a subgame-perfect equilibrium of the direct sourcing stage game, where $I(\cdot)$, the indicator function, is 1 when the condition is satisfied (formally shown in Lemma 3 in the Appendix).

The Repeated Game. In the repeated game, the stage game is played in each period $t \in \{0,1,2,...\}$.

Potential Equilibrium Strategies. In each of the two transaction step games the players may play the cooperative or the Nash actions. Specifically, three kinds of behavior may arise in equilibrium: 1) the buyer and both suppliers always play the Nash actions, or 2) the buyer and one supplier (supplier 1 or supplier 2) play the cooperative actions in the transaction games that involve them, while the buyer and the other supplier play Nash actions in the transaction step game; or 3) the buyer and both suppliers always play the cooperative action. We call these the direct transactional

---

4For concise representation, in subsequent discussions we suppress the argument of $\theta^t_i\{X^t_i\}$ and simply use $\theta^t_i$. 
the direct single relationship \((d_1s_1)\) or \((d_1s_2)\) and the direct dual relationship \((d_id)\) sourcing strategies, respectively. Note that in all three of these strategies, at any time, the buyer can choose freely to source from both the suppliers or just one of them. The difference lies in the choice of suppliers with which the buyer decides to play the cooperative outcome, or the supplier(s) with whom the buyer enters into a so-called long-term relationship (Taylor and Plambeck (2007b)).

Formally, \(\forall k \in \{t, s_1, s_2, d\}\), strategy \(\sigma^{d, k}(\theta_i)\), \(\theta_i \equiv \{\theta^t_i, t \geq 0\}\), prescribes the following play: if in all past play, the outcomes of the selection and transaction step actions prescribed below were observed, continue to play the corresponding selection and transaction step actions; else play action \(\alpha^{*}_{d_i}\) (the stage game equilibrium) forever.\(^5\)

**Selection Step Actions:** The allocation of business at each stage of the game is prescribed by \(\theta_i\). At time \(t\), the amount sourced from supplier 1 (the sourcing fraction) is given by the \(t^{th}\) element of sequence \(\theta_i^t\), \(\theta^t_i(X^*_t)\).

**Transaction Step Actions:** The prescribed actions are \((a^N, a^N)\) for strategy \(d_1t\), \((a^C, a^N)\) for strategy \(d_is_1\), \((a^N, a^C)\) for strategy \(d_is_2\), and \((a^C, a^C)\) for strategy \(d_id\), where the first element denotes the actions in the transaction game with supplier 1, and the second with supplier 2.

The present value of the expected normalized profit of player \(n\), \(n \in \{s_1, s_2, b_i\}\), under strategy \(\sigma\) is given by

\[
U_n(\sigma) = (1 - \delta) \sum_{t=0}^{\infty} \delta^t E\left[u^t_n(X^*_t, \theta^t_i(\sigma), a^t_{11}(\sigma), a^t_{12}(\sigma))\right].
\]

Further, define operator \(\delta^t(u^t_n) \equiv (1 - \delta) \sum_{r=t}^{\infty} \delta^{r-t} E\left[u^r_n\right]\), where the expectation is taken over each \(X^*_1\) and \(X^*_2\) using \(F^{r}\). Given a payoff stream, \(u^t_n\), the operator, \(\delta^t(u^t_n)\), denotes the normalized expected present value of this payoff stream starting from period \(t\). Applying Equation 3.1 to the four potential equilibrium strategies described above gives us the expected normalized discounted profits earned by following each of the strategies.

The buyers’ profits from any strategy depend on the degree of relational sourcing and the allocation of business amongst suppliers. In particular, all else being equal, the strategies with more relationships (dual > single > transactional) and strategies in which \(\theta_i\) is chosen "responsively", i.e after observing each \(X^*_t\), element, \(\theta^t_i\) is chosen to maximize the current period payoff, provide the highest profit. For dual sourcing, this responsive \(\theta_i\) is \(\tilde{\theta}_i \equiv \{\tilde{\theta}^t_i, t \geq 0\}\), which dictates always sourcing everything from the lower-cost supplier. However, the ability to sustain the above strategy profiles

\(^5\)These and all the other strategies proposed in this paper are Nash reversion trigger strategies, that is, on observation of a deviation from the equilibrium path, the Nash outcome is played in all future periods. Following Taylor and Plambeck (2007b) we restrict our attention to the trigger strategies.
as subgame-perfect equilibria of the repeated game depends on the incentives for the buyers and the suppliers to deviate from the strategy. The next lemma provides restrictions on \( \theta_i \) that ensure that the strategy is an equilibrium.

**Lemma 1. Equilibrium Outcomes of the Direct Sourcing Game.**

1. The strategy profile \( \sigma^{d, t}(\tilde{\theta}_i) \) is the only transactional subgame-perfect equilibrium of the repeated game.

2. The strategy profile \( \sigma^{d, k}(\theta_i) \) is a subgame-perfect equilibrium of the repeated game if and only if, for all \( t \geq 0 \) and all \( X^i_t \), the difference between each player’s expected normalized continuation profit from this strategy exceeds profit from the above transactional equilibrium by at least the values provided in the table below.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Buyer i</th>
<th>Supplier 1</th>
<th>Supplier 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma^{d, s_1}(\theta_i) )</td>
<td>( \frac{1-\delta}{\delta} \max { \tilde{\theta}^i_t G_b, (\tilde{\theta}^i_t - \theta^i_t) X^i_t - \eta_t } )</td>
<td>( \frac{1-\delta}{\delta} G_s \tilde{\theta}^i_t )</td>
<td>( \frac{1-\delta}{\delta} G_s (1 - \theta^i_t) )</td>
</tr>
<tr>
<td>( \sigma^{d, s_2}(\theta_i) )</td>
<td>( \frac{1-\delta}{\delta} \max { (1 - \theta^i_t) G_b, (\tilde{\theta}^i_t - \theta^i_t) X^i_t - \eta_t (1 - \theta^i_t) } )</td>
<td>( \frac{1-\delta}{\delta} G_s (1 - \theta^i_t) )</td>
<td>( \frac{1-\delta}{\delta} G_s (1 - \theta^i_t) )</td>
</tr>
<tr>
<td>( \sigma^{d, d}(\theta_i) )</td>
<td>( \frac{1-\delta}{\delta} \max { G_b, (\tilde{\theta}^i_t - \theta^i_t) X^i_t - \eta_t } )</td>
<td>( \frac{1-\delta}{\delta} G_s \tilde{\theta}^i_t )</td>
<td>( \frac{1-\delta}{\delta} G_s (1 - \theta^i_t) )</td>
</tr>
</tbody>
</table>

\( \tilde{\theta}^i_t \equiv \max X^i_t, \theta^i_t (X^i_t), \theta^i_t \equiv \min X^i_t, \theta^i_t (X^i_t) \) are the maximum and minimum amount of business allocated to supplier 1 in any state; \( G_s \) and \( G_b \) denote the gain from the most profitable deviations of the supplier and buyer in the subgame \( \Gamma \). \( \eta_t \equiv u_b (a^C) - u_b (a^X), \eta_t \equiv u_s (a^C) - u_s (a^X) \) are the buyer’s and supplier’s gain from cooperation.

**Proof.** The formal proof is provided in the Appendix (Page 28), and the intuition follows. In direct transactional sourcing, player actions do not influence subsequent stages of the game. Thus, the stage game equilibrium, played in every period, is the subgame-perfect equilibrium of repeated direct sourcing game. Sustaining the latter three relational strategy profiles as equilibrium outcomes requires that the immediate gains from the most profitable deviation should be smaller than the loss in the continuation benefits. The loss in continuation benefits is given by the difference in the profits earned by following the relational strategy and the profits from the transactional sourcing strategy (recall that on observation of any deviation, all players resort to following the transactional strategy). The expressions in Part 2 of the theorem capture the immediate gains from the most profitable deviation. The most profitable deviation arises when the maximum amount of business is transacted with cooperative behavior (\( \tilde{\theta}^i_t \) for supplier 1 and \( 1 - \theta^i_t \) for supplier 2). Further, for the buyer there is a deviation possible both in the selection step and in the transaction step. The more profitable deviation of these two defines the immediate gain of deviation for the buyer. □

---

This is the difference between the profit of the best-response action to the cooperative actions of the other player in game \( \Gamma \), and the profit of the cooperative action. Formally, \( G_s = \max_{a'_s, \xi} \{ u_s (a'_s, a^C_s | \xi) - u_s (a^C_s | \xi) \} \), \( G_b = \max_{a'_b, \xi} \{ u_b (a'_b, a^C_b | \xi) - u_b (a^C_b | \xi) \} \), where \( a'_s, a'_b \) are arbitrary supplier and buyer actions. \( \xi \) is a typical element of the set \( \Xi \), the collection of all initial nodes of all subgames of the extensive-form game \( \Gamma \).
Recall that the buyer’s profits are highest in the dual relationship strategy with responsive allocation, $\theta_i = \tilde{\theta}_i$. However, to sustain any relationship and allocation in equilibrium, the buyer must restrict the allocation as per the conditions in Lemma 1, departing from the responsive allocation. This tension between responsive allocation and the provision of the incentives to sustain relationships and achieve cooperative outcomes is a key characteristic of sourcing that our stylized model is designed to capture.

This tension is captured in Figure 3.3. For any given discount factor, the figure illustrates the best equilibrium profits that can be achieved.\footnote{As is typical in repeated games, we express our equilibrium conditions in terms of the discount factor. However, these conditions can equally be interpreted as conditions on all exogenous parameters: the distribution, $F_t(x)$, the general profit functions, $u_b$ and $u_s$, the gains from deviation $G_b$ and $G_s$, and the benefits from cooperation $\eta_{b(s)} = u_{b(s)}(a^n) - u_{b(s)}(a^n)$.}

Formally, as is typical in analyses of repeated games and in statements of Folk Theorems (Mailath and Samuelson (2006)), this figure illustrates the achievable payoff region of the buyer as a function of the discount factor. ∀δ, this is computed as

$$\max_{k \in \{t,s_1,s_2,d\}} \max_{\theta_i} U_{b_i}(\sigma_{d,k}(\theta_i)),$$

s.t. strategy $\sigma_{d,k}(\theta_i)$ is an equilibrium.

Two characteristics of the equilibrium conditions in Lemma 1 help us to understand this achievable payoff region. First, the equilibrium conditions for dual relationship strategies are more restrictive than the conditions for single relationship strategies (in dual sourcing, sufficient incentives need to be provided to the two suppliers, while in single relationship only to one). Second, for both dual or single relationships, the trade-off between responsive allocation and the provision of sufficient business to maintain relationship(s) is more restrictive as the discount factor is smaller and the suppliers value future business less, thus requiring larger and larger departures from the responsive allocation to sustain relationships.
For the highest values of $\delta$, region (v) in Figure 3.3, future business is valued highly by suppliers and the buyer can potentially maintain relationships with both suppliers while also allocating business responsively. Put differently, in this region, $\delta$ is high enough that the equilibrium conditions for even dual relationships are not binding. However, as $\delta$ gets smaller, the conditions become binding, and the buyer must now sacrifice the responsive allocation to maintain the two relationships and this decreases his profits (region (iv)). Next, at some point, the equilibrium conditions become so tight that no allocation can satisfy the dual relationship equilibria conditions, but single relationship equilibria may be sustained, first with responsive allocation and then potentially with restricted allocation (regions (iii) and (ii)). Eventually, only transactional sourcing can be sustained as an equilibria, region (i). In subsequent sections, we will illustrate how the tradeoffs shown in Figure 3.3 change with mediated sourcing.

3.3. Mediated Sourcing. With mediated sourcing, both buyers delegate their supplier selection and their transaction step actions to a third party, the intermediary. The intermediary chooses the supplier for each buyer and acts on behalf of buyers in the transaction step. In lieu of the sourcing services provided by the intermediary, the buyers pay the intermediary an agreed upon commission. Specifically, the intermediary gets a fraction, $\beta$, of the total buyer-side profits. This fraction $\beta$ could arise as a function of a bargaining process prior to signing up for the intermediary’s services or by any other mechanism that divides the total profits generated.

In the setup described above, buyers do not have any profit-relevant actions after they have signed up for the intermediary’s services. As such, they are no longer relevant players in the mediated sourcing game. In essence, the mediated sourcing game follows along exactly the same lines as the two direct sourcing games, except that the actions of the two individual buyers are now taken by one intermediary. In all other respects, the two described structures are identical.

The Stage Game, at Time t. First, in the information gathering step, the differences in sourcing from different suppliers are revealed, i.e. $X^1_t$ and $X^2_t$ are drawn from joint distribution $F^t(X^1_t, X^2_t)$. Next, the intermediary allocates a fraction $\nu^t_i : \{X^1_t, X^2_t\} \to [0, 1]$ of buyer $i$’s sourcing business to supplier 1, $i \in \{1, 2\}$. Note that the allocations $\nu^t_i$ correspond to the allocations $\theta^t_i$ from direct sourcing, but now the allocations are a function of the relative cost advantage of supplier 1 in supplying both buyer 1 and 2, $X^1_t$ and $X^2_t$. Put differently, the intermediary takes into account both buyers’ preferences for a supplier in the sourcing decision. We denote the allocations of buyer 1 and buyer 2’s business to supplier 1, $(\nu^t_1, \nu^t_2)$ as $\nu^t$, and the total business to supplier 1, $\nu^t_1 + \nu^t_2$, is

Note here that in the illustration, we show the single relationship equilibria with one supplier. With heterogeneous suppliers it is possible that we may have two single relationship regimes, one for each of the suppliers.
denoted by $\langle \nu^t \rangle$. Finally, actual sourcing takes place in the transaction step, and the intermediary and the suppliers play transaction games $\Gamma$. The games are identical to the ones that buyers play in direct sourcing, except that buyers are replaced by the intermediary. We denote the game between the intermediary and supplier $j$ as $\Gamma^{Ij}$ and the actions in this game as $a_{Ij}$. Finally, the suppliers, the buyers and the intermediary earn their profits. The profits are given as

$$u_I = \beta \sum_{i=1}^{2} \left( \nu_t^i \left( u_b(a_{I1}^I) + X_t^i \right) + (1 - \nu_t^i) u_b(a_{I2}^I) \right),$$

$$u_{s_1} = \langle \nu_t^i \rangle u_s(a_{I1}^I), \quad u_{s_2} = (2 - \langle \nu_t^i \rangle) u_s(a_{I2}^I).$$

The action profile $\alpha_m^*$, that prescribes $\nu^t = \tilde{\nu}^t \equiv (\tilde{\theta}_1^t, \tilde{\theta}_2^t)$; and actions $a^N$ in the transaction step games is a subgame-perfect equilibrium of the mediated sourcing stage game (Appendix, Lemma 4).

**The Repeated Game.** In the repeated game, the stage game is played in each period $t \in \{0, 1, 2, \ldots\}$.

**Potential Equilibrium Strategies.** Supplier selection is a function of the realization of the relative cost advantages, $X_t^1$ and $X_t^2$. With respect to transaction step actions, the choices follow along the same lines as those in direct sourcing. Specifically, the intermediary and the chosen supplier(s) may play Nash actions in all games, or the intermediary and one supplier may play cooperative actions in transaction games that involve them and Nash actions in the transaction games that involve the other supplier, or the intermediary and each supplier may always play the cooperative action. We call these the mediated transactional ($mt$), single relationship ($ms_1$ or $ms_2$), and dual relationship ($md$) strategies, respectively.

Formally, $\forall k \in \{t, s_1, s_2, d\}$, strategy $\sigma^{mk}(\nu), \nu \equiv \{\nu^t, t \geq 0\}$ prescribes the following play: if in all past play only outcomes of actions prescribed below were observed, continue to play the corresponding selection and transaction step actions, else play action $\alpha_m^*$ (the stage game equilibrium) in all subsequent stage games.

**Selection Step Actions:** At time $t$, the amount sourced from supplier 1 for buyers 1 and 2 is given by the $t^{th}$ component of sequence $\nu$, $\nu^t$.

**Transaction Step Actions:** The prescribed actions are $(a^N, a^N)$ for strategy $mt$; $(a^C, a^N)$ for strategy $ms_1$; $(a^N, a^C)$ for strategy $ms_2$; and $(a^C, a^C)$ for strategy $md$. The first action denotes the actions in the game with supplier 1 and the second with supplier 2.

---

9Note that while $\nu_1^t, \nu_2^t, \langle \nu^t \rangle$ and $\nu^t$ are all functions of $X_t^1$ and $X_t^2$, we often suppress the arguments in subsequent discussion.
As before, the profits are highest with dual relationship strategies, when \( \nu \) is chosen responsively, \( \nu = \tilde{\nu} \equiv \{ \tilde{\nu}^t, t \geq 0 \} \). Next we provide the necessary and sufficient conditions to sustain a strategy profile \( \sigma^{mk}(\nu) \) as an equilibrium.

**Lemma 2. Equilibrium Outcomes of the Mediated Sourcing Game**

1. The strategy profile \( \sigma^{mt}(\tilde{\nu}) \) is the only transactional subgame-perfect equilibrium of the mediated sourcing game.

2. The strategy profile \( \sigma^{mk}(\nu) \) is a subgame-perfect equilibrium of the repeated game if and only if, for all \( t \geq 0 \) and for all \( X_{1}^t \) and \( X_{2}^t \), the difference between each player’s expected normalized continuation profit from this strategy exceeds profit from the above transactional equilibrium by at least the values provided in the table below.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Intermediary</th>
<th>Supplier 1</th>
<th>Supplier 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma^{ms1}(\nu) )</td>
<td>( \frac{1-\delta}{\delta} \beta \max \left{ \langle \tilde{\nu} \rangle G_b, \sum_{i=1}^{2} \left( \left( \tilde{\theta}^i_t - \nu^i_t \right) X_{i}^t - \eta b \nu^i_t \right) \right} )</td>
<td>( \frac{1-\delta}{\delta} G_s \langle \tilde{\nu} \rangle )</td>
<td>( \frac{1-\delta}{\delta} G_s \langle \tilde{\nu} \rangle )</td>
</tr>
<tr>
<td>( \sigma^{ms2}(\nu) )</td>
<td>( \frac{1-\delta}{\delta} \beta \max \left{ (2 - \langle \tilde{\nu} \rangle) G_b, \sum_{i=1}^{2} \left( \left( \tilde{\theta}^i_t - \nu^i_t \right) X_{i}^t - \eta b (1 - \nu^i_t) \right) \right} )</td>
<td>( \frac{1-\delta}{\delta} G_s (2 - \langle \tilde{\nu} \rangle) )</td>
<td>( \frac{1-\delta}{\delta} G_s (2 - \langle \tilde{\nu} \rangle) )</td>
</tr>
<tr>
<td>( \sigma^{md}(\nu) )</td>
<td>( \frac{1-\delta}{\delta} \beta \max \left{ 2 G_b, \sum_{i=1}^{2} \left( \left( \tilde{\theta}^i_t - \nu^i_t \right) X_{i}^t - \eta b \right) \right} )</td>
<td>( \frac{1-\delta}{\delta} G_s \langle \tilde{\nu} \rangle )</td>
<td>( \frac{1-\delta}{\delta} G_s \langle \tilde{\nu} \rangle )</td>
</tr>
</tbody>
</table>

\( \langle \tilde{\nu} \rangle \equiv \max X_{1}^t, X_{2}^t \langle \nu^t \rangle \) and \( \langle \tilde{\nu} \rangle \equiv \min X_{1}^t, X_{2}^t \langle \nu^t \rangle \) are the maximum and minimum amount of business allocated to supplier 1 in any state. \( G_s (G_b) \) denotes the gain from the most profitable deviation (defined as before).

**Proof.** A formal proof is provided in the Appendix (Page 30).

Like the direct buyers, the intermediary acting on behalf of the two buyers in mediated sourcing faces a trade-off. Profits are increased by establishing relationships and by responsive allocation, but the intermediary may need to restrict his business allocation to sustain relationship(s) in equilibrium (Lemma 2). Further, as before, dual relationships are harder to sustain than single relationships, and all relationships are harder with lower values of the discount factor. Thus, the achievable payoff has a similar shape to the one illustrated for direct sourcing in 3.3. However, there is one difference between this trade-off for mediated sourcing and direct sourcing. Rather than an individual buyer sourcing for himself, the intermediary is now sourcing on behalf of both buyers. This implies that the intermediary’s allocation of business to the two suppliers is based on business accruing from the two buyers and his total costs are a function of both \( X_{1}^t \) and \( X_{2}^t \), i.e. the relative cost difference between suppliers in supplying both buyers 1 and 2. In the next section, we will see how this drives the advantages and disadvantages of mediated sourcing.
4. The “Benefits” of Intermediation

Consider the total buyer-side surplus or the “sourcing profits”, \( \pi \): in the case of direct sourcing, this is the sum of the two buyers’ profits. In the case of mediated sourcing, it is the sum of the buyers’ and the intermediary’s profits. If the buyer-side surplus is higher for the mediated sourcing strategy, then there exists a surplus division factor \( \beta \) such that both buyers and the intermediary are better off under mediated sourcing. Thus, to compare direct and mediated sourcing it is sufficient to compare the respective achievable sourcing profits. For each set of parameter values, the supply chain structure (direct or mediated sourcing) that achieves the higher sourcing profits is the preferred supply chain structure. Note that using sourcing profits for comparing strategies also brings scale parity between direct and mediated sourcing— in both cases, we are comparing the profits from sourcing two units from the suppliers.

Recall that the achievable sourcing profit regions were obtained by choosing the highest profit strategy that is also an equilibrium for a given set of parameter values. For both direct and mediated sourcing, the strategy space can be characterized by the type of relationship(s) (transactional, single relationship, dual relationship; \( k \in \{ t, s_1, s_2, d \} \)) and the allocation of business between suppliers (choice of \( \theta_i/\nu \)). Thus, to find the highest profit strategy that is an equilibrium, we need to consider the choice of relationship type and the choice of business allocation. To build our intuition, we first consider the highest equilibrium sourcing profit for a given type of relationship.

**Definition.** \( \forall \delta, i \) and \( k \), define \( \pi^{d,i,k}(\delta) = \max_{\theta_i} \pi(\sigma^{d,i,k}(\theta_i)) \), such that strategy \( \sigma^{d,i,k}(\theta_i) \) is an equilibrium of the direct sourcing game for this \( \delta \). Similarly, define \( \pi^{m,k}(\delta) = \max_{\nu} \pi(\sigma^{m,k}(\nu)) \), such that strategy \( \sigma^{m,k}(\nu) \) is the equilibrium of the mediated sourcing game for this \( \delta \). For any given type of relationship \( k \), \( \pi^{d,k}(\delta) \) and \( \pi^{m,k}(\delta) \) are the highest sourcing profits that are achievable as equilibria, considering all different possible allocations of business.

4.1. **Ability to Sustain Relationships.** The next Theorem compares the ability of direct and mediated sourcing in sustaining a given type of relationship.

**Theorem 1.** \( \forall \delta, k \in \{ s_1, s_2, d \} \), sourcing through an intermediary earns higher sourcing profits than if both buyers sourced directly with the same relationship:

\[
\forall \delta, k \quad \pi^{m,k}(\delta) \geq \pi^{d_1,k}(\delta) + \pi^{d_2,k}(\delta),
\]

with strict inequality for some \( \delta \).

**Proof.** A formal proof is provided in the Appendix (Page 30). \( \square \)
Sketch of the Proof: For any given type of relationship $k$, we can write the best sourcing profit for direct sourcing as

$$
\pi^{d_1 k}(\delta) + \pi^{d_2 k}(\delta) = \max_{\theta_1, \theta_2} \mathcal{E}^0 \left( \Pi^{k, \tau}(\theta_1, \theta_2) \right),
$$

s.t. $\theta_1, \theta_2 \in \mathcal{D}(\theta_1, \theta_2)$,

where the set $\mathcal{D}$ denotes the feasible set defined by the equilibrium conditions for direct sourcing strategy $k$. Interestingly, the mediated sourcing profit, $\pi^{mk}(\delta)$, can be written with exactly the same objective function, but with a different feasible set, $\mathcal{M}$:

$$
\pi^{mk}(\delta) = \max_{\nu_1, \nu_2} \mathcal{E}^0 \left( \Pi^{k, \tau}(\nu_1, \nu_2) \right),
$$

s.t. $\nu_1, \nu_2 \in \mathcal{M}(\nu_1, \nu_2)$.

This suggests that the difference between mediated and direct sourcing can be understood by examining the set of equilibrium conditions, $\mathcal{D}(\theta_1, \theta_2)$ and $\mathcal{M}(\nu_1, \nu_2)$. It is most instructive to compare the conditions that come from the suppliers’ incentives, for example, consider supplier 1’s incentives.

<table>
<thead>
<tr>
<th>Direct, $\mathcal{D}(\theta_1, \theta_2)$</th>
<th>Mediated, $\mathcal{M}(\nu_1, \nu_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D.1) $\mathcal{E}^{t+1}(\theta_1) - d\theta_1 \geq \gamma_1^t$</td>
<td>(M) $\mathcal{E}^{t+1}(\nu_1 + \nu_2) - d(\nu_1 + \nu_2) \geq \gamma_1^t + \gamma_2^t$</td>
</tr>
<tr>
<td>(D.2) $\mathcal{E}^{t+1}(\theta_2) - d\theta_2 \geq \gamma_2^t$</td>
<td></td>
</tr>
</tbody>
</table>

where, $\gamma_i^t \equiv \mathcal{E}^{t+1}(1 - F^r(0))\frac{u_{s, \mathcal{A}}(a^n)}{u_{s, \mathcal{C}}(a^n)}$, $d \equiv \frac{1-\delta}{\lambda} \frac{G_s}{u_{s, \mathcal{A}}}.$

Essentially, for direct sourcing, buyers 1 and 2 must each individually ensure that their stream of orders, $\theta_1$ or $\theta_2$, is such that the supplier has an incentive to continue the relationship. Conditions (D.1) and (D.2) reflect this. In mediated sourcing, on the other hand, the intermediary must only ensure that the stream of orders arising from the combination of the streams of orders on behalf of buyer 1 and 2, $\nu_1 + \nu_2$, must be such that the supplier has an incentive to continue the relationship. Condition (M) reflects this. Essentially, the condition for maintaining a mediated relationship is the sum of the conditions for maintaining equivalent direct relationships. Thus, the equilibrium conditions for direct relationships are a subset of the conditions for mediated relationships, and mediated sourcing always (weakly) outperforms direct sourcing for a given relationship. Also, note by looking at the RHS of the above equations that the combined stream of orders, while potentially larger than any single buyer’s order stream, must also cross a higher threshold. Put differently,

$^{10} \Pi^{k, \tau}(x, y) \equiv (x^r + y^r)u_b(a_1^k) + x^r X_1^* + y^r X_2^* + (2 - (x^r + y^r))u_b(a_2^k), a_i^k$ denotes $a_{i,j}$ or $a_{i,j}$ depending on the context.

$^{11}$Supplier 2’s incentives, if applicable (i.e. if $k = d$ or $s_2$), follow along the same lines.
the intermediary does indeed have more scale than any individual buyer, but this scale cuts both ways, providing more incentives to stay in the relationship, but also proportionally more gains from cheating or deviating from the relationship. Thus, the advantage of mediated sourcing that drives the above result is not simply arising from the greater scale of an intermediary. Next, we will illustrate two mechanisms that the intermediary uses to exploit the slack in the constraints for better maintaining the relationship.

To better understand the above effect, consider the following three cases:

Case I: The discount factor is high enough that neither of the constraints, (D.1) or (D.2), are binding. Now, direct buyers can choose the responsive sourcing stream and achieve the highest profits. In such a setup, we show that the constraint (M) will also not be binding, and mediated sourcing will also earn the same profits. Thus, direct and mediated sourcing perform equally well.

Case II: Next, consider the case where the discount factor is a bit lower, and one of the two constraints, (D.1) or (D.2), becomes binding while the other has some slack. This happens when the buyers are heterogeneous in their “long-run preferences” over suppliers, i.e. $\gamma_1 \neq \gamma_2$ or equivalently $\mathbb{E}^{t+1}(I(X_1^t \geq 0)) \neq \mathbb{E}^{t+1}(I(X_2^t \geq 0))$, i.e. the discounted probability that buyer 1 prefers supplier 1 is not equal to the discounted probability that buyer 2 prefers supplier 1. For example, when buyer 2 long-run prefers supplier 1 more than buyer 1, $\exists \delta$, where constraint (D.1) is binding and (D.2) is not binding. Now while one order stream, $\theta_1$, is constrained in a specific fashion, the other, $\theta_2$, is not constrained and can be set to the responsive order stream— the unconstrained optimal. In the case of mediated sourcing, the only constraint, constraint M, is the sum of the constraints D.1 and D.2, and as a result, it is not binding and the order streams on behalf of buyer 1 and the order stream on behalf of buyer 2 both can be set to their responsive or unconstrained maximization values. Essentially, if one buyer long-run prefers a supplier more than the other buyer, mediated sourcing makes it possible to use this buyer’s bias to compensate for the other buyer’s weaker interest. In direct sourcing, the buyer that prefers the particular supplier would find it in its interest to provide more business than strictly necessary, whereas the other buyer would be forced to provide more business than it wants to just to sustain the relationship. Pooling the order streams eliminates this inefficient situation, and the level of business accruing to the supplier on behalf of both buyers can be adjusted to the minimum level sufficient for sustaining the relationship, achieving a responsive allocation. In this fashion, an intermediary can exploit the differences between buyers in their long-run preferences over suppliers to outperform direct sourcing.

Case III: Finally, consider a case where the discount factor is such that both constraints, (D.1) and (D.2), are binding. This arises for low enough $\delta$ or when the buyers are symmetric. Now, constraint
M will also be binding, but the order streams in mediated sourcing will still earn higher profits by being more responsive. Say in direct sourcing, the constrained optimal order streams are $\theta_1^*$ and $\theta_2^*$. Now construct order streams, $\nu_1$ and $\nu_2$ as follows: when $X_t^i \geq X_t^i$, set $\nu_t^i = \min \{1, \theta_t^i + \theta_t^{i\prime} \}$ and $\nu_t^i = \theta_t^i + \theta_t^{i\prime} - \min \{1, \theta_t^i + \theta_t^{i\prime} \}$, where $i$ denotes one buyer and $i\prime$ the other. Hence, by construction, $\forall X_t^1, X_t^2$, $\nu_t^1 + \nu_t^2 = \theta_t^i + \theta_t^{i\prime}$. This order stream is constructed such that from the supplier’s point of view, the orders coming from the two separate buyers or from the intermediary are identical. However, the intermediary bases the composition of the orders on the current realization of the relative cost advantages. In particular, the intermediary ensures that whatever quantity of orders must be sent to the supplier, its composition is such that to the maximum possible degree, it is composed of orders on behalf of the buyer who has a cost advantage of sourcing from this supplier in this sourcing period. Again, the intermediary uses one buyer’s stronger preference for a supplier, $X_t^i \geq X_t^i$, to compensate for the other buyer’s weaker preference. However, this time the difference in preference arises out of the random draws on the relative cost advantage, $X_t^i \neq X_t^j$, or what we call myopic preferences. Thus, an intermediary can exploit the myopic bias of one buyer for a supplier to ensure that the allocation of business is such that the composition of the business allocated to the suppliers is the most advantageous. On the other hand, direct buyers do not have the flexibility to change the composition of the orders going to a supplier, and thus, they often end up choosing a suboptimal composition of orders.

To summarize, mediated sourcing performs better than direct sourcing by adjusting the level of sourcing business allocated to a supplier when the buyers have heterogeneous long-run preferences over suppliers, or by responsively adjusting the composition of sourcing business allocated to a supplier when the buyer’s have different myopic preferences over suppliers. Essentially, with heterogeneous long-run preferences over suppliers, one buyer wants to allocate more business than necessary to ensure cooperative behavior, whereas the other may want to allocate less business than necessary. An intermediary that pools the order streams from both buyers can use one buyer’s above-requirement allocation to compensate for the other buyer’s below-requirement business. Similarly, with different myopic preferences, the supplier can be provided the same incentives for cooperative behavior as in direct sourcing but the composition of that business can be adjusted responsively.

**Corollary. Relationship between Buyers’ Preferences over Suppliers**

1. **Identically Distributed Preferences:** If $X_t^1, X_t^2 \sim F^t(x)$, $\forall \delta, k$, mediated sourcing is better at maintaining a given relationship than a direct buyer: $\pi^{mk}(\delta) \geq \pi^{d1k}(\delta) + \pi^{d2k}(\delta)$, with strict inequality for some $\delta$.

2. **Perfectly Correlated Preferences:** If $X_t^1 = \alpha X_t^2$, 


(a) $\alpha \neq 1$, $\forall \delta, k$ mediated sourcing is better at maintaining a given relationship than a direct buyer: $\pi_{mk}(\delta) \geq \pi_{d1k}(\delta) + \pi_{d2k}(\delta)$, with strict inequality for some $\delta$.

(b) If $\alpha = 1$, $\forall \delta, k$ mediated sourcing has no advantage over direct sourcing, $\pi_{mk}(\delta) = \pi_{d1k}(\delta) + \pi_{d2k}(\delta)$.

(3) **Deterministic Preferences**: If $X_t^i = x_i$, when $t = 2T$, and $X_t^i = -x_i$, when $t = 2T + 1$, where $T \in \{0, 1, 2, \ldots\}$,

(a) if $x_1 \neq x_2$, $\forall \delta, k$ mediated sourcing is better at maintaining a given relationship than a direct buyer: $\pi_{mk}(\delta) \geq \pi_{d1k}(\delta) + \pi_{d2k}(\delta)$, with strict inequality for some $\delta$.

(b) If $x_1 = x_2$, $\forall \delta, k$ mediated sourcing has no advantage over direct sourcing, $\pi_{mk}(\delta) = \pi_{d1k}(\delta) + \pi_{d2k}(\delta)$.

If buyer preferences are identically distributed, or if on average both buyers prefer the same supplier, there are no long-run differences between buyer preferences, $\forall t$, $\gamma_t^1 = \gamma_t^2$, but in each period there is still a chance that the realizations of each buyer’s preferences over suppliers will be different, $\Pr \{X_t^1 \neq X_t^2\} > 0$, and the intermediary can exploit myopic differences as described above. Further, even if the draws are perfectly correlated, with $\alpha \neq 1$, the two draws will still be different and the intermediary can again exploit the difference. However, if $\alpha = 1$, the two realizations will always be identical and there are no benefits from changing the level or composition of orders to a supplier. Finally, if there is no risk involved, that is the shocks are deterministic, but there is still a difference in the buyers’ preferences over suppliers in every period $x_1 \neq x_2$, the intermediary can continue to exploit the resultant differences in myopic and long-run preferences as described above. The above corollary starkly demonstrates that the effects highlighted above accrue from differences in buyer preferences over suppliers. These could arise from myopic differences in preferences over suppliers and/or from systematic or long-run heterogeneity in preferences over suppliers— but as long as there is a possibility that the realized preferences of buyers over suppliers are different at some point in time, mediated sourcing can better maintain relationships. This illustrates that our argument extends beyond the pooling of randomness in preferences to the pooling of random, systematic and temporal differences in preferences.

4.2. The Preferred Supply Chain Structure. In the above section, we illustrated how intermediaries are better at maintaining any given relationship. However, the choice of the preferred supply chain structure depends on the achievable sourcing profits that take into account both the ability to maintain a given relationship and the choice of which relationship to maintain. In this section, we consider both of these effects and identify the preferred supply chain structures.
For any $\delta$, the best achievable sourcing profit in direct sourcing, $\pi^d(\delta)$, is given as $\max_k \pi^d_{1k}(\delta) + \max_k \pi^d_{2k}(\delta)$. Similarly, the best sourcing profit under mediated sourcing is $\pi^m(\delta) = \max_k \pi^m_{k}(\delta)$, where $k \in \{t, s_1, s_2, d\}$.

**Theorem 2.** Mediated sourcing outperforms direct sourcing, i.e. $\forall \delta \pi^m(\delta) \geq \pi^d(\delta)$, with strict inequality for some $\delta$, if the same strategy $k$ is the solution to both $\max_k \pi^d_{1k}(\delta)$ and $\max_k \pi^d_{2k}(\delta)$. This condition always holds when the buyers are ex-ante symmetric in their preferences over suppliers i.e. $\forall t, F^1_t = F^2_t$.

**Proof.** The formal proof is provided in the Appendix (Page 30).

Figure 4.1 illustrates the comparison between direct and mediated sourcing as described in Theorem 2. For the highest values of the discount factor, (region (v)) in both direct and mediated sourcing, the firms can achieve first-best profit, since the responsive allocation stream satisfies the dual relationship equilibrium conditions. For lower values, (region (iv)) one of the buyers’ responsive allocation streams is no longer sufficient for sustaining the dual relationship. In direct sourcing, this buyer must now shift to a less responsive allocation stream, but the intermediary can use the slack in the other buyer’s responsive allocation to still satisfy the supplier (long-run differences). Although, for even lower discount factors (region (iii)), both buyer’s responsive allocation streams may now be insufficient for the supplier(s), mediated sourcing can still exploit the changing preferences that

![Figure 4.1. Mediated Sourcing Outperforms Direct Sourcing](image)
lead to myopic differences to earn higher sourcing profits. For even lower values of the discount factor, the same effects repeat for single relationships (region (ii), (i)).

The above result highlights that if the same relationship structure is used by the two direct buyers, the mediated buyer will be able to better maintain that relationship. However, it is possible that the two direct buyers may prefer to maintain relationships with different sets of cooperative suppliers. In such cases, the intermediary will have to choose one of the two sets of cooperative suppliers or relationship structures, whereas the direct buyers can each choose their preferred relationship structure. Thus, direct sourcing may perform better, as direct buyers have more selectivity in choosing their relationships; in particular, they are not obliged to each have the same set of relationships, as is the case when an intermediary acts on their behalf. For example, in direct single sourcing, each buyer must choose supplier 1 or 2 as the cooperative supplier. This can be the same supplier for both buyers or a different supplier for each buyer. If this is the same supplier for both, the above theorem applies and mediated sourcing outperforms direct sourcing. If the preferred supplier, is different for the two buyers, then with direct sourcing both buyers can choose their desired partner. But, the intermediary, being constrained to choosing one supplier for the two buyers, might find itself in a disadvantaged position. Thus, independent decisions on the type of relationship (k) of the two buyers in direct sourcing effectively gives the buyers more selectivity in choosing the preferred supplier, whereas the intermediary being limited to choosing one type of relationship for the two buyers has lower selectivity. The next Theorem formalizes this effect of selectivity.

**Theorem 3.** If there exist δ, where the two direct buyers do not desire the same relationship type, direct sourcing can outperform mediated sourcing for a subset of these δ. Specifically, if for some δ, all of the following conditions hold for all \( t \geq 0 \), then \( \pi^d (\delta) > \pi^m (\delta) \) at this δ.

\[
E \left[ X'_t | \eta_b \geq X'_t \geq \eta_b \right] + \eta_b \cdot (1 - F'_t (\eta_b) - F'_t (\eta_b)) > 0, \quad E \left[ X'_t | \eta_b \geq X'_t \geq \eta_b \right] + \eta_b \cdot (1 - F'_t (\eta_b) - F'_t (\eta_b)) < 0; \\
\delta^{t+1} ((1 - F'_t (0)) (1 - \gamma) + F'_t (0) - F'_t (-\eta_b)), \quad \delta^{t+1} ((F'_t (0)) (1 - \gamma) - F'_t (0) + F'_t (\eta_b)) \geq 1 - \gamma, \text{ where } \gamma \equiv \frac{w_{u} (s^{u})}{s_{a} (s^{u})}; \\
\delta^{t+1} \left( E \left[ X'_t | \eta_b \geq X'_t \geq \eta_b \right] + \eta_b \cdot (1 - F'_t (\eta_b)) \right), \quad \delta^{t+1} \left( \eta_b \cdot F'_t (\eta_b) - E \left[ X'_t | \eta_b \geq X'_t \geq 0 \right] \right) \geq (1 - \delta) \bar{G}_{u}/d.
\]

**Proof.** The formal proof is provided in the Appendix (Page 32). □

The conditions in the above theorem ensure that the two direct buyers desire to enter into different relationships and that these different direct relationships create more value than mediated sourcing. Note that the above effect arises only as the mediated single relationship is constrained to be either cooperative or non-cooperative, but the intermediary can’t choose to source part of the order cooperatively and the remaining part non-cooperatively from the supplier it has a relationship with. If the intermediary could have such "partial cooperation" with one supplier, corresponding to
different behavior when sourcing for the two client buyers, this disadvantage of intermediation would not arise and an intermediary would always outperform direct sourcing, as illustrated in Theorems 1 and 2. Taken together, our analyses demonstrate that mediated sourcing is better at maintaining relationships, while direct sourcing is better at letting buyers choose which supplier to get into a relationship with. In particular, there are three key phenomena that differentiate direct and mediated sourcing; the ability to use long-run and myopic differences that favor mediated sourcing, and the better selectivity of direct sourcing.

These three phenomena are distinct from the transactional and informational advantages of intermediation. There are no information asymmetries or information aggregation effects in our setup. Further, the intermediary is not using the aggregated scale of buyer transactions to defray fixed transaction costs. The key drivers of our effects are incomplete contracting and the difference in buyer preferences over suppliers at any given point in time.

We conjecture that these effects may provide an explanation for the phenomenal recent growth in mediated sourcing. With an increasingly volatile business environment, we think that there is increasing uncertainty in buyer preferences over suppliers, which leads to more changes in buyer preferences and consequently higher differences in buyer preferences. We also believe that as firms are outsourcing increasingly critical inputs and more complex parts of their businesses, sourcing is characterized more and more by incompleteness of contracts, which increases the value of maintaining relationships which, as per our analysis, is a key advantage of intermediaries. Finally, our effects are agnostic to the scale of the sourcing company, and thus might also explain the adoption of mediated sourcing by larger companies than predicted by existing theory.

5. Extensions

In our model of mediated sourcing, we assume that buyers transfer all their profit-relevant actions to the intermediary, and thus have no control over buyer-side surplus or sourcing profits. Arguably, this assumption unfairly favors the mediated sourcing model—by assuming a perfect transfer of the actions from the buyers to the intermediary, we assume that the addition of the intermediary to the supply chain does not create any new incentive conflicts, or that the incentives of the buyer and the intermediary are perfectly aligned. However, it is possible for the buyer and intermediary to work at cross-purposes and this incentive conflict would destroy some of the value created by mediation.

To address this concern, we developed and rigorously analyzed an alternate model of mediated sourcing that explicitly models the buyer-intermediary transaction as another generic extensive form
In this model, in addition to allowing inefficiency in the supplier-intermediary transaction, we also allow for an additional inefficiency in the buyer-intermediary transaction. Specifically, we assume that the Nash behavior in the buyer-intermediary transaction decreases the sourcing surplus as compared to the model presented in the paper. Only when the buyer and intermediary behave cooperatively in their interaction is there no additional loss in efficiency; for all other actions, the buyer-intermediary transaction reduces the sourcing profits as compared to the original mediation model.

Our analysis indicates that all the effects mentioned in this paper that drive the advantages of mediation continue to hold with this extension. However, when buyer-intermediary incentives are misaligned, then, as expected, there is an increased potential for opportunism in mediated sourcing compared to direct sourcing that lowers some of the gains from mediation; however, surprisingly, we find that there is also a “policing effect” that actually increases the gains from mediation. Essentially, the increased potential for opportunism also serves as a bigger deterrent against opportunism for some players. Specifically, intermediary opportunism can be punished by actions from both the buyers and suppliers.

In our model, we allowed for buyer preferences over suppliers to change over time, but the suppliers are indifferent between buyers. Suppliers may actually also have preferred buyers and these preferences may change over time or in response to changing buyer preferences. We also developed an extension to the model presented in the paper that allows for both buyer and supplier preferences to change over time. Our analysis indicates that even in this setting the effects described in the paper continue to operate, and mediated sourcing continues to outperform direct sourcing in establishing relationships. Note that while we have labeled one party as the supplier and the other as buyer, our model is agnostic to actual product flows. Thus, the results presented here are all equally valid if the roles are reversed.

Our work opens up important new avenues of study. In this paper, we provide a rigorous analytical model to demonstrate a novel advantage of intermediaries. Our purpose in this paper is not to rigorously compare the significance of this advantage compared to the transactional and informational advantages of the intermediary. We might speculate on the basis of anecdotal evidence, though, that such effects do exist. Therefore, further work needs to be done to assess the extent of our effect; estimating its contribution would require an empirical setting with variation in the economic characteristics of the products sourced and variation in the sourcing strategy employed.

The detailed models, their analyses, and the formal results discussed in all extensions described in Section 5 are available from the authors upon request.
Finally, we must point out that while the context of this study is supply chain intermediation, we believe the effects mentioned here apply to any two-sided market with incomplete contracts (Rochet and Tirole (2006)).

REFERENCES


**Appendix A. Formalization of Section 3.1 (Model Preliminaries)**

**A.1. Notation for The Engagement Game** $\Gamma$. Let $\Xi$ be the collection of initial nodes of the subgames of game $\Gamma$, with $\xi^0$ being the initial node. The subgame of $\Gamma$ with initial node $\xi \in \Xi$ is denoted by $\Gamma_\xi$. It is partially ordered by precedence relation, where $\xi < \xi'$, if $\xi'$ is a node in $\Gamma_\xi$. A set of terminal nodes is denoted by $Y$, with typical element $y$. An action for player $s$ (b) specifies a move for player $s$ (b) at each information set owned by that player. At the end of the period, the players observe terminal node $y$ reached as a result of play. A unique terminal node is reached under a path of play implied by $a$. Given a node $\xi \in \Xi$, $u_s(b|\xi)$ is player $s$’s (b’s) payoff from $\Gamma_\xi$, given the moves in $\Gamma_\xi$ implied by $a$.

**Appendix B. Formalization and Proofs for Section 3.2 (Direct Sourcing)**

**B.1. Notation for The Direct Sourcing Stage Game at time** $t$, $G_{it}$. The collection of the initial nodes of the subgames of game $G_{it}$ is $\Xi^{G_{it}} \equiv \{\xi^{0_i}\} \cup \{\Xi^{T_{i1}} \times \Xi^{T_{i2}}\}$ where $\xi^{0_i}$ is the sourcing fraction selection node, and $\Xi^{T_{ij}}$ is the set of initial nodes of the subgames of game $\Gamma$ played between buyer $i$ and supplier $j$, i.e. we will add superscript $ij$ to all nodes, so the initial node of game $\Gamma$, $\xi^0$ will become $\xi^{0ij}$ and so the initial node of the transaction step will be $\xi^{01} \xi^{02}$. The set of terminal nodes is $Y^{G_{it}} = Y^{T_{i1}} \times Y^{T_{i2}}$.

**B.2. Notation for The Repeated Direct Sourcing Game** $G_{i\infty}$. The set of period $t$, $t \geq 0$, ex-ante histories is given by $H^t = (\mathcal{X}_i \times \mathcal{A})^t$, identifying the state $(X_i^t)$ and action profile ($\mathcal{A}$) in each previous period; $\mathcal{X}_i^t$ is the support of $F_i^t$, $\mathcal{A} \equiv \{\theta_i\} \times A \times A$. The set of period $t$, $t \geq 0$, ex-post histories is given by $\tilde{H}^t = (\mathcal{X}_i \times \mathcal{A})^t \times \mathcal{X}_i^{t+1}$, identifying the state and action profile in each previous period and identifying the current state. Let $H = \cup_{t=0}^{\infty} H^t$, $\tilde{H} = \cup_{t=0}^{\infty} \tilde{H}^t$, we
set $\mathcal{H}^0 = \{\emptyset\}$, hence $\mathcal{H}^0 = \mathcal{X}^0$. The pure strategy for player $n$ is a mapping $\sigma_n : \mathcal{H} \rightarrow \mathcal{A}_n$, associating an action with each ex-post history.

B.3. Additional Lemmas.

**Lemma 3. Direct Sourcing: The Stage Game Equilibrium**

$\alpha^*_d$ is a subgame-perfect Nash equilibrium of $G^{it}$.

**Proof.** By definition, an action profile $\alpha^*_d$ is a subgame-perfect equilibrium if for every node $\xi \in \Xi^{G^{it}}$, the profile $\alpha^*_d|_\xi$ is a Nash equilibrium of subgames $G^{it}_\xi$. Stage game $G^{it}$ starts with a choice of supply fractions– initial node $\xi^{0\iota}$, i.e. $G^{it} = G^{it}_{\xi^{0\iota}}$, and is followed by subgames of game $\Gamma$. We know that $a^N$ is a subgame-perfect equilibrium of the transaction step subgame. Hence, due to additive separability of players’ utilities, for every node $\xi \in \{\Xi^{r1} \times \Xi^{r2}\}$, $\alpha^*_d|_\xi = (a^N, a^N)|_\xi$, and so $\alpha^*_d|_{\xi^{0\iota}}$ is a subgame-perfect equilibrium of the transaction step subgame. Hence, we only need to show that $\alpha^*_d|_{\xi^{0\iota}}$ is a Nash equilibrium of $G^{it}$, i.e. $\theta^t_i (u_0(a^N) + X^t_i) + (1 - \theta^t_i) u_0(a^N) \geq \theta^t_i (u_0(a^N) + X^t_i) + (1 - \theta^t_i) u_0(a^N)$. The prescribed choice, $\theta^t_i = \tilde{\theta}^t_i \equiv \{X^t_i \geq 0\}$, satisfies this inequality. \hfill $\square$

B.4. **Proof of Lemma 1 (Section 3.2, Page 11).** Part 1. $\alpha^*_d$, with $\theta^t_i = \tilde{\theta}^t_i$, is a subgame-perfect equilibrium of $G^{it}$ (Lemma 3) and hence is a subgame-perfect equilibrium of $G^{\infty}$. No other $\theta^t_i$ can be maintained as, in any period, the buyer could deviate to $\tilde{\theta}^t_i$ and improve his profit.

Part 2. In order to establish this, we need to show that for all histories $\tilde{h}^t, \xi$, $t \geq 0$, $(1 - \delta) u_0^t (a^k|\xi) + \varepsilon^{t+1} (\sigma^k|_{\tilde{h}^t,y} (\alpha^k|\xi)) \geq (1 - \delta) u_0^t (\alpha'_n, \alpha^k_n|\xi) + \varepsilon^{t+1} (\sigma^k|_{\tilde{h}^t,y} (\alpha'_n, \alpha^k_n|\xi))$ for all $\alpha'_n$ and all players $n$, where $\alpha^k$ is an action profile prescribed by $\sigma^k$ following the ex-post history $\tilde{h}^t$.

We start with histories $\tilde{h}^t$ that include a deviation. Following such a history, the strategy $\sigma^k$ is prescribing the stage game equilibrium to be played forever after, hence $\varepsilon^{t+1} (\sigma^k|_{\tilde{h}^t,y} (\alpha^k|\xi)) = \varepsilon^{t+1} (\sigma^k|_{\tilde{h}^t,y} (\alpha'_n, \alpha^k_n|\xi))$. Further, from Lemma 3 we know $\forall \xi: u_0^t (\alpha^*_d|\xi) \geq u_0^t (\alpha'_n, \alpha^*_n|\xi)$.

Taken together, these ensure that condition (L1) is satisfied.

Further, for histories $\tilde{h}^t$ that do not include a deviation, the strategy is prescribing $\alpha^k$ to be played if no deviations are observed, and the stage game equilibrium following any deviation. Next, we divide all initial nodes of subgames, $\xi \in \Xi^{G^{it}}$, into two classes: the ones on and the ones off the equilibrium path.

For all $\xi$ that are off the equilibrium path, $\alpha^k|\xi = \alpha^*_d|\xi$ and, hence, $\varepsilon^{t+1} (\sigma^k|_{\tilde{h}^t,y} (\alpha^k|\xi)) = \varepsilon^{t+1} (\sigma^k|_{\tilde{h}^t,y} (\alpha'_n, \alpha^k_n|\xi))$, as the strategy prescribes the stage game equilibrium to be played

$$13 I(W \geq w) = \begin{cases} 1, & \text{if } W \geq w; \\ 0, & \text{if } W < w. \end{cases}$$
forever after. Then we only need to show that $u_n^t(\alpha_{d_i}^s|\xi) \geq u_n^t(\alpha_{d_i}^t, \alpha_{d_i}^m|\xi)$, which is established in Lemma 3. Taken together, these ensure that condition (L1) is satisfied.

Next, consider all $\xi$ that belong to the equilibrium path. For the non-cooperating supplier, denoted by $s_N$, (in strategy $d_i s_1$ it is supplier 2, in $d_i s_2$ it is 1) $\delta_{t+1}
left(u_{s_N}^t\left(\sigma^k(\tau_{i+j}(\alpha^k|\xi))\right)\right) = \delta_{t+1}
left(u_{s_N}^t\left(\sigma^k(\tau_{i+j}(\alpha_{s_{N}}^s, \alpha_{s_{N}}^m|\xi))\right)\right) - even if she deviates, this would not influence the continuation of cooperation among cooperating players. Thus we only need to ensure $1 - \delta u_{s_N}^t(\alpha^k|\xi)$, which holds as $\alpha_{s_{N}}^k|\xi$, prescribes $a^N$ to be played with the non-cooperative supplier which a subgame-perfect equilibrium of $\Gamma$. For the cooperative supplier(s), supplier $j$ has deviations only inside $\Gamma_{ij}$, we need to show: for all $a_{s_i}^t$, $\xi$, $\delta_{t+1}
left(u_{s_j}^t\left(\sigma^j\right)\right) - \delta_{t+1}a_{s_i}^t\left(\sigma^j\right)\left(\sigma^j(a_i|\xi)\right) \geq (1 - \delta) \frac{\vartheta_j}{\vartheta_j} \left(u_s(a_{s_i}^t, a_{s_i}^m|\xi) - u_s(a^C|\xi)\right)$, where $\vartheta_1 = \theta_i^j$ and $\vartheta_2 = 1 - \theta_i^j$. Hence, $\delta_{t+1}
left(u_{s_j}^t\left(\sigma^j\right)\right) - \delta_{t+1}a_{s_i}^t\left(\sigma^j\right)\left(\sigma^j(a_i|\xi)\right) \geq (1 - \delta) \frac{\vartheta_j}{\vartheta_j} \left(u_s(a_{s_i}^t, a_{s_i}^m|\xi) - u_s(a^C|\xi)\right)$ and $\vartheta_j = \max_{X_i^j} \vartheta_j$, will ensure that the above holds for all $a_{s_i}^t$, $\xi$.

In each period $t$, the buyer can deviate at the initial node $\xi_{ij}^t$, which is immediately detectable: $\forall X_i^t: \delta\left(\delta_{t+1}
left(u_{b_i}^t(\sigma^k)\right) - \delta_{t+1}a_{s_i}^t\left(\sigma^j\right)\left(\sigma^j(a_i|\xi)\right)\right) \geq (1 - \delta) \left(\theta_i^j \left(u_b(a^N) + X_i^t\right) + (1 - \theta_i^j) u_b(a^N)\right) - (1 - \delta) \left(\theta_i^j \left(u_b(a_{s_i}^t(\sigma^k)) + X_i^t\right) + (1 - \theta_i^j) u_b(a_{s_i}^t(\sigma^k))\right)$. The best deviation is $\theta_i^j = \tilde{\theta_i}^j$, which is reflected in the statement of the Lemma. If there are no deviations in the selection step, $\delta\left(\delta_{t+1}
left(u_{b_i}^t(\sigma^k)\right) - \delta_{t+1}a_{s_i}^t\left(\sigma^j\right)\left(\sigma^j(a_i|\xi)\right)\right) \geq (1 - \delta) \vartheta_{b_i} \left(u_b(a_{s_i}^t, a_{s_i}^m|\xi) - u_b(a^C|\xi)\right)$, for all $a_{s_i}^t$, $\xi$, where by $\vartheta_{b_i}$, we denote the amount of cooperation the buyer has; in $d_i d$ it is $\theta_i^j + 1 - \theta_i^j = 1$, in $d_i s_1$ it is $\theta_i^j$ and in $d_i s_2$ it is $1 - \theta_i^j$. As $G_b = max_{a_{s_i}^t, a_{s_i}^m\xi} \left(u_b(a_{s_i}^t, a_{s_i}^m|\xi) - u_b(a^C|\xi)\right)$, it boils down to $\delta\left(\delta_{t+1}
left(u_{b_i}^t(\sigma^k)\right) - \delta_{t+1}a_{s_i}^t\left(\sigma^j\right)\left(\sigma^j(a_i|\xi)\right)\right) \geq (1 - \delta) \vartheta_{b_i} G_b$. This establishes all inequalities of Lemma1.

**Appendix C. Formalization and Proofs for Section 3.3 (Mediated Sourcing)**

**C.1. Notation for The Mediated Sourcing, The Stage Game, $G^{It}$**. Denote the collection of the initial nodes of the subgames of $G^{It}$ as $\Xi^{G^{It}} \equiv \{\xi^{u'}\} \cup \{\Xi^{\Gamma^{I_1}} \times \Xi^{\Gamma^{I_2}}\}$ where $\xi^{u'}$ is the sourcing fraction selection node , and $\Xi^{\Gamma^{I_1}}$ is the set of the initial nodes of the subgames of games $\Gamma$ played between the intermediary and supplier $j$, i.e. we add superscript $I^j$ to all nodes. The set of terminal nodes is $Y^{G^{It}} = Y^{\Gamma^{I_1}} \times Y^{\Gamma^{I_2}}$. Each $\Gamma^{I_1}$ is a merge of two transaction step games, as the intermediary now needs to source for two buyers from two possible suppliers. In $\Gamma^{I_1}$ compared to the supplier that took the actions in game $\Gamma$, both suppliers $j$ can now take the very same actions for each buyers’ order, and whenever the buyers were supposed to act, the intermediary is now taking two such actions.
C.2. Notation for The Repeated Mediated Sourcing Game $G^{1\infty}$. The set of period $t \geq 0$ ex-ante histories is given by $\mathcal{H}^t = (\mathcal{X} \times \mathcal{A})^t$, identifying the state $\left( X_1^t, X_2^t \right)$ and the action profile $(\mathcal{A})$ in each previous period, $\mathcal{X}^t$ is the support of $F^t$, $\mathcal{A} \equiv \{ \nu \} \times A \times A \times A \times A$. The set of period $t \geq 0$ ex-post histories is given by $\mathcal{H}^t = (\mathcal{X} \times \mathcal{A})^t \times \mathcal{X}$, identifying the state and the action profile in each previous period and identifying the current state. Let $\mathcal{H} = \bigcup_{t=0}^{\infty} \mathcal{H}^t$, $\tilde{\mathcal{H}} = \bigcup_{t=0}^{\infty} \tilde{\mathcal{H}}^t$, we set $\mathcal{H}^0 = \{ \emptyset \}$, hence $\tilde{\mathcal{H}}^0 = \mathcal{X}^0$. The pure strategy for player $n$ is a mapping $\sigma_n : \mathcal{H} \rightarrow \mathcal{A}_n$, associating an action with each ex-post history.

C.3. Additional Lemmas.

**Lemma 4. Mediated Sourcing: The Stage Game Equilibrium**

$\alpha^*_m$ is a subgame-perfect Nash equilibrium of $G^{1t}$.

*Proof.* From the additive separability of the utilities of the intermediary and the suppliers with respect to actions $a_{1j}$ and $a_{2j}$ in the merged games, and that $a^N$ is a subgame-perfect equilibrium of $\Gamma$, it follows that $\alpha^*_m|_{\kappa^0}$ is a subgame-perfect equilibrium of $G^{1t}_{\kappa^0}$. Thus, we only need to show optimality in the supplier selection choice, or $\sum_{i=1}^{2} \left( \nu_i (u_b(a^N) + X_i^t) + (1 - \nu_i) u_b(a^N) \right) \geq \sum_{i=1}^{2} \left( \nu_i (u_b(a^N) + X_i^t) + (1 - \nu_i) u_b(a^N) \right)$. The prescribed choice satisfies this inequality. So, $\alpha^*_m$ is a subgame-perfect equilibrium of $G^{1t}$. □

C.4. Proof of Lemma 2 (Section 3.3, Page 15). The proof follows along the same lines as the proof of Lemma 1, noting that suppliers have $\vartheta_1 = \nu_1 + \nu_2$ and $\vartheta_2 = 2 - (\nu_1 + \nu_2)$ orders on hand on which they can deviate. In $md$, the intermediary is sourcing $\vartheta_1 + \vartheta_2 = 2$ orders cooperatively, in $ms_1 - \vartheta_1$, in $ms_2 - \vartheta_2$, the maximal possible deviations in transaction steps follow.

**APPENDIX D. PROOFS FOR SECTION 4 (THE BENEFITS OF INTERMEDIATION)**

D.1. Proof of Theorem 1 (Section 4, Page 16). 1. For given strategy $k$, as $a^k_{1j} = a^k_{2j} = a^k_{i,j} \equiv a^k_j$, $j \in \{1, 2\}$, the sourcing profits for direct and mediated structures are given by:

$$
\pi_{d1k}(\delta) + \pi_{d2k}(\delta) = \max_{\varrho_1, \varrho_2} \mathcal{E}_0 \left( \left( \varrho_1^0 + \varrho_2^0 \right) u_b \left( a^k_1 \right) + \varrho_1 X_1^t + \varrho_2 X_2^t + (2 - (\varrho_1 + \varrho_2)) u_b \left( a^k_2 \right) \right),
$$

$$
\pi_{mk}(\delta) = \max_{\nu_1, \nu_2} \mathcal{E}_0 \left( \left( \nu_1^0 + \nu_2^0 \right) u_b \left( a^k_1 \right) + \nu_1 X_1^t + \nu_2 X_2^t + (2 - (\nu_1 + \nu_2)) u_b \left( a^k_2 \right) \right).
$$

2. Further, we need to ensure that all cooperative players have sufficient incentives to maintain this strategy. In order for supplier $s_i$ to cooperate in respective games with buyer $i$ or the intermediary, as per Lemmas 1, 2, the following constraints should be satisfied for all $t$ ($\gamma = \frac{u_s(a^N)}{u_s(a^C)}$, $d = \frac{1 - \delta}{\delta} \frac{G_s}{u_s(a^C)}$):
buyer-side constraints in direct sourcing can be expressed as suppliers. Hence, the supplier-side constraints, as defined in part 2, are satisfied. Further, the described in part 2, are just the sum of the constraints that direct buyers have to satisfy for this supplier. Hence, if for given

\[ \pi^t \]

\[ \nu_1 + \nu_2 \]

Further, we can show that the intermediary might be able to further improve profits. From

\[ \delta^t \]

\[ \theta_{1t}^* \]

\[ \theta_{2t}^* \]

From \( \delta^t \) onwards, and \( G_{b_1}^{kt} \) is the highest gain buyer \( i \) can gain in period \( t \) from deviation given that strategy \( k \) is played. Having \( \nu_1 + \nu_2 = \theta_{1t}^* + \theta_{2t}^* \) as described in part 2, the constraint of the intermediary can be written as \( \pi^{mk} (\delta | t + 1) - \pi^{mt} (\delta | t + 1) \geq \frac{1-\delta}{\delta} (G_{b_1}^{kt} + G_{b_2}^{kt}) \). In part 3 we established that \( \pi^{mk} (\delta | t + 1) \geq \pi^{dlt} (\delta | t + 1) + \pi^{dsk} (\delta | t + 1) \), further in transactional sourcing intermediary and the buyers make the same sourcing profit, \( \pi^{mt} (\delta | t + 1) = \pi^{dlt} (\delta | t + 1) + \pi^{dsk} (\delta | t + 1) \). Hence, the constraint of the intermediary is also satisfied.

Further, we can show that the intermediary might be able to further improve profits. From expressions in part 4, the supplier-side constraints are just the sum of the constraints that direct buyers have to satisfy for this supplier. Hence, if for given \( \delta \), \( \theta_{1t}^* \) is chosen so that the constraint of supplier 1 is binding, but \( \theta_{2t}^* \) satisfies the respective constraint with slack, then the suppliers-side constraint in mediated sourcing is satisfied with slack. Hence, the intermediary can further improve the sourcing profit by choosing \( \nu_1^*, \nu_2^* \) so as to remove the undesired slack and improve the profit.

D.2. Proof of Corollary to Theorem 1 (Section 4, Page 19). All statements of the Corollary follow from the gain of mediation established in Part 3 of the proof of Theorem 1.
D.3. **Proof of Theorem 3 (Section 4, Page 22).** In our setup we assume that whenever the intermediary enters into a relationship with the supplier it needs to always source cooperatively from this supplier (independent of the buyer it is sourcing for). This Theorem allows us to demonstrate how restrictive this assumption can be. We will do so by constructing a specific set of conditions which lead to worse performance of mediated sourcing.

1. Suppose that in single sourcing with responsive allocation of orders between cooperative and non-cooperative suppliers, we are going to refer to these profits as \( \bar{\pi}_{d_i,s_j} \), buyers sourcing directly prefer to establish a relationship with different suppliers, i.e. for some \( i \in \{1, 2\} \) \( \bar{\pi}_{d_i,s_1} > \bar{\pi}_{d_i,s_2} \) and \( \bar{\pi}_{d_i,s_2} > \bar{\pi}_{d_i,s_1} \). Writing out these conditions in terms of the parameters of the model, we get the first set of constraints stated in the theorem.

2. We need to insure that at least for some \( \delta \) buyers will use the single sourcing with responsive allocation, i.e. for some \( \delta \), \( \pi_{d_i}(\delta) = \bar{\pi}_{d_i,s_1} \) and \( \pi_{d_i}(\delta) = \bar{\pi}_{d_i,s_2} \). The first set of conditions is ensuring that single responsive sourcing with the preferred supplier can be maintained for a wider range of \( \delta \) than any of the dual sourcing strategies. Having supplier-side constraints for maintaining dual sourcing both in direct and mediated sourcing, one can derive that for delta lower than \( \delta^d \), the solution to \( \frac{1-\delta}{\delta} \frac{G_s}{u_s(a^c)} = 1 - \gamma \) no dual sourcing strategies can be maintained both in direct and mediated sourcing. Similarly, we can find the lowest delta where the preferred single responsive strategy for each of the buyers can be sustained, \( \delta_{d_i,s_1} \) and \( \delta_{d_i,s_2} \). The second set of conditions of the theorem simply requires that \( \delta_{d_i,s_1}, \delta_{d_i,s_2} \leq \delta^d \). The last set of constraints insures that for given \( \delta \), buyer-side constraints are satisfied.

3. Lastly, for all \( \delta_{d_i,s_1}, \delta_{d_i,s_2} \leq \delta \leq \delta^d \) such that the conditions stated in the theorem are satisfied direct sourcing will perform better than mediated, as the best profit mediated sourcing can achieve for delta in this range is: \( \bar{\pi}^m = \max \left\{ \bar{\pi}_{d_i,s_1} + \bar{\pi}_{d_i,s_1}, \bar{\pi}_{d_i,s_2} + \bar{\pi}_{d_i,s_2} \right\} < \bar{\pi}_{d_i,s_1} + \bar{\pi}_{d_i,s_2} \).
Collaborative Cost Reduction and Component Procurement
Under Information Asymmetry

Sang-Hyun Kim (presenter)
Yale School of Management, Yale University, New Haven, CT 06511
sang.kim@yale.edu

Serguei Netessine
INSEAD, Boulevard de Constance, 77305 Fontainebleau, France
serguei.netessine@insead.edu
Collaborative Cost Reduction and Component Procurement
Under Information Asymmetry

Abstract
During development of a highly innovative product there is often considerable uncertainty about component production cost, and it is of interest for both the manufacturer and the supplier to engage in a collaborative effort to reduce this uncertainty and lower the expected cost. Despite the obvious benefits this brings, the supplier may be reluctant to collaborate as he fears revealing his proprietary cost information. We investigate how information asymmetry and procurement contracting strategies interact to influence the supply chain parties' incentives to collaborate. We consider a number of procurement contracting strategies, and identify a simple strategy, Expected Margin Commitment (EMC), that effectively promotes collaboration. The manufacturer prefers EMC if collaboration leads to a large reduction in unit cost and/or demand variability is low. Otherwise, a screening contract is preferred. We also find that, paradoxically, ex-post efforts to enhance supply chain efficiency may hinder ex-ante collaboration that precedes production.

1 Introduction
Many manufacturing firms rely on the expertise and the resources provided by their suppliers when they develop new products or upgrade existing products, and how well they manage these relationships critically impacts the product's success. For instance, it has been documented that the subcontracting structure of Japanese automobile manufacturers—in which suppliers actively participate in every development and production process—was one of the key differentiators that enabled their competitive advantage over U.S. manufacturers (McMillan 1990). With increasing sophistication and complexity of the products that accompany technology breakthroughs, the manufacturers and the suppliers need to collaborate more than ever in order to survive in the marketplace.

Examples of supply chain collaboration may be found in every stage of the product life cycle, ranging from such activities as co-branding initiatives to long-term strategic alliances (Rudzki 2004). However, given that approximately 80% of a product's cost is determined during product development (Blanchard 1978), it is no surprise that major collaborative efforts are made in an early product development stage.
In particular, as a recent survey by Aberdeen Group (2006) reveals, firms identify cost reduction achieved during product development as one of the primary reasons for engaging in collaborative relationships. This view is supported by the following quote by an operations director from Copeland Corporation, an Ohio-based manufacturer of AC/heating equipment (Kinni 1996, p. 105): “The only way we could reach the current state of manufacturing efficiency is through sharing and understanding both companies’ processes... [Copeland’s component supplier] Osco is a member of the New Product Team and is intimately involved in all aspects of the casting design and machining process. The best way to achieve the lowest-cost raw material and finished component is to leverage the design process by utilizing the supplier’s expertise and achieving the lowest true cost for the component.”

While firms strive to attain the highest level of efficiency through collaboration, it can be an elusive goal. Forming a successful collaborative relationship rests on two key factors: the inter-firm information structure and product characteristics. The former is crucial since it influences the firms’ willingness to establish and sustain the relationship. Benefits of collaboration notwithstanding, the reality that each firm’s ultimate goal is maximizing its own profitability. Firms are inherently opportunistic, implying that they are averse to sharing proprietary information (such as the cost structure) and will take advantage of the other’s if they are presented with it. As an executive from an auto parts supplier put it, “if one doesn’t say anything, all the savings are ours” (Anderson and Jap 2005). From these reports and other evidence, it is clear that information asymmetry exists even in collaborative relationships, and in fact, it plays a crucial role in shaping the firms’ incentives to collaborate.

In addition, product characteristics such as the strategic importance of a procured component, modularity of component architecture, and uncertainty in production cost, quality, and delivery lead time also play big roles in determining successful outcome of collaboration (Pyke and Johnson 2003). In this paper we focus on the impact of uncertainty, both of demand and of the cost of producing a strategically important component. Unpredictable consumer demand is an especially important concern for the firms manufacturing innovative products with short life cycles, such as smartphones, whose fast pace of feature evolutions and shifting consumer tastes create a high level of inventory risks due to forecasting limitations and high rates of obsolescence. Uncertainty in the component cost arises as the supplier faces a multitude of production options early in the product development phase. For instance, the supplier may consider adopting untested technology in order to fulfill the manufacturer’s requirement for the end product’s functionality. In fact, reduction of uncertainty that accompanies new technology adoption is cited as one of the main reasons that firms collaborate (Handfield et al. 1999, Ragatz et al. 2002). Therefore, while innovations in product design and in production processes are necessary, they present a cost risk to the supplier and ultimately to the manufacturer, who bears a portion of the same risk when financial transactions are made.
The factors we have mentioned—uncertainties in demand and cost, information asymmetry, and incentives to collaborate—are all intricately related. As the supplier’s initial uncertainty about the component cost stems in large part from imprecise product design requirements,\(^1\) it is likely to be reduced by forging a close working relationship with the manufacturer during product development, which would help understand each other’s expectations and limitations better. As the product specification ambiguities clear up, the supplier is able to select technologies suitable for production. Additionally, a better cost predictability is typically accompanied by reduction of the expected unit cost, which then lowers the component procurement cost for the manufacturer. Such a benefit is supported by many studies including Handfield et al. (1999), who report in their survey of 71 manufacturers that collaboration led to reduction of the material purchase costs by 2.6%-50%.

However, despite the obvious benefits that collaborative cost reduction brings, it may not always be a good proposition for the supplier. As collaboration typically involves mutual information exchanges, the supplier unavoidably reveals some of his cost structure to the manufacturer (see Womack et al. 1991, p. 149). For example, the supplier may have to inform the manufacturer that he will use a particular material to build a component, but the price of the material may be known publicly. Hence, the supplier faces a dilemma: despite the benefits, is it worth participating in the collaborative effort and risk exposing a better estimate of his cost to the manufacturer? How does the choice of a procurement contract impact the supplier’s and the manufacturer’s incentives to collaborate? How do uncertainties in demand and cost impact collaboration decisions? These are the questions that we aim to answer.

In this paper we develop a stylized game-theoretic model that formalizes the process by which the manufacturer’s and the supplier’s voluntary contributions to collaborative efforts lead to cost reduction. Using this model, we find that both firms’ incentives to collaborate critically depend on the procurement contracting strategy that the manufacturer employs. In addition, we identify demand variability as one of the important environmental factors that influence a successful outcome of collaboration. Specifically, we obtain the following insights from our analysis.

- Although the adverse selection literature points to the screening contract as the most efficient mechanism to deal with information asymmetry, in the setting that we consider, it may not be the optimal procurement contract to use since it hinders the supplier’s ex-ante incentive to collaborate, thereby creating and aggravating a hold-up problem.

- Price commitment, which is frequently mentioned in the literature as an effective means to alleviate the negative consequences of the hold-up problem, does not promote collaboration in our setting.

\(^1\)During the product development stage, suppliers usually receive only rough estimates of design specification parameters from the OEMs (Nellore et al. 1999).
Instead, we show that committing to a fixed margin over the expected cost, which we call Expected Margin Commitment (EMC), is a better instrument in achieving the same goal.

- Demand variability is a key factor that determines which contracting strategy should be employed by the manufacturer. The manufacturer prefers EMC to the screening contract when (a) collaboration can potentially lead to a large reduction in the unit cost and/or (b) demand variability is low.

- Paradoxically, ex-post supply chain efficiency improvement—achieved through more accurate demand forecasting or lead time reduction—is in conflict with ex-ante collaboration; when efficiency improvement is so large that the supply chain turns into a make-to-order production system, neither party exerts collaborative efforts.

The rest of the paper is organized as follows. After a literature review, we lay out the assumptions of the model and introduce notations used throughout the paper. Next, we present the analysis of the benchmark integrated supply chain case. In the subsequent sections, we study how collaboration incentives are impacted by the procurement contracting strategies and identify the conditions under which one strategy is preferred to others.

2 Literature Review

The topic of supply chain collaboration has received much attention in the business press, but surprisingly few works exist in the academic literature, especially in the operations management (OM) area. Notable exceptions are the papers that investigate the benefits of Collaborative Planning, Forecasting and Replenishment (CPFR), including Aviv (2001, 2007). CPFR is mainly concerned with promoting information sharing and joint process improvement during production and fulfillment stages. While there are overlaps, our paper differs from the CPFR literature in that we study collaboration that occurs during the product development stage that precedes production. As indicated in many reports (e.g., Aberdeen Group 2006), such an early-stage collaboration is commonplace in many industries. In this paper we specifically focus on collaborative cost reduction, motivated by widespread practice of such initiatives (for example, Stallkamp 2005 details the supplier cost reduction program called SCORE at Chrysler). Bernstein and Kök (2009) investigate cost reduction in an assembly network and their model shares some similarities to ours, but they do not consider collaboration or information asymmetry. Roels et al. (2010) is one of the few OM papers that study the topic of collaboration, but their model is developed in the context of service provisioning, whereas ours applies to product development in a manufacturing environment.

As we analyze the dynamics that occur during product development, this paper is related to the
new product development (NPD) literature. For surveys of the literature, see Krishnan and Ulrich (2001) and Krishnan and Loch (2005). In this literature, however, the topic of inter-firm collaboration has not received much attention. The only exception, to the best of our knowledge, is Bhaskaran and Krishnan (2009), whose broad theme is similar to ours but they investigate a set of research questions quite different from ours. While they acknowledge that agency issues caused by opportunistic behaviors of the collaboration partners is a real challenge, they sidestep this discussion. In contrast, information asymmetry plays a central role in our paper. Moreover, one of the unique features of this paper is the interaction between collaborative cost reduction decisions and procurement contracting, the topic that is unaddressed in Bhaskaran and Krishnan (2009).

Supply chain contracting in the presence of information asymmetry, especially that of adverse selection, has become an established area of research in OM in recent years. Articles such as Ha (2001) and Corbett et al. (2004) are some of the representative works in this stream of research. Among them, Iyer et al. (2005) is quite related to this paper since they also consider the use of a screening contract in the context of product development. However, the features and the focuses of the two papers do not overlap much; whereas they study the issue of resource sharing under the complementary/substitutability assumptions in a static setting, we study how collaboration incentives are impacted by various types of procurement contracts (not just a screening contract) in a dynamic setting. More recently, several authors have investigated dynamic adverse selection problems arising in strategic sourcing, such as Li and Debo (2009) and Taylor and Plambeck (2007). Although there are some similarities, these models differ from ours in many dimensions, including motivations, modeling approaches, and managerial insights.

Our model can also be viewed as a variant of the models that combine adverse selection and moral hazard, since, in our model, the supplier exerts a discretionary, non-verifiable effort and subsequently possesses private information about his cost. There are a number of papers in the procurement contracting literature with a similar focus, including Laffont and Tirole (1986) and Baron and Besanko (1987). However, our model does not fit exactly into the traditional framework and therefore differs from these works because, in ours, the manufacturer is not represented as a “principal” in a strict sense. Instead, even though it is the manufacturer who devises the contract terms and offer them to the supplier, their relationship is more equal in the beginning when they engage in a simultaneous-move game in which they both decide how much efforts should be expended. In this respect, our model shares some similarities to the models that consider the principal-agent problems in teams (McAfee and McMillan 1991, Olsen 1993) and those that consider double moral hazard (Bhattacharyya and Lafontaine 1995, Baiman et al. 2000). However, many unique features of our model—including the joint decisions in the presence of adverse selection and double moral hazard, operational considerations such as inventory risk, and the dynamics created by the interaction between demand and cost uncertainties—distinguish our model from
the existing works.

One of the key elements of our model is the contract offer timing decision, which naturally brings up the hold-up problem (Klein et al. 1978) and the issue of evaluating operational flexibility vs. the value of commitment. In the OM literature, Taylor (2006) examines this issue in a setting where a manufacturer may offer a contract either before or after demand is realized to a retailer who possesses private information about demand. Despite a similar theme, the results in Taylor (2006) and in this paper are driven by different dynamics; for example, in this paper, one of the important determinants of whether to commit to a contract term is the interaction between demand variability and cost uncertainty. The work that comes closest to ours in addressing the timing issue is Gilbert and Cvsa (2002). Our paper differs from theirs in many respects, however, especially in our focus on information asymmetry, the role of uncertainty originating not only from demand but also from cost, and the decisions driven by inventory risks, as captured by the newsvendor framework.

3 Model Assumptions

3.1 Basic Assumptions

We focus on the two stages that precede sales: product development occurs in Stage 1, and production occurs in Stage 2. A manufacturer (“she”) designs and builds a highly innovative product, which has a short life span due to fast technological obsolescence but requires a long production lead time. As a result, inventory risk is a significant concern for the manufacturer, and she decides the product quantity in advance using the newsvendor logic. Because the manufacturer lacks in-house expertise to develop a key component, she outsources the task to a supplier (“he”), who possesses the necessary capability. We assume that each end product requires one unit of this component.

In Stage 1 the manufacturer and the supplier engage in collaborative component development. At the beginning, the supplier does not have sufficient knowledge on how to manufacture the component most efficiently, since it has to be custom-made for the end product that features novel functionalities. For example, the supplier would have to choose from a multitude of options on raw materials and parts, second-tier supplier selections, and competing ideas for the component architecture. Consequently, the unit cost $c$ of producing a component is highly uncertain at the start of Stage 1. This uncertainty can be reduced by collaborating with the manufacturer, who guides the supplier to build the component that satisfies her functional requirements. However, the manufacturer can only provide a rough guidance since her requirements are incomplete; not fully understanding the fine details of component manufacturing (e.g., does the right technology exist that enables the desired functionality?), the manufacturer starts

\footnote{Lee and Whang (2002) motivates their newsvendor-based model using the examples from similar product categories.}
product design by leaving many questions open, hoping that they will be resolved as the development process unfolds. Therefore, the two parties learn of each other’s expectations and limitations through collaboration, which typically involves multiple iterations of trial and error. (For simplicity, however, we do not explicitly model such a dynamic learning process; see Section 3.2.) In addition to uncertainty reduction, a higher level of collaboration lowers the expected unit production cost. Hence, collaboration lowers both the expected unit cost and uncertainty around it, presenting potential benefits to the manufacturer and the supplier. In our model, we identify these changes in the unit cost as the main outcome of collaboration and specifically focus on them.

Stage 2 begins after component development is complete. Significant uncertainty about the unit cost still remains, but at the start of this stage, the supplier privately learns unit cost realization, which is not relayed to the manufacturer. At this point the manufacturer may offer a procurement contract to the supplier (more details on this later). Production and assembly start afterwards. We normalize the cost of acquiring other parts and assembling the end product to zero, since they do not play significant roles in our analysis. Since the production lead time is long, the manufacturer has to order production quantity \( q \) in advance, when demand uncertainty exists. (In Section 8.1 we relax this make-to-stock production assumption and consider the make-to-order system.) We assume that the end product is sold at the end of Stage 2 at a predetermined price \( r \). Not only is the fixed price assumption in line with most other papers in the OM literature, it is consistent with practice. In the auto industry, for example, manufacturers typically set a target retail price first and then, working with the suppliers, figure out the ways to lower the cost below this target and be profitable (Womack et al. 1991, p. 148). For completeness, we relax the fixed price assumption in Section 8.4 and examine the consequences.

The manufacturer uses \( r \) as the basis of generating a forecast of the end product demand \( D \), which is a random variable with the mean \( \mu \), pdf \( f \), and cdf \( F \). This distribution is common knowledge. We assume that \( F \) is defined on a nonnegative support with \( F(0) = 0 \), and that it exhibits an increasing generalized failure rate (IGFR) property, which is satisfied by many well-known distributions. The notations \( \overline{F}(\cdot) \equiv 1 - F(\cdot) \) and \( J(y) \equiv \int_0^y x f(x) dx \), which represents the incomplete mean of \( D \), are used throughout the paper. For simplicity, we assume that unsold units are discarded after the end product becomes obsolete, i.e., we do not consider the secondary market. Introducing a salvage value for the product does not change the insights.

### 3.2 Collaboration Level and Unit Production Cost

To quantify the outcome of collaboration, we introduce the parameter \( \theta \in [0, 1] \) that measures the extent to which the unit cost is reduced through collaboration. We refer to \( \theta \) simply as the “collaboration level”. At \( \theta = 0 \) the firms are completely disengaged (“arm’s length relationship”), whereas \( \theta = 1 \) corresponds
to the maximum level of collaboration that can be achieved. The collaboration level $\theta$ results from joint efforts made by the manufacturer and the supplier, respectively denoted as $e_m$ and $e_s$. These efforts reflect the amount of investment, time, and resources that each firm puts in the collaborative process.

To capture the idea that collaboration creates positive synergy between the two firms, we assume that $e_m$ and $e_s$ are complementary with respect to $\theta$. To succinctly represent this relationship, we employ the Cobb-Douglas function with constant returns to scale: $\theta = e_m^\alpha e_s^{1-\alpha}$, where $0 < \alpha < 1.\textsuperscript{3}$ The exponents $\alpha$ and $1-\alpha$ are the elasticities of $\theta$ with respect to $e_m$ and $e_s$. Complementarity is ensured since $\frac{\partial^2 \theta}{\partial e_m \partial e_s} > 0$.

By construction, positive collaboration level ($\theta > 0$) is obtained if and only if both parties exert nonzero efforts. We assume a deterministic relationship between the efforts and $\theta$ to enable tractable analysis; a similar assumption is found in Roels et al. (2010) and other papers that consider double moral hazard. As we discuss in Section 8.3, adding a stochastic component to $\theta$ does not alter the insights.

Exerting an effort is costly to both the manufacturer and the supplier, and for simplicity, we assume that the disutility of effort is linear: $k_m e_m$ and $k_s e_s$. The disutility may include, among others, expenses incurred for communication, personnel exchanges, prototype testing, etc. We use the shorthand notation $K \equiv (\frac{k_m}{\alpha})^\alpha (\frac{k_s}{1-\alpha})^{1-\alpha}$ as this expression frequently appears in our analysis. It represents the composite cost-contribution ratio of exerting efforts. All functional forms introduced thus far are assumed to be common knowledge.

Reflecting our focus on unit cost reduction as the main outcome of collaboration, we define the relationship between $\theta$ and the unit cost $c$ as follows. The conditional unit cost $c|\theta$ is a random variable defined on a finite support with the conditional cdf $G(\cdot|\theta)$ and the pdf $g(\cdot|\theta)$. Since the mapping $G^{-1}(z|\theta)$ uniquely identifies the unit cost realization for a fixed $\theta$ at the $z\textsuperscript{th}$ quantile, $z \in [0,1]$, we present our model in the transformed $(\theta,z)$-space instead of the original $(\theta,c)$-space, as doing so simplifies analysis. Throughout the paper we refer to $G^{-1}(\cdot|\theta)$ as the unit cost function. As it turns out, analysis becomes intractable when we combine the general distribution functions $F$ and $G$. For this reason, we develop our model under the following simplifying assumptions on the unit cost distribution.\textsuperscript{4}

**Assumption 1**  
(i) $c|\theta$ is uniformly distributed with a constant lower support bound $c$ and an upper support bound $\overline{c}_\theta$, which varies with $\theta$.

\textsuperscript{3}Constant returns to scale implies that an $x\%$ increase in both $e_m$ and $e_s$ results in the same percentage increase in $\theta$. Although this is a somewhat strong assumption, we adopt it in order to simplify analysis. The same assumption is frequently found in the economics literature, especially since it offers intuitive interpretations (e.g., Varian 2003, p. 83). Note that Roels et al. (2010), like in our paper, employ the Cobb-Douglas function but they assume decreasing returns to scale, i.e., the function has a form $x^a y^b$ with $a + b < 1$. In our model, however, the objective functions in the optimization problems may not be unimodal if $a + b$ is sufficiently smaller than one, unnecessarily complicating the analysis. In our paper the distinction between constant vs. decreasing returns to scale is of small concern, since only the relative scale of $\theta$ matters and the main insights are not impacted by the exact shape of the functional form of $\theta$, as long as it exhibits complementarity between $e_m$ and $e_s$.

\textsuperscript{4}Together, (i)-(iii) in Assumption 1 imply the more general conditions $\frac{\partial}{\partial \theta} E[c|\theta] = \frac{\partial}{\partial \theta} \int_0^1 G^{-1}(z|\theta) dz < 0$ and $\frac{\partial}{\partial \theta} [G^{-1}(z_2|\theta) - G^{-1}(z_1|\theta)] < 0$. 

8
Figure 1: An example of a unit cost function satisfying Assumption 1.

(ii) $G^{-1}(z|\theta)$ decreases linearly in $\theta$ for all $z \in (0, 1]$.

(iii) $\delta \equiv \Delta_1/\Delta_0 < 1$, where $\Delta_0 \equiv \bar{c}_\theta - \underline{c}$.

Although these assumptions are somewhat restrictive, they offer the essential features of the unit cost function we wish to capture. See Figure 1 for an illustration. Uniform distribution assumption in (i) is frequently employed in the models in which two or more random variables interact, as in ours (see, for example, Li and Debo 2008, Cachon and Swinney 2009; non-uniform distributions do not qualitatively change the insights but considerably complicate the analysis). Under this assumption, the expected unit cost at a given $\theta$ is equal to $E[c|\theta] = G^{-1}(\frac{1}{2}\theta)$. Under the assumptions (ii) and (iii), both the mean unit cost and the spread $G^{-1}(z_2|\theta) - G^{-1}(z_1|\theta)$ for $0 \leq z_1 < z_2 \leq 1$, i.e., the gap between any two equiquantile curves, decrease in $\theta$. This formalizes the idea that a higher level of collaboration leads to lower expectation and lower uncertainty of the unit cost. The quantity $\delta$ in (iii) represents the fractional residual unit cost at $\theta = 1$. Equivalently, $1 - \delta$ is the percentage of cost reduction that can be attained at full collaboration.

Among the assumptions stated above, perhaps the two more restrictive ones are that $G^{-1}(\cdot|\theta)$ decreases linearly and that the lower support bound is fixed to a constant. The former is in fact inconsequential because only the relative values of $\theta$ are of our interest; replacing the linear functions with nonlinear monotonic curves only changes the scale. The latter assumption is employed mainly for the purpose of simplifying the analysis (without it, tractability is lost; we discuss the consequences of relaxing this assumption in Section 8.2). By having the lower bound fixed, we essentially assume that the supplier has a clear idea about the baseline unit cost $\underline{c}$ to build the component and that it is only the upside uncertainty that can be reduced through collaboration. Under these assumptions, we can express the unit cost function as

$$G^{-1}(z|\theta) = \underline{c} + \Delta_0 (1 - (1 - \delta)\theta) z.$$ (1)
Information asymmetry emerges at the end of Stage 1, after collaboration is completed and the efforts by both parties determine the value of $\theta$. At that point, the supplier privately learns the realized cost, or equivalently, the supplier’s “type” $z|\theta$. This information is not relayed to the manufacturer, as the supplier keeps it to himself with the intention of using it to his advantage. The manufacturer continues to have only limited knowledge about the supplier’s type, i.e., she knows the distribution of the unit cost at $\theta$ as specified in (1) but not the realized value.

Additionally, we make two technical assumptions on the range of parameter values in order to ensure the problem is well-behaved and to enable clean exposition by reducing the number of special cases that require separate discussions but of less import. First, we assume $c + 2\Delta_0 < r$, which leads to positive order quantities in all cases we consider. Second, we restrict our attention to the case

$$K < r \int_0^1 \left[ J \left( F^{-1} \left( 1 - \frac{c + \delta\Delta_0 z}{r} \right) \right) - J \left( F^{-1} \left( 1 - \frac{c + \Delta_0 z}{r} \right) \right) \right] dz. \quad (2)$$

This condition is satisfied when $\delta$ and the effort costs $k_m$ and $k_s$ are sufficiently small so that investing in cost reduction is attractive to them. Such a situation is conducive to the manufacturer and the supplier to engage in collaborative efforts since the unit cost can be significantly reduced with relatively small effort disutility.

### 3.3 Collaboration Effort Decisions and Contracting

As described above, the collaboration level $\theta$ is jointly determined in Stage 1 by the manufacturer’s and the supplier’s efforts $e_m$ and $e_s$. We assume that the effort levels are unverifiable and therefore non-contractible.\(^5\) This assumption is consistent with practice, in which complexity of a product development process prevents a third party (i.e., court) from enforcing a contract based on the effort levels that typically consist of many actions that are impractical to document and be presented as evidence. Moreover, neither party has a unilateral power to dictate the level of the other’s effort, as each has to rely on the other’s complementary expertise to develop the component. In contrast, as the designer and the producer of the end product who initiates the supply chain activities, the manufacturer has a greater influence over the procurement contract terms. Based on these observations, we model the game structure in the following stylized way. For the collaborative component development, we assume that the manufacturer and the supplier engage in a simultaneous-move game under which they decide their effort levels competitively, taking into account the costs and the mutual benefits they bring. For the component procurement, on the other hand, the manufacturer decides the terms of the trade and offers a take-it-or-leave-it contract to the

\(^5\)Bhattacharyya and Lafontaine (1995), Roels et al. (2010), and numerous others adopt the same assumption in similar double moral hazard settings.
supplier. Note that this leader-follower assumption does not give a complete leverage to the manufacturer since the supplier has an informational advantage, i.e., he keeps his realized unit cost private.

Motivated by the majority of procurement practices, we assume that the contract type is that of the price-quantity pair, \((w, q)\). That is, the manufacturer specifies the unit price and the quantity of the component in a contract, making sure that the supplier will agree to the proposed terms. Depending on when the manufacturer offers the contract, the contract may or may not consist of a single price-quantity pair. If she offers the contract once at the beginning of Stage 2 (immediately before production starts), at which point the collaboration level \(\theta\) is set and information asymmetry about the unit cost is in place, the optimal contract consists of a menu of price-quantity pairs \(\{(w(z|\theta), q(z|\theta))\}\), for \(0 \leq z \leq 1\) and \(0 \leq \theta \leq 1\). Each pair in this screening contract maps to the supplier’s realized type \(z|\theta\). As is well known in the mechanism design literature, the manufacturer can structure the menu so that the supplier chooses the pair specifically designed for him, thereby truthfully revealing his type. Then the manufacturer can extract all of the supplier’s surplus except for his information rent, which represents the inefficiency created by information asymmetry.

Although the practice of offering a procurement contract immediately before production starts is routinely observed, it is not the only option available to the manufacturer. In particular, she may decide to commit to a contract term in an early stage of the relationship, i.e., when they start to collaborate on component development. In Section 6 we investigate these commitment strategies in depth. Variants of these strategies are observed in practice. For instance, Japanese auto manufacturers and their suppliers agree on a payment amount based on the projected cost improvement that they expect to achieve through joint efforts (Womack et al. 1991). A similar practice was adopted by Chrysler as part of its pre-sourcing effort (Dyer 2000). Volume commitments are also frequently used as a way to improve supply chain relationships (Corbett et al. 1999).

The sequence of events is summarized as follows. At the start of Stage 1, the manufacturer decides which contracting strategy to adopt before commencing collaborative component development. She may either commit to a contract term value at the outset, or delay offering a contract until after collaboration is completed. Once the strategy is set, the manufacturer and the supplier simultaneously exert their collaborative cost reduction efforts \(e_m\) and \(e_s\). There is still uncertainty remaining about the unit cost after the collaboration level \(\theta = e_m^\alpha e_s^{1-\alpha}\) is determined. Afterwards, the supplier learns the true unit cost, but he keeps this information from the manufacturer. At the start of Stage 2, the manufacturer may offer a contract term, depending on the strategy set in the beginning. The supplier then manufactures and delivers the components in the quantity \(q\) and receives payment at the unit price \(w\), as specified in the procurement contract. The manufacturer in turn assembles the end products and sells them in the market at the retail price \(r\).


4 Integrated Supply Chain

We first establish the first-best benchmark under the assumption that the manufacturer and the supplier are integrated as a single firm. A manager of the integrated firm sets the optimal allocation of collaborative efforts between the “manufacturer” division and the “supplier” division as well as the production quantity. Since the unit cost is uncertain in the beginning, it is optimal for the manager to delay the quantity decision until after the cost is realized. Consistent with the assumption in the previous section, the collaboration level $\theta$ is already determined at this point. Thus, the integrated firm faces the problem

\[
(B) \quad \max_{e_m, e_s} \int_0^1 \left( rE[\min\{D, q(z|\theta)\}] - G^{-1}(z|\theta)q(z|\theta) \right) dz - k_m e_m - k_s e_s
\]

s.t. $q(z|\theta) = \arg\max_q \left\{ rE[\min\{D, q\}] - G^{-1}(z|\theta)q \right\}$,

with the constraint $0 \leq \theta = e_m^{\alpha} e_s^{1-\alpha} \leq 1$. This is a stochastic program with recourse (Birge and Louveaux 1997). The optimal solutions, denoted by the superscript $B$ (for “benchmark”) are specified as follows. Note that, in the remainder of the paper, we mainly focus on the optimal efforts and the resulting collaboration level, the variables of our main interest, at the expense of suppressing the discussions on optimal purchase price and quantity.

**Proposition 1 (First-best)** The integrated firm chooses the efforts $e^B_m = \left( \frac{\alpha}{k_m} \frac{k_s}{1-\alpha} \right)^{1-\alpha}$ and $e^B_s = \left( \frac{k_m}{\alpha} \frac{1-\alpha}{k_s} \right)^{\alpha}$, resulting in $\theta^B = 1$.

As expected, the integrated firm opts for the maximum collaboration level $\theta = 1$. The expected unit cost is lowest at this level, and therefore, the firm’s expected profit is largest as its profit margin goes up and the inventory risk (measured in the overage cost) is reduced. The optimal allocation of efforts is also quite intuitive. Observe that the cost-contribution ratios $\frac{k_m}{\alpha}$ and $\frac{k_s}{1-\alpha}$ play key roles. The relative magnitude of these ratios is

\[
\frac{e^B_s}{e^B_m} = \frac{k_m/\alpha}{k_s/(1-\alpha)}.
\]

As the manufacturer’s cost-contribution ratio $\frac{k_m}{\alpha}$ increases, more effort is allocated to the supplier, and vice versa. If the manufacturer and the supplier are symmetric, i.e., $k_m = k_s$ and $\alpha = 1/2$, the optimal efforts are identical: $e^B_m = e^B_s = 1$. Anchoring on these insights as a benchmark, we now consider what happens in a decentralized supply chain where information asymmetry exists and procurement contracting plays a key role.
5 Collaboration Under the Screening Contract

In a decentralized supply chain, the manufacturer specifies the unit price and the quantity of the component in a procurement contract that she offers to the supplier. The main challenge for the manufacturer in this setting is how she structures the contract in the presence of information asymmetry; without perfect knowledge of the unit cost, her ability to offer the contract terms that are favorable to herself is limited. The contract theory points to the screening mechanism as the most efficient form of contracting in this situation. That is, the optimal contract consists of a menu of price-quantity pair, where each pair is tailored to a unit cost realization, or each supplier type.

In order for the manufacturer to implement the screening mechanism, she should offer the contract to the supplier after the latter privately learns his type. Under the sequence of events outlined in Section 3, therefore, the contract is offered at the start of Stage 2 after the collaborative cost reduction is completed and the final value of \( \theta \) is known. Hence, the menu of contracts has the form \( \{(w(z|\theta), q(z|\theta))\} \), where \( w(z|\theta) \) and \( q(z|\theta) \) are the price and the quantity designated for the supplier type \( z \) for a given value of \( \theta \). The optimal menu, denoted by the superscript *, is determined from the following optimization problem.

\[
\max_{\{(w(z|\theta), q(z|\theta))\}} \int_0^1 (rE[\min\{D, q(z|\theta)\}] - w(z|\theta)q(z|\theta)) \, dz
\]

s.t.
\[
\pi_s(z, z|\theta) \geq 0, \quad \forall z \in [0, 1], \quad (IR-S)
\]
\[
\pi_s(z, z|\theta) \geq \pi_s(z, \hat{z}|\theta), \quad \forall z, \hat{z} \in [0, 1]. \quad (IC-S)
\]

Here, \( \pi_s(z, \hat{z}|\theta) \equiv [w(\hat{z}|\theta) - G^{-1}(z|\theta)]q(\hat{z}|\theta) \) is the \( z \)-type supplier’s Stage 2 profit when he accepts the price-quantity pair \((w(\hat{z}|\theta), q(\hat{z}|\theta))\) that is intended for the \( \hat{z} \)-type supplier. The optimal screening mechanism rests on the revelation principle, which states that one can always design an optimal contract under which the supplier voluntarily chooses the price-quantity pair that is designated for him, effectively revealing his true type. This is reflected in the incentive compatibility constraint (IC-S) above. The participation constraint (IR-S) ensures that the resulting profit for the supplier is nonnegative regardless of the realized cost. The solution of \((S_2)\) is as follows.

Lemma 1 The optimal screening contract consists of a menu of price-quantity pairs \( \{(w^*(z|\theta), q^*(z|\theta))\} \) where \( q^*(z|\theta) = F^{-1} \left( 1 - \frac{1}{r} \left[ c + 2\Delta_0 \left( 1 - (1 - \delta)\theta \right) z \right] \right) \) and \( w^*(z|\theta) = c + \Delta_0 \left( 1 - (1 - \delta)\theta \right) \left( z + \int_0^1 q^*(x|\theta) \, dx \right) q^*(z|\theta)\left( 1 - \frac{1}{r} \left[ c + 2\Delta_0 \left( 1 - (1 - \delta)\theta \right) z \right] \right) \).

The Stage 2 expected profits of the manufacturer and the supplier are \( r \int_0^1 J(q^*(z|\theta)) \, dz \) and \( \Delta_0 \left( 1 - (1 - \delta)\theta \right) \int_0^1 zq^*(z|\theta) \, dz \), respectively.

Despite the truth-revealing nature of the optimal screening contract, the first-best cannot be achieved because the manufacturer’s imperfect knowledge of the unit cost leaves the supplier with a positive information rent except for the highest-type supplier (i.e., \( z = 1 \)). The supplier’s expected profit in
Lemma 1 represents the expected information rent over all possible realizations of $z$.

Anticipating these Stage 2 outcomes, the manufacturer and the supplier simultaneously decide the optimal effort levels $e_m$ and $e_s$ in Stage 1, resulting in the collaboration level $\theta = e_m^{\alpha}e_s^{1-\alpha}$. In doing so, they maximize their Stage 1 utilities $U_m(e_m|e_s) = r \int_0^1 J(q^*(z|\theta))dz - k_m e_m$ and $U_s(e_s|e_m) = \Delta_0 (1 - (1 - \delta)\theta) \int_0^1 q^*(z|\theta)dz - k_s e_s$. The equilibrium collaboration level of this game is specified in the next proposition. We use the superscript $S$ (for “screening”) to denote the equilibrium outcomes of this game.

**Proposition 2** *(Equilibrium collaboration level under the screening contract)* The Nash equilibrium of the effort game under the screening contract exists. Let

$$
\Psi(\theta) = \begin{cases} 
\Delta_0 (1 - \delta) [2q^*(1|\theta) + 2\psi(\theta)]^\alpha \psi(\theta)^{1-\alpha} \geq 0 & \text{if } \psi(\theta) \geq 0, \\
-\Delta_0 (1 - \delta) [2q^*(1|\theta) + 2\psi(\theta)]^\alpha [-\psi(\theta)]^{1-\alpha} < 0 & \text{if } \psi(\theta) < 0.
\end{cases}
$$

If $\Psi(\theta) < K$ for all $\theta \in [0,1]$, then $\theta^S = 0$. If $\Psi(\theta) > K$ for all $\theta \in [0,1]$, then $\theta^S = 1$. Otherwise, $\theta^S \in [0,1]$ is found from the equation $\Psi(\theta) = K$.

While Proposition 2 identifies the equilibrium solution for $\theta$, the expressions there do not permit easy interpretations. In addition, an analytical proof of uniqueness of the equilibrium is not readily available, although it is confirmed by numerical examples. However, essential insights can be obtained by investigating the solution behavior in the following limiting case, for which uniqueness of the equilibrium can be verified analytically.

**Corollary 1** Let $\hat{y}$ be the unique root of the function

$$
\hat{\psi}(y) = -y \left( \frac{F(y) - \frac{c}{r}}{\frac{F(x) - \frac{c}{r}}{x}} \right) f(x)dx.
$$

In the limit $K \to 0$, the equilibrium collaboration level $\theta^S$ approaches $\hat{\theta}^S = \min \left\{ \left( \frac{c+2\Delta_0-2F(\hat{y})}{2\Delta_0(1-\delta)} \right)^+, 1 \right\}$. Moreover, $\theta^S \leq \hat{\theta}^S$ for $K > 0$.

$\hat{\theta}^S$ is the upper bound of the equilibrium collaboration level $\theta^S$ which is attained in the limit $K \to 0$. According to the corollary, $\hat{\theta}^S$ can take any value between 0 and 1, depending on the parameter values and the shape of the demand distribution $F$ (we discuss this further in Section 7). In addition, $\hat{\theta}^S < 1$ if $\delta = 0$ (which is easily verified), implying that full collaboration is never attained if collaborative cost reduction is so effective that the residual cost uncertainty can be completely removed. These observations provide strong evidence that, in general, the equilibrium collaboration level under the screening contract tends to be lower than the first-best level $\theta = 1$. Provided that the supplier’s expected cost savings is
greatest at $\theta = 1$ and therefore the manufacturer and the supplier can take the maximum advantage of efficiency gain at that point, the fact that the equilibrium collaboration level tends to be less than one implies that they make suboptimal effort decisions; they forego an opportunity to generate an extra surplus in the supply chain which could have been shared among them.

As we alluded in the Introduction, this deviation from the first-best arises because the manufacturer’s attempt to minimize the impact of her informational disadvantage backfires. Namely, the supplier is reluctant to contribute her share of collaborative effort for fear of being held up by the manufacturer. To be more specific, consider the chain of events after the supplier increases his share of collaborative effort. Higher effort $e_s$ leads to a higher collaboration level $\theta$, which corresponds to a lower mean and uncertainty of the unit cost, as specified by Assumption 1. While smaller average unit cost benefits the supplier as it leads the manufacturer to increase the order quantity, smaller uncertainty does not. Recall that the supplier’s expected profit consists of his information rent. With lower uncertainty about the unit cost, the supplier’s informational advantage is eroded, and as a response, the manufacturer is able to structure the screening contract so that she can extract the supplier’s surplus more effectively. Therefore, collaboration is a double-edged sword for the supplier: on one hand, it will increase his profitability through a volume increase as it lowers the expected unit cost, but on the other hand, he has to guard against ceding his informational advantage to the opportunistic manufacturer. This tradeoff restrains the supplier from fully collaborating. The next result, which compares the allocation of equilibrium efforts $e_{Sm}$ and $e_{Ss}$ with that under the first-best, further identifies the supplier’s unwillingness to collaborate as the primary reason for the suboptimal outcome under the screening contract.

**Proposition 3 (Comparison of effort allocations)** If $0 < \theta^S < 1$, (i) $e_{Ss} < e_{Bs}$ and (ii) $e_{Ss}/e_{Sm} < \frac{1}{2}e_{Bs}/e_{Bm}$.

Part (i) of the proposition makes it clear that the supplier’s effort level in equilibrium is lower than the benchmark level (as long as $0 < \theta^S < 1$). In addition, part (ii) says that the supplier’s relative share of the joint effort under the screening contract is smaller than that of the benchmark case. As an illustration, assume that the manufacturer and the supplier are symmetric with respect to their effort cost-contribution ratios, i.e., $\frac{k_s}{\alpha} = \frac{k_m}{1-\alpha}$. Then the efforts are evenly allocated in the benchmark case ($e_{Bs}/e_{Bm} = 1$), whereas under the screening contract, the supplier’s effort is less than a half of the manufacturer’s ($e_{Ss}/e_{Sm} < \frac{1}{2}$).

In sum, while the screening mechanism used for procurement enables the manufacturer to effectively deal with information asymmetry, it creates a hold-up problem for the supplier. As a result, the supplier has a low incentive to contribute to the collaborative cost reduction effort that precedes procurement. This dynamic suggests that procurement decisions are not to be made independently of the collaborative

---

6It is optimal for the newsvendor manufacturer to increase the order quantity in response to the reduced unit cost of the supplier, since it enables the manufacturer to lower her marginal cost, i.e., the purchase price.
effort decisions during product development. The question is then, what types of procurement contract promote collaboration?

6 Collaboration Under Contract Term Commitments

In the previous section we identified the hold-up problem as the source of inefficient collaboration outcome. A remedy commonly suggested in the literature to resolve this problem is price commitment, under which the manufacturer commits to a price before costly investments are made. Motivated by this, we investigate if contractual commitments, including price commitment, are effective in alleviating inefficiency in our setting as well.

6.1 Price and Quantity Commitments

As we demonstrated in the previous section, the screening mechanism equips the manufacturer with an imperfect but an effective way to deal with information asymmetry but at the expense of discouraging the supplier from collaborating on the joint cost reduction effort. The hold-up problem cannot be avoided as long as the manufacturer makes use of a screening contract, since it should be offered after the unit cost uncertainty is resolved to the supplier—a point in time when collaboration is already completed. This reasoning suggests that abandoning the screening mechanism, and in doing so, breaking up the price-quantity pair in the contract and offering one or both before collaboration starts, may convince the supplier to collaborate more. By committing to a contract term, the manufacturer is able to convey to the supplier that she will not act as opportunistically as she would have with a screening contract.

Since a procurement contract specifies price and quantity, there can be three types of the commitments: price commitment, quantity commitment, and price-quantity commitment. Of the three, price commitment has received most attention in the literature, but we investigate all three (a) for completeness and (b) to illustrate distinct effects of committing to a price and/or a quantity. The disadvantage of contract term commitment is obvious. The manufacturer would have to leave a larger portion of the rent to the supplier than she would without the commitment, since the truth-revealing mechanism can only be implemented if both the price and the quantity contract terms are offered ex-post, i.e., after the unit cost is revealed to the supplier. Therefore, instituting the screening mechanism vs. the commitment strategy can be viewed as a tradeoff between (potentially) incentivizing the supplier to collaborate more and extracting more rents from him.

To see how the commitment strategy works, consider price commitment. At the beginning of Stage 1 the manufacturer offers a price $w$ to the supplier. At this point in time no information asymmetry exists, since the unit cost is yet to be realized and both the manufacturer and the supplier know only its
distribution. Next, each party exerts an effort simultaneously, resulting in the equilibrium collaboration level $\theta$ which determines the mean unit cost. The uncertainty in the unit cost is resolved afterwards, and subsequently the manufacturer offers to the supplier a quantity $q$.\footnote{A combination of $w$ and a menu of quantities, i.e., $\{q(z|\theta)\}$, is insufficient to implement the truth-revealing screening mechanism since at least two contract terms are needed in a menu to satisfy both the individual rationality constraint and the incentive compatibility constraint. If such a contract were offered, the supplier would always choose the same quantity in the menu $\{q(z|\theta)\}$ that maximizes his profit regardless of his type.} Quantity commitment and price-quantity commitment follow similar sequences of events.

Under price commitment, reduced unit cost uncertainty no longer presents a risk to the supplier since the manufacturer lacks a device (i.e., pricing) to take advantage of the reduction later. Therefore, intuition guides us to believe that price commitment will eliminate the hold-up problem and induce the supplier to exert a higher level of collaborative effort, potentially leading to full collaboration. As the following proposition reveals, however, this reasoning tells only a half of the story.

**Proposition 4** Under all three contract commitments, i.e., price, quantity, and price-quantity commitments, neither party exerts collaborative effort in equilibrium: $e_m = e_s = \theta = 0$.

As the proposition asserts, none of the commitment strategies result in full collaboration. In fact, a complete opposite happens: in equilibrium, neither party exerts effort, and therefore, the collaboration level is at the lowest level, $\theta = 0$. This unexpected conclusion is driven by the fact that collaborative cost reduction needs inputs from both the manufacturer and the supplier. Under price commitment, it is true that the supplier is more incentivized to exert effort than he would have been if he were subject to a screening contract. However, the manufacturer is not; with her payment price $w$ fixed at a constant, she does not receive any benefit of collaborative cost reduction since her profit margin $r - w$ is fixed and her quantity is effectively fixed, too, as the optimal quantity is determined by the constant underage cost $r - w$ and the constant overage cost $w$. (Recall that the assembly cost and other costs unrelated to procuring the supplier’s component are normalized to zero.) Since exerting an effort incurs disutility but does not bring any profit increase, the manufacturer does not contribute, i.e., she sets $e_m = 0$. As a response the supplier sets $e_s = 0$, too, since collaboration requires mutual efforts; no synergy can be created with only one party’s effort.

Quantity commitment also fails to bring positive efforts, but for a different reason. In this case the hold-up problem is again the culprit. The optimal price $w$ that the manufacturer sets in Stage 2 is lower if the unit cost uncertainty smaller, implying that the supplier’s profit margin goes down with higher $\theta$. Since the quantity $q$ is fixed, it also means that the supplier’s profit (margin times the quantity) is highest at $\theta = 0$. Hence, the supplier refuses to collaborate and sets $e_s = 0$, and as a response, it is optimal for the manufacturer to set $e_m = 0$, too. Therefore, not all commitments alleviate the hold-up
problem; with quantity commitment, the problem is actually exacerbated. The same result is obtained for price-quantity commitment by a similar reasoning.

Thus, committing to either or both contract terms at the outset of the relationship does not promote collaboration—quite to the contrary, it sti‡ es collaboration. This is because, while collaboration requires both parties’ efforts, the commitment strategies we described above incentivize only either one or neither. Under price commitment, it is the manufacturer who refuses to put in e¤ort. On the other hand, under quantity commitment, it is the supplier, and under price-quantity commitment, it is both. In order for both to be motivated, then, a middle ground should be reached on which the manufacturer can internalize the benefit of cost reduction and at the same time the supplier is not concerned about being held up. In the next subsection we propose a simple contracting scheme that achieves this goal.

6.2 Expected Margin Commitment

Under expected margin commitment (EMC), the manufacturer commits to pay a constant margin \( v \) above the expected unit cost \( \int_0^1 G^{-1}(z|\theta) \, dz \), no matter what collaboration level \( \theta \) results from their mutual efforts. EMC is similar to but different from price commitment, since, while commitment is made on price, the price is not fixed—it decreases with \( \theta \). This is appealing to both the manufacturer and the supplier. The manufacturer receives the benefit of cost reduction since her margin \( r - w(\theta) \) improves while the inventory risk (represented by the overage cost \( w(\theta) \)) becomes smaller. From the supplier’s perspective, EMC encourages exerting an effort since the order quantity increases with \( \theta \) (which can be verified from the proposition below) while his expected profit margin is protected, as it is equal to the constant value \( v \). This is in contrast to the screening contract, under which the supplier’s margin is eroded with collaboration. Hence, EMC has a potential to neutralize the hold-up problem. Taking the two together, we see that collaboration becomes attractive to both parties under EMC.

However, this does not necessarily imply that the manufacturer always prefers EMC, since it does not enable her to extract rents from the supplier as efficiently as she could have with a screening contract. We consider this tradeo¤ further in the next section. First, let us characterize the equilibrium collaboration level under EMC. We use the superscript \( M \) to denote the equilibrium outcomes under EMC.

**Proposition 5 (Equilibrium collaboration level under expected margin commitment)** The Nash equilibrium of the collaborative effort game under EMC exists. Let \( q^1(\theta) \equiv F^{-1} \left( 1 - \frac{1}{\gamma} \left[ v + \zeta + \frac{\Delta \theta}{2} \left( 1 - (1 - \delta) \theta \right) \right] \right) \) and \( \Gamma(\theta) \equiv \frac{1}{2} \Delta \theta (1 - \delta) \left( \frac{v}{\tau_f(q(\theta))} \right)^{1-\alpha} \). If \( \Gamma(\theta) < K \) for all \( \theta \in [0, 1] \), then \( \theta^M = 0 \). If \( \Gamma(\theta) > K \) for all \( \theta \in [0, 1] \), then \( \theta^M = 1 \). Otherwise, \( \theta^M \) is found from the equation \( \Gamma(\theta) = K \).

Note that this proposition is incomplete since it does not specify the optimal value of \( v \), which involves analytical difficulty. However, it is intuitive that the optimal value of \( v \) should be determined
from the binding participation constraint, i.e., the manufacturer should choose the minimum margin for the supplier that ensures his participation in the trade. This is consistent with the equilibrium result in the screening contract case, and indeed, it is what we observe from numerical experiments whenever $K$ is sufficiently small. With the binding constraint the optimal value of $v$ is equal to $v^M = \frac{\Delta_0}{2} (1 - (1 - \delta)\theta^M)$.\textsuperscript{8} In the subsequent analyses we assume this is true, except for the next result which does not rely on this assumption.

**Corollary 2** $\theta^M = 1$ in the limit $K \to 0$.

That is, the equilibrium collaboration level under EMC always approaches its maximum value when the effort costs are negligible. This is in contrast to the analogous result in Corollary 1 for the screening contract case, where we found that the upper bound $\hat{\theta}^S$ may be less than one depending on parameter values. Hence, Corollary 2 provides evidence that EMC tends to bring a higher collaboration level than the screening contract does, as we suspected.

As we mentioned above, however, the manufacturer may not always prefer EMC to the screening contract despite the former’s ability to promote collaboration, because it requires her to leave a larger fraction of surplus to the supplier. We investigate this tradeoff in the next section, with a goal of identifying the conditions under which one contracting approach dominates the other.

### 7 Optimal Contracting Strategies

In this section we compare the performances of the two contracting strategies we studied in the previous sections, namely, the screening contract and EMC, from the manufacturer’s perspective. We focus on the role of demand variability, an important product characteristic that drive many procurement decisions in practice. In our setting, demand variability not only influences the terms of procurement contracts (i.e., price and quantity), but also the supply chain parties’ incentives to collaborate on cost reduction. We elaborate on this below.

The first hint at how demand variability impacts the collaboration level comes from Corollary 1, which specifies the upper bound $\hat{\theta}^S$ of the equilibrium collaboration level under the screening contract. As we found there, the shape of the demand distribution $F$ determines whether $\hat{\theta}^S$ is equal to zero, one, or a value in between. To make this observation more concrete, let us assume that demand is normally distributed and examine how $\hat{\theta}^S$ varies with the standard deviation $\sigma$. The result of this sensitivity analysis is summarized in the next proposition.

\textsuperscript{8}This assumes the ex-post participation constraint $\pi_s(z|\theta^M) \geq 0$, $\forall z \in [0, 1]$, which is consistent with the assumption in the screening contract case. With this, we rule out the possibility that the supplier walks away from the trade if his realized cost is too high.
Proposition 6  Suppose that demand is normally distributed with the mean $\mu$ and the standard deviation $\sigma$. Then $\frac{\partial \tilde{\theta}^S}{\partial \sigma} > 0$ for $0 < \tilde{\theta}^S < 1$.

That is, the upper bound of the equilibrium collaboration level under the screening contract increases with demand variability. This finding hints that a similar statement can be made about the collaboration level for a general case, i.e., $\frac{\partial \tilde{\theta}^S}{\partial \sigma} > 0$ is likely as well. Indeed, this is confirmed by numerical examples. On the surface, this sounds intuitive—more uncertainty brings higher level of collaboration. After all, many studies in the literature tout supply chain collaboration as an important strategic tool to minimize the negative consequences of demand variability. For example, Lee et al. (2004) identify the collaborative demand forecast sharing as one of the four strategies for mitigating the bullwhip effect. However, this naïve intuition does not apply to our setting, since in our model demand information is symmetric; demand forecast sharing is a built-in assumption in our model.

Instead, the result in Proposition 6 arises from a subtle interaction between demand variability and collaborative cost reduction. The reasoning is as follows. Larger demand variability brings a higher demand-supply mismatch risk to the manufacturer, and this prompts her to find a way to compensate for the expected loss. An obvious remedy is to recoup her loss by lowering the payment to the supplier and extract more surplus from him. However, the manufacturer’s ability to do so is limited by the supplier’s unit cost; the higher the unit cost, the smaller the amount of surplus that the manufacturer can take away from the supplier. Hence, it is optimal for the manufacturer to restructure the terms of the screening contract so that the supplier finds it more appealing to put in his share of collaborative effort, lowering the unit cost in the process and thus creating more surplus that the manufacturer can extract from him. The net effect is higher collaboration level. Therefore, a higher collaboration level results from the manufacturer’s self-interested motive, rather than from a goal of creating mutual benefits.

Combining this observation that high demand variability fosters collaboration under the screening contract with that from Corollary 2, namely that the supply chain parties tend to collaborate more readily under EMC, we infer that the difference in $\tilde{\theta}$ between the screening contract and EMC is larger when demand variability is low, while the opposite is true when demand variability is high. An example in Figure 2(a) supports this hypothesis. In this example, $\tilde{\theta}^M = 1$ is always attained in equilibrium under EMC while $\tilde{\theta}^S$ steadily increases with $\sigma$ under the screening contract. However, as Figure 3(a) illustrates, this is not a universal result; there are situations where $\tilde{\theta}^M = \tilde{\theta}^S = 0$ if $\sigma$ is sufficiently small and $\tilde{\theta}^M = \tilde{\theta}^S = 1$ if $\sigma$ is sufficiently large. The difference between these two examples is the degree of cost reduction that can be attained via collaboration; $\delta = 0.8$ in Figure 2, i.e., 20% reduction is achieved by full collaboration, whereas $\delta = 0.95$ in Figure 3, i.e., 5% reduction is achieved. These numerical observations suggest that it is the combination of demand variability and the degree of cost reduction that determines the collaboration level. The remaining question is: under what circumstance is it better
for the manufacturer to adopt EMC contract over the screening contract, and vice versa? The following result, which focuses on special cases, provides an analytical support for the answer to this question.

Proposition 7 Let $\pi^S_m$ and $\pi^M_m$ be the manufacturer’s Stage 2 expected profits in the limit $K \to 0$ under the screening contract and EMC, respectively. For $\pi^S_m$, $\pi^M_m$, and $\bar{y}$, which is specified in Corollary 1, the following holds.

(i) If $F(\bar{y}) \leq \frac{1}{r}(c + 2\delta_0)$, then $\pi^S_m > \pi^M_m$.

(ii) If $F(\bar{y}) \geq \frac{1}{r}(c + 2\Delta_0)$, there is a unique $\delta \in (0, 1)$ such that $\pi^S_m < \pi^M_m$ for $\delta < \tilde{\delta}$ and $\pi^S_m > \pi^M_m$ for $\delta > \tilde{\delta}$.

If demand is normally distributed, it can be shown that the condition in (i) is satisfied when demand variability is sufficiently large, and similarly, (ii) is satisfied when variability is sufficiently small. Under the condition in (i), $\theta^S = \theta^M = 1$. Under the condition in (ii), $\theta^S = 0$ whereas $\theta^M = 1$. According to the proposition, the manufacturer’s Stage 2 expected profit is always higher under the screening contract when demand variability is sufficiently large. On the other hand, when variability is sufficiently small, a similar statement can be made only when $\delta$ is over the threshold value $\tilde{\delta}$; otherwise, EMC produces higher expected profit. Extending this result to include the remaining case ($\frac{1}{r}(c + 2\delta_0) < F(\bar{y}) < \frac{1}{r}(c + 2\Delta_0)$) and accounting for the effort disutilities incurred in Stage 1, we numerically verify the insights suggested by Proposition 7: the manufacturer prefers EMC to the screening contract when (a) a large percentage of cost reduction (small $\delta$) can be achieved through collaboration and/or (b) demand variability is low. This is illustrated in Figure 2(b), which shows that EMC dominates the screening contract especially when

![Figure 2](image_url)

(a) Equilibrium collaboration level

(b) Manufacturer’s utility

Figure 2: In this example, $\delta = 0.8$, $r = 1$, $c = 0.19$, $\Delta_0 = 0.4$, $\alpha = 0.5$, and $k_m = k_s = 0.12$. Demand is normally distributed with the mean 100.
demand variability is small. In contrast, if the degree of achievable cost reduction is relatively small and demand variability is large, then the screening contract becomes more attractive to the manufacturer; see Figure 3(b) that shows the screening contract dominating EMC for large values of $\sigma$.\footnote{In Figure 3(b) it is also observed that the screening contract dominates EMC for very small $\sigma$. This happens because $\theta^M = \theta^S = 0$ in that region (see Figure 3(a)); effort disutility is too high under both contracts. Since the screening contract enables the manufacturer to extract the supplier's surplus more effectively, given the same value of $\theta$, the manufacturer's profit is higher when she uses the screening contract.}

8 Discussion of Assumptions and Robustness of the Results

8.1 Collaboration in a Make-to-Order Production System

Our analysis has been based on the assumption that production lead time is long and therefore the manufacturer has to procure the component well before demand is realized. While this assumption is quite reasonable for many product categories that we used as a motivation, there are situations where the manufacturer is able to operate in a make-to-order production system. This is possible when the total lead time of productions and assemblies is relatively short and the customers’ willingness-to-wait is high. The benefit of having such a system is clear: since the manufacturer does not have to position the products before demands arrive, supply matches demand perfectly, and the inventory risk is eliminated. On the other hand, what is the impact of having a make-to-order system on collaborative cost reduction? The next proposition answers this.

Proposition 8 (Collaboration in a make-to-order production system) The optimal collaboration level is $\theta = 1$ if the supply chain is integrated. In a decentralized supply chain, the manufacturer and the supplier
choose $e_m = 0$ and $e_s = 0$ under both non-commitment and EMC, leading to the equilibrium collaboration level $\theta = 0$. Furthermore, the manufacturer is indifferent between the two contract choices.

Notice that, in the proposition, we used the term “non-commitment” in place of “screening contract”. This is because the screening mechanism cannot be implemented when there is no demand-supply mismatch; since the order quantity is determined after demand is realized, it is optimal to set $q = D$ regardless of the supplier type, and hence, the optimal price-quantity pairs cannot be tailored for each supplier type.

Given our observation from the previous section that $\theta^S$ increases with demand variability (Proposition 6 and the subsequent discussions), it is not unexpected that $\theta = 0$ emerges in equilibrium under non-commitment, as the supply exactly matches the demand in a make-to-order production system and this is equivalent to having zero demand variability in a make-to-stock system. However, it is not immediately clear why the same should be true under EMC; as we found earlier, EMC is effective in incentivizing the supplier to exert a collaborative effort, often leading to $\theta = 1$. This surprising result in fact arises because the make-to-order system acts as a quantity commitment. With the expected margin fixed to a constant under EMC and the expected volume also equal to a constant $E[q] = E[D] = \mu$, exerting a collaborative effort only incurs the effort disutility without affecting the supplier’s expected profit, and hence, it is optimal for him to set $e_s = 0$. Consequently, $\theta = 0$ regardless of the value of $e_m$, and as a response, the manufacturer also sets $e_m = 0$. Thus, neither contracting strategy promotes collaboration in a make-to-order production system.

This finding suggests that an ex-post improvement of supply chain efficiency may bring an unintended consequence: it hinders ex-ante collaboration. One of the main goals of supply chain management is mitigating the impact of demand-supply mismatch, which can be achieved by investing in technologies and resources to improve forecasting accuracy and reducing production lead times. Transforming a supply chain into a make-to-order system is a consummate outcome of such efforts, and the literature touts many benefits associated with it. Our analysis identifies one caveat of these arguments, namely, that the supply chain members’ incentives to collaborate are lowered when they anticipate that the supply chain will operate in the most efficient manner, i.e., when demand-supply mismatch is eliminated. Interestingly, the supply chain members end up in this Prisoner's Dilemma-like situation regardless of procurement contract options.\textsuperscript{10} Although better matching between supply and demand through enhanced forecasting and lead time reduction will contribute to a profit increase, it comes at the expense of discouraging the supply chain members from exerting collaborative efforts during product development. As a result, they may not be able to receive the full benefit of efficiency improvement.

\textsuperscript{10}It can be shown that other forms of commitments that were discussed in Section 6.1, i.e., price commitment, quantity commitment, and price-quantity commitment, also fail to result in $\theta > 0$.\hfill 23
### 8.2 Non-Constant Lower Bound of the Unit Cost Distribution Support

One of the restrictive assumptions of the main model is that the lower bound of the unit cost distribution support is fixed at the constant value \( c \), a simplifying assumption that enable tractable analysis. We now investigate the consequences of relaxing this assumption. In doing so, we maintain the premise that the unit cost function varies linearly in \( \theta \) and that collaboration leads to lower values of both the mean and the spread of the unit cost. Through numerical experiments, we verify that the main findings from the earlier sections remain intact. Namely, EMC tends to promote collaboration better than the screening contract does, and EMC is preferred when demand variability is relatively small and when the degree to which the mean and the spread of the unit cost are reduced through collaboration is relatively large. Hence, no structural changes occur as a result of relaxing the assumption of a constant lower bound.

An interesting special case permitted under the relaxed assumption is the limit \( \delta \to 1 \), i.e., when collaboration leads to reduction of the mean of the unit cost but not of the spread. In this case, the unit cost function \( G^{-1}(z|\theta) \) decreases in \( \theta \) at a uniform rate for all \( z \). It can be proved that, in this case, both the manufacturer and the supplier exert maximum efforts to induce the full collaboration level \( \theta = 1 \) regardless of whether the screening contract or EMC is employed. This is because, in this limit, reduction of the mean leads to expected cost savings and a volume increase for both parties, while at the same time, the supplier does not lose his informational advantage; the net benefits are positive for both. Since both contracts lead to \( \theta = 1 \), the manufacturer prefers the screening contract to EMC since the former is more effective in extracting rents from the supplier for a given level of collaboration.

### 8.3 Uncertainty in Collaboration Level

In specifying the link between the efforts and the collaboration level, we have assumed the deterministic relationship \( \theta = \varepsilon \alpha \mu^1 - \alpha \). Although this captures the essence of the complementary nature of the efforts, in reality, an exogenous shock that neither party can control may result in an uncertain collaboration outcome. To investigate what effects such uncertainty has on our findings, assume an alternate specification \( \theta = \varepsilon \alpha \mu^1 - \alpha \), where \( \varepsilon \) is a random variable with a known distribution with \( E_{\varepsilon}[\varepsilon] = 1 \). With this addition, we modify the sequence of events as follows: (1) the manufacturer and the supplier choose the effort levels simultaneously; (2) \( \varepsilon \) is realized and observed by both parties, determining the collaboration level \( \theta \); (3) the unit cost \( c|\theta \) is realized and is known only to the supplier. In addition, comparison of the collaboration level should be based on the expectation \( E_{\varepsilon}[\theta] \).

Adding this stochastic variable significantly complicates analysis, which is expected given that our model then includes three sources of randomness: demand, the unit cost, and the effort outcome. However, numerical examples reveal that the qualitative insights we obtained in the previous section remain intact. This is in fact not surprising, since the structure of the game does not change except that an additional
expectation operator $E_\varepsilon[\cdot]$ is added to the first-order equations that describe the equilibrium. Therefore, the main conclusions of our model are robust to this additional source of uncertainty.

### 8.4 Retail Pricing

As we mentioned in Section 3.1, it is a common practice that cost reduction efforts are made after a target retail price is set. However, depending on the presence of competing products in the market and consumers’ sensitivity to price, the manufacturer may consider optimally choosing the retail price $r$ after the cost reduction initiative during product development is complete. To investigate the impact of allowing such endogenous retail pricing, assume a linear additive demand curve $D = \beta_0 - \beta_1 r + \epsilon$, where $\epsilon$ is the stochastic part of the demand that is independent of the retail price $r$. We modify the sequence of events such that the manufacturer determines the optimal $r$ at the start of Stage 2, i.e., immediately after collaboration is complete.

We focus on the role of consumers’ price sensitivity, represented by the parameter $\beta_1$; the larger $\beta_1$, the higher the sensitivity. From numerical examples, we find that the collaboration level $\theta$ tends to increase with $\beta_1$, especially under the screening contract. The reason behind this result is qualitatively similar to that of the demand variability result that we discussed in Section 7. With an increasing price sensitivity of the consumers, the manufacturer has to lower the price, which in turn leads to smaller production quantity (as the underage cost becomes smaller). The net outcome is a lower expected profit, and to compensate for the loss, the manufacturer structures the screening contract so that it is palatable for the supplier to collaborate more and create a larger surplus to extract from. In addition, as supported by the examples, EMC tends to promote collaboration better than the screening contract does. Therefore, we conclude that the insights obtained from the earlier analysis remain quite robust even under endogenous retail pricing, which enriches the model and adds another dimension to our discussion.

### 9 Conclusion

In this paper we study how supply chain members’ incentives to collaborate during product development is impacted by information asymmetry and procurement contracting strategies. Despite a high level of interests among the practitioners of supplier management and strategic sourcing, the topic of collaboration has received relatively little attention in the OM literature. We aim to fill this gap by focusing on one of the most important aspects of collaboration, namely, firms’ desire to balance the benefit of collaboration with the need to protect their proprietary information. To this end, we develop a game-theoretic model that captures the incentive dynamics that arise when a manufacturer and a supplier exert collaborative efforts to reduce the unit cost of a critical component during product development, but at the same time,
the supplier is unwilling to fully share his private cost information. We find that the manufacturer’s choice of a procurement contracting strategy critically impacts the supplier’s and the manufacturer’s incentives to collaborate. In addition, we identify demand variability as one of the important environmental variables that influence the collaborative outcome.

We investigate the consequences of using a screening contract, which is appealing to the manufacturer since it is effective in extracting a large fraction of the supplier’s surplus ex-post. However, knowing that the manufacturer’s ability to do so is bounded by uncertainty in the unit cost, the supplier is reluctant to contribute a large amount of effort to the joint cost reduction initiative since collaboration leads to a better estimate of the cost range, thereby eroding his informational advantage. In other words, the supplier’s desire to protect private information about his cost structure lowers his incentive to collaborate and reduces the effectiveness of the screening contract. To resolve this hold-up problem and convince the supplier to collaborate, the manufacturer may instead commit to a contract term before collaboration starts. However, not all commitments work. In particular, the frequently-cited price commitment fails to incentivize either party to collaborate. This happens because, while price commitment does resolve the hold-up problem for the supplier, it leaves the manufacturer with no share of the cost reduction benefit; since the manufacturer does not collaborate, neither does the supplier, as no synergy is created without the efforts by both parties. As an alternative, we propose Expected Margin Commitment (EMC), under which the supplier is guaranteed to earn a fixed margin above the expected unit cost. Our analysis shows that this form of commitment is indeed quite effective in promoting collaboration, and it dominates the screening contract approach in many situations, especially when: (a) a large degree of cost reduction can be attained through collaboration and/or (b) demand variability is relatively small.

As an extension of the model, we also investigate the nature of collaboration incentives in a make-to-order production system. The make-to-order system represents a high level of production efficiency and is achieved by forecasting accuracy improvement and lead time reduction. Surprisingly, such ex-post efficiency improvement is tempered by the ex-ante inefficiency: neither party is willing to collaborate on cost reduction during product development, and hence, production has to proceed with a high unit cost. EMC and other commitments do not alleviate this problem, unlike in a make-to-stock system. Therefore, we conclude that ex-post production efficiency improvement may not represent the full efficiency gain when it depends on the outcome of the ex-ante collaborative efforts.

The insights obtained from our analysis are to be understood in the context of the model assumptions. Relaxing some of these assumptions and including other real-world considerations into the model will undoubtedly enrich the managerial insights and present an opportunity to test the robustness of our findings. For example, in this paper we focus on unit cost reduction as the outcome of collaboration, based on many industry reports that identifies cost reduction as the most important reason that the firms...
establish collaborative relationships. Of course, there are other benefits of collaboration. They include, among others, reduction of product time-to-market (Bhaskaran and Krishnan 2009) and improvement of supplier reliability (Wang et al. 2009). Although this paper focuses on one aspect, a more complete picture of the incentive dynamics where collaboration plays a key role will emerge once the impacts of these and other factors are better understood in future researches.

References


Sloan Management Review. 40(4), 71-82.


Oxford University Press, New York.

[16] Gilbert, S. M., V. Cvsa. 2002. Strategic commitment to price to stimulate downstream innovation 

buyer participation. Naval Research Logistics. 48(1), 41-64.


Wiley.


Manufacturing & Service Operations Management. 3(4), 293-305.

effect. Management Science. 50(12), 1875-1886.


Appendix: Proofs of Selected Results

Lemma 2 Let \( \psi(\theta) = -q^*(1|\theta) + \int_0^1 z q^*(z|\theta) dz \), where \( q^*(z|\theta) \) is defined in Lemma 1. The root of \( \psi(\theta) \), denoted as \( \bar{\theta} \), exists in the interval \( \left( -\frac{r-c-2\Delta_0}{2\Delta_0(1-\delta)}, \frac{1}{1-\delta} \right) \) and is unique. Moreover, \( \psi(\theta) > 0 \) for \( \theta < \bar{\theta} \) and \( \psi(\theta) < 0 \) for \( \theta > \bar{\theta} \).

Proof. Inverting \( q^*(z|\theta) = F^{-1} (1 - \frac{1}{r} r [c + 2\Delta_0 (1 - (1 - \delta)\theta) z]) \) yields \( z = \frac{r F(q) - c}{2\Delta_0(1-(1-\delta)\theta)} \) and \( dz = \frac{-rf(q)}{2\Delta_0(1-(1-\delta)\theta)} dq \). Then

\[
\int_0^1 z q^*(z|\theta) dz = \frac{r^2}{4\Delta_0^2(1 - (1 - \delta)\theta)^2} \int_{q^*(0|\theta)}^{q^*(1|\theta)} (F(q) - \frac{c}{r}) q f(q) dq.
\]

Noting that \( q^*(0|\theta) = F^{-1} (1 - \frac{c}{r}) \) and \( 2\Delta_0 (1 - (1 - \delta)\theta) = r F(q^*(1|\theta)) - c \), we can rewrite this relation with the change of variables \( q \rightarrow x \) and \( q^*(1|\theta) \rightarrow y \) as

\[
\int_0^1 z q^*(z|\theta) dz = \frac{1}{(F(y) - \frac{c}{r})^2} \int_{y}^{F^{-1}(1 - \frac{c}{r})} \left( F(x) - \frac{c}{r} \right) x f(x) dx.
\]

Then \( \psi(\theta) \) can be rewritten as \( \tilde{\psi}(y)/ (F(y) - \frac{c}{r})^2 \), where \( \tilde{\psi}(y) \) is defined in Corollary 1. Hence, \( \psi(\theta) \) and \( \tilde{\psi}(y) \) have the same sign, and therefore, our goal of showing that there exists \( \bar{\theta} \) such that \( \psi(\theta) > 0 \) for \( \theta < \bar{\theta} \) and \( \psi(\theta) < 0 \) for \( \theta > \bar{\theta} \) is achieved by showing that there exists \( \bar{y} \) such that \( \tilde{\psi}(y) > 0 \) for \( y < \bar{y} \) and \( \tilde{\psi}(y) < 0 \) for \( y > \bar{y} \). Note that the lower and upper bounds of \( y \) that are defined for \( \tilde{\psi}(y) \), i.e., 0 and \( F^{-1}(1 - \frac{c}{r}) \), correspond to \( \theta = -\frac{r-c-2\Delta_0}{2\Delta_0(1-\delta)} \) and \( \theta = \frac{1}{1-\delta} \), which are obtained from the relation \( y = q^*(1|\theta) = F^{-1} (1 - \frac{1}{r} r [c + 2\Delta_0 (1 - (1 - \delta)\theta) z]) \). Taking a derivative, \( \tilde{\psi}'(y) = -\eta_1(y)\eta_2(y) \), where \( \eta_1(y) \equiv F(y) - y f(y) - \frac{c}{r} \) and \( \eta_2(y) \equiv F(y) - e/c \). Let \( y_1 \) and \( y_2 \) be the roots of \( \eta_1(y) \) and \( \eta_2(y) \), respectively. Note that \( y_2 = F^{-1}(1 - \frac{c}{r}) \) is equal to the upper bound of \( y \). Since \( F \) has the IGFR property, \( F(y) - y f(y) < F(y) \) that appears in \( \eta_1(y) \) is decreasing and, as a result, \( y_1 < y_2 \), if \( y_1 \) exists. The existence is confirmed by continuity of \( \eta_1(y) \) along with \( \eta_1(0) = 1 - \frac{c}{r} > 0 \) and \( \lim_{y \to \infty} \eta_1(y) = -y_2 f(y_2) < 0 \). Moreover, \( y_1 \) is unique, as proved in Theorem 1 of Lariviere and Porteus (2001). We therefore conclude that there is a unique \( y_1 < y_2 \) such that \( \eta_1(y) > 0 \) for \( 0 < y < y_1 \), \( \eta_1(y_1) = 0 \), and \( \eta_1(y) < 0 \) for \( y_1 < y < y_2 \). This in turn implies \( \tilde{\psi}'(y) > 0 \) for \( y_1 < y < y_2 \). In addition, observe that: (i) \( \tilde{\psi}(0) = \int_0^{y_2} \left( F(x) - \frac{c}{r} \right) x f(x) dx > 0 \), (ii) \( \tilde{\psi}'(0) = -\left( 1 - \frac{c}{r} \right)^2 < 0 \), (iii) \( \lim_{y \to \infty} \tilde{\psi}(y) = 0 \),
and (iv) \( \lim_{y \to y_2} \tilde{\psi}(y) = 0 \). Summarizing, \( \tilde{\psi}(y) \) initially (at \( y = 0 \)) starts from a positive value with a negative slope, flattens out at \( y_1 \), increases as \( y \) goes from \( y_1 \) to \( y_2 \), converging to zero. This implies that there is a unique \( \tilde{y} \in (0, y_1) \) such that \( \tilde{\psi}(y) > 0 \) for \( 0 \leq y < \tilde{y} \), \( \tilde{\psi}(\tilde{y}) = 0 \), and \( \tilde{\psi}(y) < 0 \) for \( y > \tilde{y} \), the result we set out to prove. 

**Proof of Corollary 1.** In the limit \( K \to 0 \), \( \Psi(\theta) = K \) is reduced to \( \psi(\theta) = 0 \). In the proof of Lemma 2 we showed that \( \psi(\theta) = 0 \) can be rewritten as \( \tilde{\psi}(y) = 0 \), which has a unique solution \( \tilde{y} \), with the change of variables \( y = F^{-1}(1 - \frac{1}{r} [\xi + 2\Delta_0 (1 - (1 - \delta)\theta)]) \), or equivalently, \( \theta = \frac{e^{2\Delta_0 - rF(y)}}{2\Delta_0(1 - \delta)} \). The rest of the proof is straightforward and hence omitted.

**Proof of Proposition 6.** For notational convenience, we drop the superscript \( S \). Let \( \Phi \) and \( \phi \) be the cdf and the pdf of the standard normal distribution. Then, if \( 0 < \tilde{\theta} < 1 \), \( \tilde{\psi}(\tilde{y}) = 0 \) and \( \tilde{\theta} \) defined in Corollary 1 can be written as

\[
\begin{align*}
\tilde{\psi}(y) &= -\tilde{y} \left( \Phi \left( \frac{\tilde{y} - \mu}{\sigma} \right) - \frac{c}{r} \right)^2 + \int_{\tilde{y} = \mu}^{\Phi^{-1}(1 - \frac{1}{r})} \left( \Phi (\zeta) - \frac{c}{r} \right) \left( \mu + \sigma \zeta \right) \phi (\zeta) d\zeta = 0, \quad (3) \\
\tilde{\theta} &= \frac{\xi + 2\Delta_0 - r\Phi \left( \frac{\tilde{y} - \mu}{\sigma} \right)}{2\Delta_0(1 - \delta)}, \quad (4)
\end{align*}
\]

where we have let \( \zeta = \frac{\pi - \mu}{\sigma} \) in (3). Implicit differentiation of (3) with respect to \( \sigma \) yields

\[
0 = -\left( \frac{\partial \tilde{y}}{\partial \sigma} \right) \left( \Phi \left( \frac{\tilde{y} - \mu}{\sigma} \right) - \frac{c}{r} \right)^2 + \tilde{y} \left( \frac{\partial \tilde{y}}{\partial \sigma} \right) \left( \Phi \left( \frac{\tilde{y} - \mu}{\sigma} \right) - \frac{c}{r} \right) \frac{1}{\sigma} \phi \left( \frac{\tilde{y} - \mu}{\sigma} \right) - \frac{c}{r} \Phi \left( \frac{\tilde{y} - \mu}{\sigma} \right)
\]

\[
= -\left( \frac{\partial \tilde{y}}{\partial \sigma} \right) \left( \Phi \left( \frac{\tilde{y} - \mu}{\sigma} \right) - \frac{c}{r} \right)^2 + \tilde{y} \left( \frac{\partial \tilde{y}}{\partial \sigma} \right) \left( \Phi \left( \frac{\tilde{y} - \mu}{\sigma} \right) - \frac{c}{r} \right) \frac{1}{\sigma} \phi \left( \frac{\tilde{y} - \mu}{\sigma} \right)
\]

\[
- \frac{\tilde{y} - \mu}{\sigma} \left( \Phi \left( \frac{\tilde{y} - \mu}{\sigma} \right) - \frac{c}{r} \right)^2 + \int_{\tilde{y} = \mu}^{\Phi^{-1}(1 - \frac{1}{r})} \left( \Phi (\zeta) - \frac{c}{r} \right) \phi (\zeta) d\zeta.
\]

Rearranging (3), the last integral can be expressed as

\[
\int_{\tilde{y} = \mu}^{\Phi^{-1}(1 - \frac{1}{r})} \left( \Phi (\zeta) - \frac{c}{r} \right) \phi (\zeta) d\zeta = \frac{\tilde{y}}{\sigma} \left( \Phi \left( \frac{\tilde{y} - \mu}{\sigma} \right) - \frac{c}{r} \right)^2 - \frac{\mu}{\sigma} \int_{\tilde{y} = \mu}^{\Phi^{-1}(1 - \frac{1}{r})} \left( \Phi (\zeta) - \frac{c}{r} \right) \phi (\zeta) d\zeta.
\]

Substituting this,

\[
0 = -\left( \frac{\partial \tilde{y}}{\partial \sigma} \right) \left( \Phi \left( \frac{\tilde{y} - \mu}{\sigma} \right) - \frac{c}{r} \right)^2 + \frac{\mu}{\sigma} \left[ \left( \Phi \left( \frac{\tilde{y} - \mu}{\sigma} \right) - \frac{c}{r} \right)^2 - \int_{\tilde{y} = \mu}^{\Phi^{-1}(1 - \frac{1}{r})} \left( \Phi (\zeta) - \frac{c}{r} \right) \phi (\zeta) d\zeta \right].
\]
The following can be shown by integration by parts: \( \int_a^b \phi(\zeta) \Phi(\zeta) d\zeta = \frac{1}{2} \left( \Phi(b)^2 - \Phi(a)^2 \right) \). Using this relation and after a few steps of algebra, \( \int_{\frac{\mu}{\sigma}}^{\Phi^{-1}(1-\varepsilon/r)} \left( \Phi(\zeta) - \frac{\mu}{\sigma} \right) \phi(\zeta) d\zeta = \frac{1}{2} \left( \Phi \left( \frac{\mu}{\sigma} \right) - \frac{\mu}{\sigma} \right)^2 \). Substituting this back into the above equality, we get \( \frac{\partial \tilde{y}}{\partial \sigma} - \frac{\tilde{y} - \mu}{\sigma} = \frac{\mu}{2\sigma} \eta_1(\tilde{y}) \), where \( \eta_1(y) \equiv \mathcal{F}(y) - yf(y) - \frac{c}{r} = \Phi \left( \frac{\mu}{\sigma} \right) - \frac{\mu}{\sigma} \phi \left( \frac{\mu}{\sigma} \right) - \frac{\mu}{\sigma} \). Recalling from the proof of Lemma 2 that \( \tilde{y} < y_1 \), where \( y_1 \) is the unique solution of \( \eta_1(y) = 0 \) such that \( \eta_1(y) > 0 \) for \( y < y_1 \). Hence, \( \eta_1(\tilde{y}) > 0 \). It is also shown in the same proof that \( \tilde{y} < y_2 \), where \( y_2 \) solves \( \eta_2(y) = 0 \). Since \( \eta_2(y) \) is a decreasing function, \( \eta_2(\tilde{y}) > 0 \). In sum, we have \( \eta_1(\tilde{y}) > 0 \) and \( \eta_2(\tilde{y}) > 0 \). Then, differentiating (4) yields

\[
\frac{\partial \tilde{y}}{\partial \sigma} = \frac{r}{2\Delta_0(1-\delta)} \left( \frac{\partial \tilde{y}}{\partial \sigma} - \frac{\tilde{y} - \mu}{\sigma} \right) \frac{1}{\sigma} \phi \left( \frac{\tilde{y} - \mu}{\sigma} \right) = \frac{r}{2\Delta_0(1-\delta)} \frac{\mu}{2\sigma^2 \eta_1(\tilde{y})} \phi \left( \frac{\tilde{y} - \mu}{\sigma} \right) > 0,
\]

where we used the result \( \frac{\partial \tilde{y}}{\partial \sigma} - \frac{\tilde{y} - \mu}{\sigma} = \frac{\mu}{2\sigma} \eta_2(\tilde{y}) \) obtained above. ■

**Proof of Proposition 7.** Under EMC, \( \theta^M = 1 \) in the limit \( K \to 0 \), according to Corollary 2. Substituting this yields the expected Stage 2 profit \( \pi^M_m = rJ(q^M) \) where \( q^M = F^{-1} \left( 1 - \frac{1}{r} (c + \delta \Delta_0) \right) \). Under the screening contract, \( \pi^S_m = r \int_0^1 J(q^*(z|\theta^S)) dz \) where \( q^*(z|\theta) = F^{-1} \left( 1 - \frac{1}{r} [c + 2\Delta_0 (1 - (1 - \delta)\theta) z] \right) \).

Define \( \chi_\theta(z) \equiv J(q^*(z|\theta)) - J(q^M) \). Observe that

\[
\chi_\theta(z) = \frac{\partial}{\partial z} \left( \int_0^{q^*(z|\theta)} xf(x)dx - \int_0^{q^M} xf(x)dx \right) = q^*(z|\theta)f(q^*(z|\theta)) \frac{\partial q^*(z|\theta)}{\partial z} = q^*(z|\theta)f(q^*(z|\theta)) \left( -\frac{2\Delta_0 (1 - (1 - \delta)\theta)}{rf(q^*(z|\theta))} \right) = -\frac{2\Delta_0 (1 - (1 - \delta)\theta)}{r} q^*(z|\theta) < 0,
\]

and \( \chi_{\theta}''(z) > 0 \), which follows from the fact that \( q^*(z|\theta) \) decreases in \( z \). Therefore, \( \chi_\theta(z) \) is a convex decreasing function. Consider case (i). According to Corollary 1, \( \theta^S = 1 \) if \( \mathcal{F}(\tilde{y}) \leq \frac{1}{2} (c + 2\Delta_0) \). Hence, \( \pi^S_m - \pi^M_m = r \int_0^1 \chi_1(z) dz \) in this case. Since \( \chi_1(z) \) is convex decreasing and \( \chi_1(1/2) = 0 \), we have \( \int_0^{1/2} \chi_1(z) dz > -\int_1^{1/2} \chi_1(z) dz \) and therefore \( \pi^S_m - \pi^M_m = r \int_0^1 \chi_1(z) dz > 0 \). Consider case (ii). According to Corollary 1, \( \theta^S = 0 \) if \( \mathcal{F}(\tilde{y}) \geq \frac{1}{2} (c + 2\Delta_0) \). Hence, \( \pi^S_m - \pi^M_m = r \int_0^1 \chi_0(z) dz \) in this case. Since \( \chi_0(z) \) is convex decreasing and \( \chi_0(\delta/2) = 0 \), \( \chi_0(z) > 0 \) for \( z < \delta/2 \) and \( \chi_0(z) < 0 \) for \( z > \delta/2 \). Therefore, \( \pi^S_m - \pi^M_m = r \int_0^1 \chi_0(z) dz = r \left( \int_0^{\delta/2} \chi_0(z) dz + \int_{\delta/2}^1 \chi_0(z) dz \right) \) becomes negative as \( \delta \to 0 \) and positive as \( \delta \to 1 \), the latter following from the finding in (i). Combining this with \( \frac{\partial}{\partial \delta} \chi_0(z) = \frac{\partial}{\partial \delta} \left( -\int_0^M xf(x)dx \right) = -q^M f(q^M) \frac{\partial q^M}{\partial \delta} > 0 \), we conclude that \( \pi^S_m - \pi^M_m \) crosses zero exactly once from negative to positive as \( \delta \) goes from 0 to 1. ■
Title
Subsidizing the Distribution Channel: Donor Funding to Improve the Availability of Products with Positive Externalities

Presenter
Prof. Terry Taylor, U.C. Berkeley, Haas School of Business

Contact information
Email: taylor@haas.berkeley.edu
Address: 545 Student Services #1900, Berkeley, CA 94720-1900
Phone: 510-643-5435

Co-author
Prof. Prashant Yadav, MIT-Zaragoza International Logistics Program

Contact information
Email: pyadav@mit.edu
Address: Zaragoza Logistics Center - Edificio Náyade 5 C/ Bari, 55 (PLAZA) 50197 Zaragoza
Phone: +34 625 324 670
Subsidizing the Distribution Channel: Donor Funding to Improve the Availability of Products with Positive Externalities

Abstract

Large populations in the developing world lack adequate access to products, such as essential medicines, whose use confers positive externalities. Because of the critical role played by the private-sector distribution channel, donors (e.g., the Global Fund) are beginning to devote substantial resources to fund subsidies that encourage the channel to improve the availability of these products. A key question for a donor is whether it should subsidize the purchases or sales of the private channel. We find the answer depends crucially on whether consumers are homogeneous or heterogeneous in their valuation of the product. For the case of heterogeneous consumers’ valuations, we provide evidence that subsidizing sales leads to greater expected donor utility, consumption, and social welfare. When consumers’ valuations are homogeneous, we establish the opposite conclusion: subsidizing purchases leads to greater expected donor utility, consumption, and social welfare.

Key words: global health supply chains; developing country supply chains; subsidies; externalities
1 Introduction

In developing countries, products for health, food and other basic needs are distributed through both a government-run system and a private-sector distribution channel. In many countries, the private distribution channel plays a much more important role in providing access to these products (Bustreo et al. 2003, International Finance Corporation 2007), in part because of the limited geographic reach of the government-run outlets. In some countries, over 70% of the population accesses products of this type via the private distribution channel (Sabot et al. 2008). Private distribution channels have been remarkably successful in distributing consumer products such as soft drinks in poor and remote parts of the world (Yadav et al. 2010) and thus offer enormous possibilities for expanding access to products with a societal benefit. However, for products such as medicines and vaccines, where the benefits of consuming the product do not accrue solely to the consumer, private supply chains fail in providing high levels of availability of the product (Cameron et al. 2009). For example, a study in Senegal found that only 8% of the private shops had the recommended drug for malaria in stock and in those shops the stocking level was very low (Sabot et al. 2008).

There are a large number of products whose use benefits both the individual using the product and society at large (i.e., the products possess positive externalities). Examples include mosquito nets, condoms, and various medicines (e.g., oral rehydration salts for treatment of diarrhea) and vaccines. Among products with positive externalities, many have limited shelf lives. Notable examples include vaccines such as those for human papillomavirus and hepatitis B, Artemisinin Combination Therapies (ACTs) for malaria treatment, and some fortified foods. Some of these products, such as malaria drugs, have a short demand season due to the seasonal peaks observed in malaria incidence coinciding with the rainy season. Other products, such as vaccines and fortified foods, are distributed in campaigns which are run for a fixed duration. The disease incidence, transmission intensity, peak amplitudes and lengths of the disease season, and food yield and famine can vary significantly from year to year. Consequently, the demand for malaria drugs, vaccines and other specific health products is highly uncertain, especially at the level of the small geographic region (e.g., a village) which an individual retailer serves.

For many products with positive externalities, especially medicines, vaccines and fortified foods, bilateral donors such as the U.S. Government, multilateral agencies such as the World Bank and the Global Fund to Fight AIDS, Tuberculosis and Malaria, large non-governmental organizations such as the Clinton Health Access Initiative, and private philanthropic organizations such as the Bill and Melinda Gates Foundation, intervene to improve access to populations in poor countries. Because of the important role played by the private distribution channel, a primary way donors seek to achieve this objective is by designing and then funding product subsidies that encourage the channel to make decisions (e.g., stocking and/or pricing decisions) that increase the availability of the product to end
A key question in designing a subsidy is how it should be administered in the supply chain. One option is to administer the subsidy upstream: reduce the cost of each unit the channel acquires via a *purchase subsidy*. A second option is to administer the subsidy downstream: increase the revenue for each unit the channel sells via a *sales subsidy*. Both forms of donor-funded subsidies are used in practice. For example, the Affordable Medicines Facility-malaria (AMFm) for ACTs is a purchase subsidy, which reduces the acquisition cost of these medicines for the distribution channel (Arrow et al. 2004, Adeyi and Atun 2010). Voucher schemes have been used to implement sales subsidies for products such as insecticide-treated bednets (ITNs) used in the prevention of malaria. The voucher provides a means by which the subsidy provider can verify a retailer’s sales to end consumers. A consumer presents a voucher when purchasing an ITN and receives a discount. For each redeemed voucher the retailer submits, the retailer receives a subsidy payment (Mushi et al. 2003).

The purpose of this paper is to compare the effectiveness of a purchase subsidy versus a sales subsidy from the perspective of the donor and society. We focus on this question at the point where products are made available to consumers, the retail outlet. We find that the attractiveness of the subsidy type depends crucially on whether consumers are homogeneous or heterogeneous in their valuation of the product. For the case where consumers’ valuations are heterogeneous, we provide evidence that the sales subsidy leads to greater expected donor utility, consumption, and expected social welfare. When consumers’ valuations are homogeneous, we establish the opposite conclusion: the purchase subsidy leads to greater expected donor utility, consumption, and expected social welfare.

The remainder of the paper is organized as follows. §2 reviews the relevant literature. §3 describes the model. §4 examines the case with homogeneous consumers and §5 the case with heterogeneous consumers. §6 provides concluding remarks.

### 2 Literature Review

Because this paper takes the perspective of a donor, the literature that examines the underlying motivations of donors is relevant (Arrow 1972). Schwartz (1970), Becker (1974) and Sugden (1982) argue that individuals and private charities, in making philanthropic donation decisions, are utility maximizers, where the donor’s utility depends on her contributions to the utility of others. Our paper is in this spirit.

Several papers in the economics literature starting with Pigou (1932) have examined the design of subsidies to encourage consumption of products with positive externalities. In this stream of literature, the donor’s utility depends on the recipient’s consumption of the product (e.g., Ben-Zion and Spiegel 1983). Daly and Giertz (1972) argue that if the donor’s utility depends on the recipient’s
consumption choices, then the donor prefers a product subsidy to a cash transfer. In a product diffusion model, Kalish and Lilien (1983) examine how a donor should vary her product subsidy over time to accelerate consumer adoption. In a field study, Dupas (2010) finds that short-term subsidies positively impact long-term adoption of ITNs. Our work differs from these papers in that they focus on the impact of subsidies on consumption levels, assuming product availability, whereas we focus on the impact of subsidies on product availability, which in turn impacts consumption.

In an epidemiological model of malaria transmission, immunity, and drug resistance, Laxminarayan et al. (2006, 2010) study the impact of reductions in the retail price of ACTs on consumption. From simulation results under plausible parameters, they conclude that donor-funded price reductions are welfare enhancing. To capture the richness of disease progression and resistance, Laxminarayan et al. (2006, 2010) abstract away from the details of the distribution channel. We complement their macro-level approach with a micro-level approach that focuses on capturing these details: demand uncertainty, supply-demand mismatch, and the impact of subsidies on stocking and pricing decisions in the distribution channel.

Our micro-level approach is part of a stream of work in operations management that looks at the impact of incentives on the behavior of firms in a supply chain. In contrast to our focus on how donors should design subsidies to increase product availability, this literature looks at how firms should design contracts to maximize their own profits. See Cachon (2003) for an excellent review. In a classical setting, a manufacturer designs contract terms for its retailer, who faces uncertain demand in a single-period selling season. Lariviere and Porteus (2001) examine how the manufacturer should set the per-unit purchase price. A reduction in this purchase price (a so-called trade deal) is analogous to our purchase subsidy. Taylor (2002), Krishnan et al. (2004), and Aydin and Porteus (2009) examine rebates, a payment the manufacturer makes to the retailer for each unit the retailer sells to end consumers, which is analogous to our sales subsidy. Distinguished by their setting of deterministic demand that occurs continuously over a time horizon, Drèze and Bell (2003) compare trade deals with scan-back rebates (payments based on retailer sales during a specified promotion period) and find that the manufacturer prefers to offer a rebate because this eliminates the retailer’s stockpiling of inventory (so-called forward buying) which occurs with trade deals.

3 Model

We focus on availability at the point where products are made available to consumers, with special attention to rural areas, because here the problem of availability is most acute, in part due to the limited reach of government-run distribution systems outside of urban areas. Because traveling longer distances is difficult for consumers, the retail market is segmented geographically, often with only one retail outlet (e.g., a drug shop or general store) for the product in a given market area, which
consists of one or a small number of villages (see Goodman et al. 2009 for evidence in Tanzania).

We model product availability through the decisions of a retailer that sells to end consumers. A donor (she) offers a subsidy to the retailer (he) with the intention of increasing availability. Consumer demand depends on the market state $M$, a random variable with distribution $F$, density $f$, mean $\mu$, and continuous support on $[\underline{m}, \overline{m}]$, where $0 \leq \underline{m} < \overline{m} \leq \infty$. For simplicity in exposition, we suppose that $\underline{m} = 0$ and $F(0) = 0$. Assume that $F$ is an Increasing Failure Rate (IFR) distribution. Let $F^(-1)(\cdot) \equiv 1 - F(\cdot)$. Let $F^(-1)(x)$ denote the inverse of $F$ for $x \in [0, 1]$, let $F^(-1)(x) = 0$ for $x > 1$ and $F^(-1)(x) = \overline{m}$ for $x < 0$.

The sequence of events is as follows: First, the donor offers a per-unit sales purchase subsidy $a \geq 0$ and/or a per-unit sales subsidy $s \geq 0$. The retailer acquires $Q$ units, incurring per-unit acquisition cost $c > 0$ less the per-unit purchase subsidy $a$. The market state uncertainty is resolved, and the retailer observes the market state $M = m$. The retailer sets the retail price $p \geq 0$, and demand $D_k(m, p)$ is realized, where $k \in \{e, o\}$. Finally, for each unit the retailer sells, the retailer receives the retail price $p$ from the end consumer, the donor pays the retailer the sales subsidy $s$, and the donor receives utility $v \geq 0$. We refer to $v$ as the donor’s value; $a$ is mnemonic for the acquisition cost reduction conferred by the purchase subsidy. Instead of explicitly modeling the cost to administer each type of subsidy, we discuss the impact of such costs in §§4.3 and 5.3.

Consumption of a unit generates utility both for the immediate consumer and for the larger society, and it is this externality benefit that motivates the donor’s interest in expanding the availability of the product. Let $e \geq 0$ denote the value that consumption of a unit generates for society, excluding the immediate consumer; we refer to $e$ as the externality. Thus, if the immediate consumer values a unit of consumption at $\rho_0$, then the social value derived by the consumption of that unit is $e + \rho_0$. For most donors, the externality is an upperbound on the value that a donor places on consumption of a unit $v \leq e$, although we do not impose this as a requirement of our analysis.

We consider two demand models

$$D_e(m, p) = \begin{cases} \frac{(m - p)}{b} & \text{if } p \leq m \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$D_o(m, p) = \begin{cases} m & \text{if } p \leq r \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

In (1), demand decreases linearly in the retail price. This corresponds to the setting where consumers are heterogeneous in their valuation for the product, specifically, where consumer valuations are uniformly distributed on $[0, m]$ with density $1/b$. In (2), demand is of a step-function form. This corresponds to the setting where consumers share a homogeneous valuation for the product, given by $r$.

In practice, consumer valuations may be relatively homogeneous or heterogeneous. Three factors
favor homogeneity in the consumers’ valuations. First, a retail market that is small and geographically concentrated will tend to exhibit limited heterogeneity. Second, severe income constraints in some markets prevent large differentials in the willingness to pay for the product. Third, many products with positive externalities are sold alongside similar products that lack these externalities. For example, the use of the long-standing antimalarial sulfadoxine pyremethamine (SP) contributes to the rise of drug-resistant strains of malaria, while the more recently developed ACTs are much more effective in discouraging resistance. The socially-inferior product’s price serves as a reference point (Kahneman et al. 1986) which conditions consumers’ valuations of the socially-superior product. To the extent that consumers are ill-informed as to the benefits of the socially-superior product, or do not believe that the benefit accrues to them personally, they may be unwilling to pay more for the superior product. As a consequence, consumer valuations for the superior product may be relatively concentrated around the price of the competitor product. When none of these factors is prominent, consumers’ valuations may be relatively heterogeneous. If, in addition, the benefits that the product confers depend heavily on idiosyncratic characteristics of the consumer (e.g., the severity of the consumer’s illness), this heterogeneity may be pronounced.

As noted above, we are especially interested in a retailer serving a rural market, because here the problem of availability is most acute. High travel costs for shop owners, who must often travel long distances (e.g., to the capital city) to purchase the product, mean that shop owners purchase infrequently, with limited opportunity to reorder during the selling season (e.g., the rainy season for malaria medicines). Because of the long lead time from the point of manufacturer to the point of retail sale, the remaining shelf-life of a product is often sufficiently short that unsold product from one season cannot be sold in the subsequent season (or campaign). These factors support our assumption of a single ordering opportunity. Our assumption that the retailer sets his market price in response to market conditions reflects the fact that, once the season is underway, it much easier for a retailer to adjust his price than his acquisition quantity. Rural retailers in developing countries have substantial discretion in adjusting their prices in response to market conditions: typically, they do not advertise or otherwise post prices in advance, and they face little in the way of regulatory scrutiny.

The homogeneous valuation case (equation (2)) is consistent with the classical newsvendor setting in which the retail price is exogenous, which has received substantial attention in the literature (e.g., Lariviére and Porteus 2001, Krishnan et al. 2004). A separate stream of papers (e.g., Petruzzi and Dada 1999, Drèze and Bell 2003, Aydin and Porteus 2009) considers consumers that are heterogeneous in their valuations. We seek to explore both settings in a unified framework, which allows us to explore the impact of consumer heterogeneity on optimal subsidy design.
4 Homogeneous Valuations

This section examines the case where consumers’ valuations are homogeneous (demand is given by (2)), which is an approximation of the case where valuation heterogeneity is relatively limited. Although our primary focus is to compare the attractiveness to the donor of the two subsidy types, for compactness, we begin by examining the retailer’s problem under a purchase and sales subsidy, because this provides insight into the retailer’s problem when he receives only one type of subsidy. Regardless of the subsidy the retailer receives, it is optimal for the retailer to choose retail price \( \pi = \rho \).

Under a purchase and sales subsidy, the retailer’s problem is to choose his acquisition quantity \( \theta \) to maximize his expected profit

\[
\max_{\theta \geq 0} \{(r + s)E \min(M, Q) - (c - a)\theta\}.
\]

(3)

The retailer pays the acquisition cost less the purchase subsidy \( c - a \) for each unit he acquires, and then receives the sum of the retail price and the sales subsidy \( r + s \) for each unit he sells. The retailer faces a standard newsvendor problem, and his optimal acquisition quantity is

\[
\theta_b = F^{-1}((c - a)/(r + s)),
\]

(4)

where the subscript is pneumonic for both purchase and sales subsidy.

We conclude this subsection with two benchmarks. Under no subsidy \( (a = s = 0) \), the retailer’s optimal acquisition quantity is \( Q = F^{-1}(c/r) \). Expected social welfare under acquisition quantity \( Q \) is the sum of the utility derived by consumers, the retailer and the rest of society for each unit consumed, \( e + r \), less the cost of acquiring the \( Q \) units

\[
W(Q) \equiv (e + r)E \min(M, Q) - cQ.
\]

(5)

The quantity which maximizes expected social welfare is \( Q = F^{-1}(c/(e + r)) \).

4.1 Purchase Subsidy

The retailer’s acquisition quantity problem under a purchase subsidy is a special case of (3) where the sales subsidy \( s = 0 \). The retailer’s optimal order quantity is

\[
Q_a = F^{-1}((c - a)/r).
\]

(6)

Under a purchase subsidy, the donor’s problem is to choose the per-unit purchase subsidy \( a \geq 0 \) to maximize her expected utility

\[
vE \min(M, Q_a) - aQ_a.
\]

The donor receives utility \( v \) for each unit the retailer sells, but must pay the retailer the purchase subsidy \( a \) for each unit the retailer acquires. In offering a purchase subsidy, the donor trades off the
benefit of a stochastically larger sales quantity against the up-front cost of a larger subsidy payment to the retailer.

From (6), the subsidy required to induce acquisition quantity \( Q_a \) is

\[
a = c - rF(Q_a). \tag{7}
\]

Because there is a one-to-one mapping between the donor’s purchase subsidy and the retailer’s quantity, we can rewrite the donor’s problem as one of choosing the quantity \( Q \in [Q, M] \) to maximize

\[
U_a(Q) \equiv vE \min(M, Q) - cQ + rF(Q)Q. \tag{8}
\]

We observe in the appendix (Lemma 2) that \( U_a(Q) \) is unimodal in \( Q \).

Proposition 1 characterizes the donor’s optimal purchase subsidy \( a^* \) and the resulting retailer’s acquisition quantity \( Q_a^* \). All proofs are in the appendix.

**Proposition 1** If

\[
v > \max((r^2/c)f(Q)Q, c - r), \tag{9}
\]

then the donor’s optimal purchase subsidy \( a^* \) is the unique solution to

\[
rv - (r + v)a^* - r^2F^{-1}((c - a^*)/r)f(F^{-1}((c - a^*)/r)) = 0; \tag{10}
\]

\( a^* > 0 \); and the optimal resulting retailer’s acquisition quantity \( Q_a^* \) is given by the unique solution to

\[
vF(Q_a^*) - c + r[F(Q_a^*) - Q_a^*f(Q_a^*)] = 0. \tag{11}
\]

Otherwise, the optimal purchase subsidy \( a^* = 0 \), and the resulting acquisition quantity is \( Q_a^* = Q \).

In the case with homogeneous valuations, the retailer’s acquisition quantity \( Q_a^* \) determines the level of product availability. As one would expect, product availability is increasing in the donor’s value \( v \) and decreasing in the retailer’s acquisition cost \( c \). The impact of the consumers’ valuation \( r \) is more subtle, as Proposition 2a demonstrates. Let \( \tau \) denote the unique solution to

\[
\tau - F^{-1}(\tau)f(F^{-1}(\tau)) = 0. \tag{12}
\]

Note that \( \tau \in (0, 1) \). Let \( W_a^* = W(Q_a^*) \) denote expected social welfare under the optimal purchase subsidy.

**Proposition 2** Suppose (9) holds. (a) Under the optimal purchase subsidy \( a^* \), the retailer’s optimal acquisition quantity \( Q_a^* \) is strictly decreasing in the consumers’ valuation \( r \) if and only if

\[
c/v < \tau. \tag{13}
\]
(b) The optimal purchase subsidy $a^*$ is strictly decreasing in the consumers’ valuation $r$ and strictly increasing in the acquisition cost $c$ and donor value $v$. (c) If (13) holds, then there exists finite $\overline{\epsilon}$ such that if the externality $\epsilon > \overline{\epsilon}$, then expected social welfare under the optimal purchase subsidy $W_a^*$ is strictly decreasing in the consumers’ valuation $r$.

The assumption in the first line of the Proposition is innocuous in that it simply restricts attention to the interesting region where the optimal purchase subsidy is non-zero.

It is natural to expect that as the consumers’ valuation of a product increases (e.g., due to improved product quality), product availability and social welfare would increase. Proposition 2 reveals that the opposite may occur. When the retailer’s acquisition cost is small relative to the donor’s value, product availability decreases as the consumers’ valuation of the product increases (Proposition 2a). When, in addition, the externality $\epsilon$ is large, expected social welfare decreases as the consumers’ valuation of the product increases (Proposition 2c). Intuitively, an increase in the consumers’ valuation $r$ (and hence the retail price $p$) makes it attractive for the retailer to stock more aggressively; consequently, the donor responds by reducing her subsidy $a^*$ (Proposition 2b). When the retailer’s acquisition cost is small and the donor’s value is large, this subsidy-reduction effect outweighs the valuation-increase effect, and the retailer reduces his acquisition quantity $Q_a^*$. When the externality $\epsilon$ is large, the negative impact of this reduced availability on social welfare outweighs the the positive impact from the higher value $r$ generated per unit consumed.

4.2 Sales Subsidy

The retailer’s acquisition quantity problem under a sales subsidy is a special case of (3) where the purchase subsidy $a = 0$. The retailer’s optimal order quantity is

$$Q_s = \overline{F}^{-1}(c/(r + s)).$$

Under a sales subsidy, for each unit the retailer sells, the donor receives utility $v$ less the sales subsidy $s$ payment to the retailer. The donor’s sales subsidy problem is to choose the per-unit sales subsidy $s \geq 0$ to maximizes her expected utility

$$(v - s)E \min(M, Q_s).$$

In designing a sales subsidy, the retailer trades off the benefit of a stochastically larger sales quantity against the cost of the subsidy payment, which is increasing in the realized demand.

From (14), the subsidy required to induce acquisition quantity $Q_s$ is

$$s = c/\overline{F}(Q_s) - r.$$ (15)

Because there is a one-to-one mapping between the donor’s purchase subsidy and the retailer’s
quantity, we can rewrite the donor’s problem as one of choosing the quantity $Q \in [Q, \infty)$ to maximize

$$U_s(Q) \equiv [r + v - c/\bar{f}(Q)]E \min(M, Q).$$

Proposition 3 characterizes the donor’s optimal purchase subsidy $s^*$ and the resulting retailer’s acquisition quantity $Q^*_s$. The proof of the proposition establishes that $U_s(Q)$ is strictly concave.

**Proposition 3** If

$$v > \max\left(\frac{r^3}{c^2}E \min(M, Q) f(Q), c - r\right)$$

then the donor’s optimal sales subsidy $s^*$ is the unique solution to

$$c^2(v - s^*) - f(\bar{F}^{-1}(c/(r + s^*))) \left[\begin{array}{c}
(r + s^*)^3 \int_0^{\bar{F}^{-1}(c/(r + s^*))} mdF(m) + c(r + s^*) \bar{F}^{-1}(c/(r + s^*))
\end{array}\right] = 0;$$

$$s^* > 0; \text{ and the optimal resulting firm’s acquisition quantity } Q^*_s \text{ is given by the unique solution to}$$

$$(r + v)\bar{F}(Q^*_s) - c \left(1 + \frac{f(Q^*_s)}{\bar{F}(Q^*_s)^2}\right) E \min(M, Q^*_s) = 0.$$  

Otherwise, the optimal sales subsidy $s^* = 0$, and the resulting acquisition quantity is $Q^*_s = Q$.

As in the case for the optimal purchase subsidy, under the optimal sales subsidy: as the acquisition cost $c$ increases (through the range where the optimal subsidy is strictly positive), the optimal subsidy increases but availability decreases; and as the donor’s value $v$ increases, both the subsidy and availability increase. In contrast to the results for the optimal purchase subsidy (Proposition 2a and 2c), under the optimal sales subsidy, availability and expected social welfare always increase in the consumers’ valuation $r$.

### 4.3 Comparison of Subsidies

In this section, we compare the two subsidy types from the perspective of the donor, society, consumers and the retailer—with an emphasis on the perspective of the donor, the designer of the subsidy. One might expect that the donor’s preference between the subsidy types would depend on the problem parameters. Theorem 1 shows that this conjecture is false.

Let $U_s^*(U^*_a, \text{ respectively})$ denote the donor’s expected utility under the optimal sales (purchase, respectively) subsidy. Let $W^*_s$ denote the donor’s expected utility under the optimal sales subsidy, and recall that $W^*_a$ denotes the analogous value under the optimal purchase subsidy. Observe that $U^*_k = U_k(Q^*_k)$ and $W^*_k = W(Q^*_k)$ for $k \in \{a, s\}$.

**Theorem 1** (a) The donor’s expected utility is higher under the optimal purchase subsidy than under the optimal sales subsidy

$$U^*_a \geq U^*_s.$$
where the inequality is strict if (9) holds. (b) If $v \leq c$, then expected social welfare is higher under the optimal purchase subsidy than under the optimal sales subsidy

$$W_a^* \geq W_s^*, \tag{19}$$

where the inequality is strict if (9) holds. (c) The retailer’s acquisition quantity under the optimal purchase subsidy is higher than under the optimal sales subsidy

$$Q_a^* \geq Q_s^*, \tag{19}$$

where the inequality is strict if (9) holds. Therefore, the retailer’s expected sales quantity is higher under the optimal purchase subsidy than under the optimal sales subsidy.

From Theorem 1a, the donor is better off offering a purchase subsidy than a sales subsidy. To build intuition, observe that in designing a subsidy, the donor is only concerned about the quantity $Q$ the retailer purchases—as this determines the expected benefit $vE \min(M, Q)$ the donor receives—and the cost of inducing that acquisition quantity. Because the purchase subsidy influences the retailer’s acquisition quantity more directly, it is more cost effective in encouraging the retailer to purchase a larger quantity. Formally, the cost of inducing any acquisition quantity $Q \geq Q$ is higher under the sales subsidy: Under a purchase subsidy, the cost of inducing acquisition quantity $Q \geq Q$ is the product of the per-unit subsidy and the retailer’s acquisition quantity

$$aQ = [c - rF(Q)]Q,$$

where the equality follows from (7). Under a sales subsidy, the expected cost of inducing acquisition quantity $Q \geq Q$ is the product of the per-unit subsidy and the retailer’s expected sales quantity

$$sE \min(M, Q) = [c - rF(Q)] \left( Q + \int_0^Q mdF(m)/F(Q) \right),$$

where the equality follows from (15). Because the subsidy cost is higher under the sales subsidy $sE \min(M, Q) > aQ$, the purchase subsidy is superior. The superiority of the purchase subsidy is larger for larger quantities (the subsidy cost difference $sE \min(M, Q) - aQ$ is increasing in $Q$), which suggests that the superiority of the purchase subsidy to the donor, $U_a^* - U_s^*$, grows as the donor’s value $v$ increases.

From Theorem 1c, the retailer’s optimal acquisition quantity (and hence, expected consumption) is higher under the optimal purchase subsidy than under the optimal sales subsidy. The intuition stems from the fact that the cost advantage the purchase subsidy has over the sales subsidy also holds in the marginal sense. Because the donor’s marginal cost of increasing the retailer’s acquisition quantity is lower under the purchase subsidy (i.e., $(\partial/\partial Q)aQ \leq (\partial/\partial Q)sE \min(M, Q)$), it is optimal for the donor to induce a larger acquisition quantity under the purchase subsidy (i.e., $Q_a^* \geq Q_s^*$).
From Theorem 1b, expected social welfare is higher under the optimal purchase subsidy than under the optimal sales subsidy, provided that the donor’s value is less than the externality \( v \leq e \).

To see the intuition, observe that expected social welfare is determined by the retailer’s acquisition quantity (see (5)). Although the donor’s subsidy (of either type) increases product availability, availability is still lower than the socially optimal level

\[ Q_k^* \leq \bar{Q} \]  

for \( k \in \{a, s\} \). Because the optimal purchase subsidy does more to improve product availability than the optimal sales subsidy, the purchase subsidy results in higher social welfare. The inadequate availability result (20) stems from the fact that, in designing the subsidy, the donor cares about getting products into the hands of consumers but is indifferent to the retailer’s profitability. (For completeness, we note that in the unlikely case that the donor values consumption at a much higher rate than the externality \( v >> e \), inequalities (19) and (20) can be reversed.)

We illustrate our results using ACTs as a case example. We consider the setting where consumer valuations’ are anchored by the price of the socially-inferior competitor product, SP. We use the retail price of SP (ACTwatch 2008) as our estimate for the consumers’ valuation \( r \). We use data from Kindermans et al. (2007), Mosha et al. (2010) and Patouillard et al. (2010) to estimate \( c \), data from Chima et al. (2003) to estimate \( v \) and \( e \), and data from ACTwatch (2008) to estimate \( \mu \). We obtain \( c = 1.63 \), \( v = 2.20 \), \( e = 13.83 \), \( r = 0.54 \), and \( \mu = 435 \). There is lack of systematic measurement of demand variability in field studies; we simply assume that \( M \) is a Normal(\( \mu, \sigma \)) random variable, truncated such that its probability mass is distributed over \( x \geq 0 \) and that \( \sigma = \mu/2 \). Without a subsidy, the retailer’s optimal acquisition quantity \( Q = 0 \). The optimal purchase subsidy \( a^* = 1.25 \) results in retailer acquisition quantity \( Q_a^* = 329 \), expected sales of 293, expected donor utility \( U_a^* = 235 \), and expected social welfare \( W_a^* = 3680 \). The optimal sales subsidy \( s^* = 1.41 \) results in significantly lower availability, donor utility and social welfare: \( Q_s^* = 239 \), expected sales of 224, \( U_s^* = 177 \) and \( W_s^* = 2835 \). This illustrates that the donor’s choice of subsidy type can have a substantial impact on the donor, consumers, and society. The results are qualitatively similar when \( \sigma \in \{\mu/4, 3\mu/4\} \).

In our analysis we have ignored the administrative costs associated with implementing either subsidy. Typically, one would expect these costs to be higher under a sales subsidy because of the effort required in verifying the retailer’s sales to end consumers. Taking into account these administrative costs would reinforce Theorem 1’s conclusion that the donor and society are better off under a purchase subsidy.

The donor would prefer to offer a purchase subsidy rather than a sales subsidy. Ignoring the administrative costs entailed with offering both subsidies, would the donor be better off offering
both subsidies types simultaneously? When offering both types of subsidies, the donor’s problem is to choose the per-unit purchase subsidy \( \alpha \geq 0 \) and per-unit sales subsidy \( \sigma \geq 0 \) to maximize her expected utility

\[
(v - s)E \min(M, Q_b) - aQ_b,
\]

where, recall, (4) specifies the retailer’s optimal acquisition quantity \( Q_b \). Let \((a^*_b, s^*_b)\) denote the donor’s optimal purchase and sales subsidy. Recall that \( a^* \) denotes the optimal purchase subsidy, which is characterized in Proposition 1.

**Proposition 4** The purchase and sales subsidy that maximizes the donor’s expected utility has no sales subsidy. The optimal purchase and sales subsidy is simply the optimal purchase subsidy: \((a^*_b, s^*_b) = (a^*, 0)\).

One might expect that the donor would be better off by employing both subsidies simultaneously rather than restricting herself to the simple purchase subsidy, at least for some problem parameters. Proposition 4 shows this conjecture is false. The rationale follows that of the intuition described above for Theorem 1a: The subsidy cost for inducing any acquisition quantity is lower using the purchase subsidy than the sales subsidy. Because the sales subsidy “lever” is inferior to the purchase subsidy “lever,” the donor—when given the opportunity to use both levers—only chooses the purchase subsidy lever.

### 5 Heterogeneous Valuations

This section examines the case where consumers’ valuations are heterogeneous (demand is given by (1)). As in §4, although we are focused on the comparison between the two subsidy types, for compactness we present the retailer’s problem under a purchase and sales subsidy. We first examine the retailer’s pricing problem and then step back to his procurement problem. After having purchased \( Q \) units, upon observing market state \( M = m \), the retailer’s retail price setting problem is to choose his retail price \( p \in [0, \infty) \) to maximize his expected revenue

\[
(s + p) \min(D_e(m, p), Q).
\]

The retailer’s retail price setting problem can be written equivalently as the problem of setting the sales quantity \( q \in [0, \min(Q, m/b)] \) to maximize the expected revenue

\[
(s + m - bq)q.
\]

The optimal sales quantity is

\[
q^* = \min(Q, m/b, (s + m)/(2b)),\quad (21)
\]
which corresponds to an optimal retail price

$$p^* = m - \min(bQ, m, (s + m)/2).$$

(22)

Under a purchase and sales subsidy, the retailer’s procurement problem is to choose his acquisition quantity $Q \geq 0$ to maximize his expected profit

$$\pi(Q) = \begin{cases} 
\int_0^{bQ} sm/b \, dF(m) + \int_{bQ}^{m} (s + m - bQ)Q \, dF(m) - (c - a)Q & \text{if } Q < s/b \\
\int_0^{s} sm/b \, dF(m) + \int_{s}^{2bQ-s} (s + m)^2/(4b) \, dF(m) + \int_{2bQ-s}^{\infty} (s + m - bQ)Q \, dF(m) - (c - a)Q & \text{otherwise.}
\end{cases}$$

Let $z(a)$ denote the unique solution to

$$\int_z^m (m - z)dF(m) = c - a$$

(23)

if $a + \mu - c > 0$, and let $z(a) = 0$ otherwise. We abuse notation by letting $Q$, $\overline{Q}$, $Q_b$ and $Q_k^*$ for $k \in \{a, s\}$ have the same meaning in the heterogeneous consumer valuation case as they had in the homogeneous valuation case. For example, $Q_b$ denotes the retailer’s optimal acquisition quantity under a purchase and sales subsidy.

**Lemma 1** Under a purchase and sales subsidy, the retailer’s optimal acquisition quantity $Q_b$ is given by the following. If $a + s + \mu - c \leq 0$, then $Q_b = 0$. Suppose instead that $a + s + \mu - c > 0$. If $s \leq z(a)$, then $Q_b = [s + z(a)]/(2b)$; further, $Q_b \geq s/b$. If $s > z(a)$, then $Q_b$ is the unique solution to

$$\int_{bQ_b}^m (m + s - 2bQ_b)dF(m) = c - a;$$

(24)

further, $Q_b < s/b$.

We conclude this subsection by discussing two benchmarks: no subsidies and social welfare maximization. Under no subsidies ($a = s = 0$), the retailer’s optimal acquisition quantity is $Q = z(0)/(2b)$. Expected social welfare under acquisition quantity $Q$ is the sum of the utility derived by consumers, the retailer and the rest of society for each unit consumed less the cost of acquiring the $Q$ units. For any acquisition quantity $Q$ and any realized market condition $M = m$, social welfare is maximized by selling $\min(Q, m/b)$ units to the $\min(Q, m/b)$ consumers with the highest valuations. Consequently, expected social welfare under acquisition quantity $Q$ and the social welfare maximizing sales quantity is

$$w(Q) \equiv \int_0^{bQ} m^2/(2b)dF(m) + \int_{bQ}^{\infty} (m - bQ/2)QdF(m) + cE \min(M/b, Q) - cQ.$$
The quantity which maximizes expected social welfare, $\bar{Q}$, is the unique solution to
\[
\int_{b\bar{Q}}^{\bar{m}} (e + m - b\bar{Q})dF(m) = c
\]
if $e + \mu - c > 0$; otherwise, $\bar{Q} = 0$. Thus, when the pricing and acquisition quantity decisions are made to maximize expected social welfare, expected social welfare is $w(\bar{Q})$.

### 5.1 Purchase Subsidy

Under a purchase subsidy, the retailer’s procurement problem is to choose his acquisition quantity $Q$ to maximize his expected profit $\pi(Q)|_{s=0}$. The retailer’s optimal acquisition quantity under a purchase subsidy is given by the special case of Lemma 1 where $s = 0$. Under a purchase subsidy, the donor’s problem is to choose the per-unit purchase subsidy $\alpha \geq 0$ to maximize her expected utility

\[
u_a \equiv vE \min(M/(2b), Q) - aQ,
\]

where the retailer’s optimal acquisition quantity $Q = z(a)/(2b)$. Offering a more generous purchase subsidy (increasing $a$) has two effects. First, it increases the retailer’s acquisition quantity: $z(a)/(2b)$ is increasing in $a$. This first effect has a second effect: having a larger quantity encourages the retailer to price more aggressively (sell a larger quantity) when market conditions are strong. To be concrete, if a donor increases his purchase subsidy from $a'$ to some higher level, the retailer will respond by setting a lower price when the market condition is strong $m > z(a')$. The retailer’s pricing decision will be unaffected when the market condition is weak $m \leq z(a')$. In designing a purchase subsidy, the donor trades off the up-front cost of paying out the subsidy against the benefit, which only occurs when market conditions are strong, of the retailer selling a larger quantity.

Proposition 5 characterizes the donor’s optimal purchase subsidy $a^*$ and the resulting retailer’s acquisition quantity $Q^*_a$.

**Proposition 5** (a) If $c \in \left[0, \int_v^{\overline{m}} (m-v)dF(m)\right]$, then the optimal purchase subsidy $a^* = 0$ and the retailer’s optimal acquisition quantity $Q^*_a = \frac{Q}{2} \geq v/(2b)$. If $c \in \left(\int_v^{\overline{m}} (m-v)dF(m), \mu + v\right)$, then $Q^*_a$ is the unique solution to
\[
\int_{2bQ^*_a}^{\overline{m}} (v + m - 4bQ^*_a)dF(m) = c;
\]

$Q^*_a \in (\frac{Q}{2}, v/(2b))$; the optimal purchase subsidy is

\[
a^* = (v - 2bQ^*_a)\overline{F}(2bQ^*_a);
\]

and $a^* > 0$. If $c \geq \mu + v$, then $a^* = 0$ and $Q^*_a = 0$. (b) The optimal purchase subsidy $a^*$ is strictly increasing in the retailer’s acquisition cost on $c \in \left(\int_v^{\overline{m}} (m-v)dF(m), \mu + v\right)$ with $\lim_{c \to (\mu + v)^-} a^* = \cdots$
Proposition 5 provides insight into how the retailer’s acquisition cost $c$ impacts the donor’s optimal purchase subsidy $a^*$. When the acquisition cost is low $c \in [0, \int_v^m (m - v) dF(M)]$, the retailer purchases a large quantity without a subsidy and a strictly positive per-unit subsidy is costly to the donor because it applies to retailer’s large acquisition quantity; consequently, the donor offers no subsidy ($a^* = 0$). As the retailer’s acquisition cost increases through the range $c \in (\int_v^m (m - v) dF(m), \mu + v)$, this increasing cost discourages the retailer from purchasing units, and the donor responds by increasing the subsidy she offers. However, when the retailer’s acquisition cost crosses the threshold $c = \mu + v$, the donor goes from offering a very generous subsidy, $\lim_{c \to (\mu + v)-} a^* = v$, to offering no subsidy, $\lim_{c \to (\mu + v)+} a^* = 0$. At this cost threshold, it becomes too costly for the donor to offer a subsidy that will impact the retailer’s acquisition decision, so the donor “gives up” and offers no subsidy. (The result for the case with homogenous valuations is similar; see Proposition 2b.)

Expected social welfare under the optimal purchase subsidy is the sum of the utility captured by the retailer and consumers plus the value that consumption generates for the rest of society, less the cost of acquiring the units

$$w_a^* = \int_0^{2bQ_a^*} 3m^2/(8b) dF(m) + \int_{2bQ_a^*}^m (m - bQ_a^*/2)Q_a^*dF(m) + cE \min(M/(2b), Q_a^*) - cQ_a^*.$$ 

The utility generated by consumption for the retailer, consumers and the rest of society depends on both how aggressively the retailer prices ex post, which in turn depends on how many units the retailer acquired ex ante. Relative to the social welfare maximizing decisions, under the optimal purchase subsidy, the retailer acquires too few units ex ante (provided that the donor’s value is less than the externality $v \leq \epsilon$), and, given this quantity, does not price aggressively enough ex post.

### 5.2 Sales Subsidy

Under a purchase subsidy, the retailer’s procurement problem is to choose his acquisition quantity $Q$ to maximize his expected profit $\pi(Q)|_{a=0}$. The retailer’s optimal acquisition quantity under a sales subsidy is given by the special case of Lemma 1 where $a = 0$. To ease notation, for the remainder of the paper, we write $z(0)$ as simply $z$. Under a sales subsidy, the donor’s problem is to choose the per-unit sales subsidy $s \geq 0$ to maximize her expected utility

$$u_s \equiv \begin{cases} 
(v - s) \left( \int_0^s m/b \ dF(m) + \int_s^{2bQ-s} (s + m)/(2b) \ dF(m) + Q^* \mathcal{F}(2bQ - s) \right) & \text{if } s < z \\
(v - s)E \min(M/b, Q) & \text{otherwise},
\end{cases}$$

where $Q = (s + z)/(2b)$ if $s < z$, and where $Q$ is the unique solution to $\int_Q^m (m + s - 2bQ)dF(m) = c$ otherwise.
The sales subsidy is similar to the purchase subsidy in that both encourage the retailer to purchase a larger quantity and to price more aggressively when market conditions are strong. The subsidies differ in the degree to which they encourage more aggressive pricing. Recall that under a purchase subsidy, the retailer’s more aggressive pricing was solely due to the retailer’s having purchased a larger quantity. Under a sales subsidy, there is a second effect that encourages the retailer to price even more aggressively. Under a sales subsidy, the retailer receives not only the retail price but also the sales subsidy for each unit he sells, which makes it attractive for the retailer to cut his price further so as to increase the volume of units that are eligible for the subsidy. Consequently, for any acquisition quantity \( Q \) and any realized market condition \( M = m \), the retailer’s optimal retail price (see (22)) is lower under the sales subsidy than under the purchase subsidy, strictly so if and only if the market condition is weak \( m < \max(bQ, 2bQ - s) \).

To summarize, offering a more generous sales subsidy stochastically increases the retailer’s sales quantity not only by encouraging the retailer to purchase more ex ante, but also by rewarding the retailer for his sales volume ex post. In designing a sales subsidy, the donor trades off this benefit against the cost of the subsidy payment, which increases in the retailer’s sales quantity.

Proposition 6 characterizes the donor’s optimal sales subsidy \( s^* \) and the resulting retailer’s acquisition quantity \( Q_s^* \). Note \((v - (v - z)f(z)/2 < \mu + v \) if and only if the donor’s value is sufficiently large \( v > zF(z)/[3 - F(z)] \).

**Proposition 6** If \( c \in [0, \mu - v] \), then the optimal sales subsidy \( s^* = 0 \) and the retailer’s optimal acquisition quantity \( Q_s^* = Q \). If \( c \in (\mu - v, \min(\mu - (v - z)f(z)/2, \mu + v)) \), then \( s^* \) is the unique solution to
\[
(v - 2s^*)f(s^*) - \int_{0}^{s^*} mF(m) - \mu + c = 0; \tag{27}
\]
the retailer’s optimal acquisition quantity is
\[
Q_s^* = (s^* + z)/(2b);
\]
further, \( s^* < bQ_s^* < z \). If \( c \in [\mu - (v - z)f(z)/2, \mu + v] \), then \( Q_s^* \) is the unique solution to
\[
\int_{bQ_s^*}^{\infty} (m + v - 2bQ_s^*)dF(m) - c - E \min(M, bQ_s^*) \times \left[ 2 + \left( c - \int_{bQ_s^*}^{\infty} (m - bQ_s^*)dF(m) \right) f(bQ_s^*)/F(bQ_s^*) \right]^2 = 0; \tag{28}
\]
the optimal purchase subsidy is
\[
s^* = 2bQ_s^* + \left( c - \int_{bQ_s^*}^{\infty} mF(m) \right)/F(bQ_s^*); \tag{29}
\]
Further, \( s^* \geq bQ_s^* \geq z \). If \( c \geq \mu + v \), then \( s^* = 0 \) and \( Q_s^* = 0 \).
The result for the optimal sales subsidy parallels that for the optimal purchase subsidy: The donor offers a non-zero subsidy $s^* > 0$ only if the retailer’s acquisition cost is moderate.

Expected social welfare under the optimal sales subsidy is the sum of the utility captured by the retailer and consumers plus the value that consumption generates for the rest of society, less the cost of acquiring the units

\[
W_s^* \equiv \begin{cases} 
\int_0^{2bQ_s^*} \frac{m^2}{(2b)}dF(m) + \int_{2bQ_s^*}^z (3m - 2bQ_s^* + z)(m + 2bQ_s^* - z)/(8b)dF(m) \\
+ \int_{Q_s^*}^{m} (m - bQ_s^*/2)Q_s^*dF(m) + e[E \min(M, 2bQ_s^* - z) + \mu - c]/(2b) - cQ_s^* & \text{if } Q_s^* < z/b \\
0 \quad & \text{otherwise.}
\end{cases}
\]

Recall that under the optimal purchase subsidy, relative to the social welfare maximizing decision, the retailer does not price aggressively enough. The sales subsidy mitigates this problem by encouraging the retailer to price more aggressively. Indeed, when the acquisition cost $e \in [\mu - (v - z)F(z)/2, \mu + v)$ so that $Q_s^* \geq z/b$, the distortion in the retail pricing decision is completely eliminated: having purchased $Q$ units, upon observing market condition $M = m$ the retailer optimally prices so as to sell the social welfare maximizing quantity, $\min(m/b, Q)$. In this case, the only source of social welfare loss is the distortion in the acquisition quantity.

We compare expected social welfare, as well as donor utility, under the two subsidy types in the next subsection.

5.3 Comparison of Subsidies

Let $u^*_D$ ($u^*_a$, respectively) denote the donor’s expected utility under the optimal purchase (sales, respectively) subsidy.

**Theorem 2** (a) There exists $\bar{v} \in [z, \infty]$ such that if $v \leq z$ or $v > \bar{v}$, then the donor’s expected utility is higher under the optimal sales subsidy than under the optimal purchase subsidy

\[
u_s^* \geq u_a^*;
\]

where the inequality is strict if $v \in (\mu - c, z]$ or $v > \bar{v}$. (b) Suppose $e \in ((e\bar{m} - E[M^2]/4)/\mu, \infty)$. There exists $\bar{v} \in [z, \infty)$ and $\underline{v} \in (\mu - c, z]$ such that if $v \leq \underline{v}$ or $v > \bar{v}$, then expected social welfare is higher under the optimal sales subsidy than under the optimal purchase subsidy

\[
w_s^* \geq w_a^*;
\]

where the inequality is strict if $v \in (\mu - c, \underline{v})$ or $v > \bar{v}$.

The sales subsidy is superior to the purchase subsidy, from the perspective both of the donor and of society, provided that the donor’s value $v$ is sufficiently high or low. More precisely, if the
The donor's value is sufficiently large ($v > \tau$) or the donor's value and the production cost are low ($c < \mu$ and $v \in (\mu - c, \mu]$), then both the donor and society are strictly better off under the optimal sales subsidy rather than the optimal purchase subsidy. The social welfare result requires the caveat that the externality $\varepsilon$ be sufficiently large.

To assess whether Theorem 2's central conclusion— that the sales subsidy is superior to the purchase subsidy, from the perspective both of the donor and society—continues to hold when the Theorem's sufficient conditions are violated, we conducted a large scale numerical study. We considered the 1452 combinations of the following parameters: $\varepsilon \in \{0.25, 0.75, 1.25, 1.75, 2.25, 2.75, 3.25, 3.75, 4.25, 4.75, 5.25\}$, $v \in \{e/4, e/2, 3e/4, e\}$, $c \in \{0.05, 0.45, 0.85, 1.25, 1.65, 2.05, 2.45, 2.85, 3.25, 3.65, 4.05\}$, $b = 0.001$, $\mu = 1$, and $\sigma \in \{\mu/4, \mu/2, 3\mu/4\}$, where, as before, $M$ is a truncated Normal$(\mu, \sigma)$ random variable. In every instance, the donor's expected utility is higher under the sales subsidy, $u^*_s \geq u^*_a$, and expected social welfare is higher under the sales subsidy, $\omega^*_s \geq \omega^*_a$. Further, in all but 1.4% of the instances, the retailer’s optimal acquisition quantity is higher under the sales subsidy $Q^*_s \geq Q^*_a$; in every instance, the retailer’s expected sales quantity is higher under the sales subsidy. This suggests that the conclusions of Theorem 1 (expected donor utility, expected social welfare and expected sales are higher under the purchase subsidy), which apply when consumers’ valuations are homogeneous, are reversed when consumers’ valuations are heterogeneous.

When consumers’ valuations are homogeneous, the purchase subsidy is superior because it is more effective in influencing the retailer’s key decision, his acquisition quantity (see the discussion following Theorem 1). The retailer’s pricing decision is dictated by the consumers’ valuation, and it is not influenced by the donor’s subsidy. When consumers’ valuations are heterogeneous, the donor is interested in influencing not only the retailer’s purchase decision, but also the retailer’s pricing decision. The sales subsidy is superior to the purchase subsidy because it provides stronger incentives for the retailer to price aggressively: When the retailer makes her pricing decision, the purchase subsidy is sunk and so has a more limited impact on the retailer’s pricing decision; in contrast, the sales subsidy actively rewards the retailer for increasing her sales volume. This more aggressive pricing benefits not only the donor, but society as well.

Again we use the case of ACTs to illustrate our results. In some ACTs settings, it may be more accurate to model consumers’ valuations as being heterogeneous instead of homogeneous. Per the discussion in §3, this will tend to be true when the market served by a retailer is relatively large, diverse, and lacks the strong presence of socially-inferior competitor product, such as SP. For the heterogeneous case, we assume that $M$ is a Uniform$(\alpha, \beta)$ random variable. Based on Cohen et al. (2010), we obtain $\alpha = 0$, and $\beta = 4.9$, and $b = 0.0096$. We use the same cost and valuation parameters from §4.3. The optimal purchase subsidy $a^* = 0.466$ results in retailer acquisition quantity $Q^*_a = 79.3$, expected sales of 67.0, expected donor utility $U^*_a = 110$, and expected social
welfare $W_a^* = 961$. The optimal sales subsidy $s^* = 0.610$ results in higher expected sales, donor utility and social welfare: $Q_s^* = 78.8$, expected sales of 72.5, $U_s^* = 115$ and $W_s^* = 1040$. Although the donor’s utility is higher under the sales subsidy, the difference is not overwhelmingly large; it is possible that the difference could be offset by the higher administrative costs required by the sales subsidy. (Although these costs may be higher under a sales subsidy, evidence from the use of sales subsidies with ITNs indicates that the administrative costs are not prohibitive (Mulligan et al. 2008). In Tanzania, ITN subsidy administrators are testing mobile phones and text messaging as a mechanism for handling vouchers, which may further reduce these costs.)

When consumers have homogeneous valuations and when the donor has the opportunity to offer both subsidy types simultaneously, she never exercises this right: It is optimal to offer only a purchase subsidy and no sales subsidy: $a_b^* \geq 0$ and $s_b^* = 0$ (Proposition 4). This result breaks when consumers have heterogeneous valuations. Specifically, when the donor’s value is large, it is optimal to offer both subsidies $a_b^* > 0$ and $s_b^* > 0$. When consumers have homogeneous valuations, in offering a subsidy, the donor is influencing one decision, the retailer’s acquisition quantity. Because the purchase subsidy is strictly superior at influencing this decision (see the discussion before and after Proposition 4), it is optimal to only offer a purchase subsidy. In contrast, when consumers have heterogeneous valuations, the donor is influencing not only the retailer’s acquisition quantity decision, but also her pricing decision. Because the two subsidy types impact each of these decisions differently, the donor benefits by using both subsidy types.

6 Discussion

This paper provides guidance to donors designing subsidies to improve the availability of products in the private-sector distribution channel. We show that the donor’s choice of the kind of subsidy to employ depends crucially on whether consumers are homogeneous or heterogeneous in their valuation of the product. For the case where consumers’ valuations are heterogeneous, we provide evidence that a sales subsidy leads to greater expected donor utility, consumption, and social welfare. When consumers’ valuations are homogeneous, we establish the opposite conclusion: the purchase subsidy leads to greater expected donor utility, consumption, and social welfare. Intuitively, the purchase subsidy is more effective when the key decision in the distribution channel is the purchasing decision, as it is in the homogeneous case. The sales subsidy is more effective when the pricing (sales quantity) decision is a critical decision, as it is in the heterogeneous case.

We now discuss how our modeling assumptions impact how this conclusion should be interpreted. First, differential administrative costs to implement each type of subsidy, which we have not explicitly modeled, should be taken into account. Second, our model only captures the cases of perfectly homogeneous consumers and heterogeneous consumers whose valuations are uniformly
distributed. The sharply contrasting results for these two settings suggests that, in general, the donor’s choice of subsidy type may depend on the degree of heterogeneity among consumers: if consumers are relatively homogeneous, the purchase subsidy is likely to be superior; if consumers are quite heterogeneous, the sales subsidy may be superior.

Third, in our model of the donor’s subsidy design problem, product availability is captured by the stocking and pricing decisions of a single firm selling to end consumers. This setup informs the donor’s subsidy design decision in three settings. First, the setup is appropriate when the donor is able to offer distinct subsidies to different retailers. Second, the setup informs the donor’s decision when a particular type of retailer (e.g., a drug shop) is the primary means by which consumers access the product in the region where the subsidy is offered, and retailers of this type are relatively homogeneous. Third, the setup informs the donor’s decision when the distribution channel serving a particular geographic area exhibits considerable vertical integration, which can be the case in developing countries where weak contractual-enforcement mechanisms favor this structure (Jaffe and Yi 2007). In such cases, the firm in the model is interpreted as being the entire distribution channel.

References


scheme for insecticide-treated nets. *Malaria J.* 7(32)


Yadav, P., O. Stapleton, L. N. Van Wassenhove. 2010. If access is the symptom, what is the cause? Comparing medicine and consumer product supply chains in the developing world. Working Paper, INSEAD.

**Appendix**

Lemma 2 is useful in the proofs of Propositions 1 and 2. Let

\[ \Psi(Q) = v - \frac{c}{F(Q)} + r \left[ 1 - \frac{Qf(Q)}{F(Q)} \right]. \]  

(32)

Because \( F \) is IFR, \( \Psi(\cdot) \) is strictly decreasing. If \( v \leq c - r \), then let \( Q_a^c = 0 \). If \( v > c - r \), then let \( Q_a^c \) denote the unique solution to \( \Psi(Q_a^c) = 0 \); because \( \Psi(0) > 0 \) and \( \Psi(\overline{\mu}) < 0 \), \( Q_a^c \in (0, \overline{\mu}) \). It is useful to observe that

\[ (\partial/\partial Q)U_a(Q) = vF(Q) - c + r[F(Q) - Qf(Q)]. \]

**Lemma 2** The donor’s expected utility under a purchase subsidy \( U_a(Q) \) is unimodal on \( Q \in [0, \overline{\mu}] \), strictly increasing on \( Q \in [0, Q_a^c] \), strictly decreasing on \( Q \in (Q_a^c, \overline{\mu}] \), and strictly concave on \( Q \in [0, Q_a^c] \).
The proof is similar to that of Theorem 1 in Lariviere and Porteus (2001), and hence is omitted.

**Proof of Proposition 1:** From Lemma 2, if \( v \leq c - r \), then \( U_s(Q) \) is strictly decreasing on \( Q \in (0, \overline{m}) \). The quantity \( Q \in \overline{Q, m} \) which maximizes the donor’s expected utility \( U_s(Q) \) is \( Q_a^* = Q \), which corresponds to a purchase subsidy of \( a^* = 0 \).

For the remainder of the proof suppose that \( v > c - r \). Then, from Lemma 2, \( U_s(Q) \) is unimodal with an interior maximizer. Therefore, there exists a unique solution \( Q \in (0, \overline{m}) \) to

\[
v \overline{F}(Q) - c + r [\overline{F}(Q) - Qf(Q)] = 0.
\]

Observe that

\[
v > (r^2/c) f(Q) \tag{34}
\]

holds if and only if \( (\partial/\partial Q) U_s(Q) > 0 \). Therefore, if (34) is violated, then \( (\partial/\partial Q) U_s(Q) < 0 \) on \( Q \in \overline{Q, m} \) and the quantity \( Q \in [Q, \overline{m}] \) which maximizes the donor’s expected utility \( U_s(Q) \) is \( Q_a^* = Q \), which corresponds to a purchase subsidy of \( a^* = 0 \). If (34) holds, then the unique solution to (33) has \( Q > Q \) and this quantity maximizes \( U_s(Q) \). That is, the quantity that maximizes the donor’s expected utility \( Q_a^* \) is the unique solution to (11). The purchase subsidy \( a^* \) corresponding to the quantity \( Q_a^* \) is strictly positive because \( Q_a^* > Q \). Further, because

\[
v \overline{F}(Q_a^*) - c + r [\overline{F}(Q_a^*) - Q_a^* f(Q_a^*)] = 0
\]

if and only if (10) holds, the optimal subsidy \( a^* \) is the unique solution to (10).

The proof of Proposition 2 is in the internet appendix.

**Proof of Proposition 3:** Note that

\[
(\partial/\partial Q) U_s(Q) = (r + v) \overline{F}(Q) - c \left(1 + \left[f(Q)/\overline{F}(Q)\right]^2 \right) E \min(M, Q) \tag{35}
\]

Because \( F \) is IFR, (35) is strictly decreasing in \( Q \). If \( v \leq c - r \), then \( U_s(Q) \) is strictly decreasing on \( Q \in (0, \overline{m}) \). The quantity \( Q \in [Q, \overline{m}] \) which maximizes the donor’s expected utility \( U_s(Q) \) is \( Q_s^* = Q \), which corresponds to a sales subsidy of \( s^* = 0 \).

For the remainder of the proof suppose that \( v > c - r \). Because (35) is strictly decreasing in \( Q \), \( (\partial/\partial Q) U_s(0) > 0 \) and \( (\partial/\partial Q) U_s(\overline{m}) < 0 \), there exists a unique solution \( Q \) to

\[
(r + v) \overline{F}(Q) - c \left(1 + \left[f(Q)/\overline{F}(Q)\right]^2 \right) E \min(M, Q)) = 0, \tag{36}
\]

and further this solution \( Q \in (0, \overline{m}) \). Observe that

\[
v > (r^3/c^2) E \min(M, Q) f(Q) \tag{37}
\]

holds if and only if \( (\partial/\partial Q) U_s(Q) > 0 \). Therefore, if (37) is violated, then \( (\partial/\partial Q) U_s(Q) < 0 \) on \( Q \in (\overline{Q, m}) \) and the quantity \( Q \in [Q, \overline{m}] \) which maximizes the donor’s expected utility \( U_s(Q) \) is
\( Q_s^* = Q \), which corresponds to a sales subsidy of \( s^* = 0 \). If (37) holds, then the unique solution to (36) has \( Q > \underline{Q} \) and this quantity maximizes \( U_a(Q) \). That is, the quantity that maximizes the donor’s expected utility \( Q_s^* \) is the unique solution to (18). The purchase subsidy \( s^* \) corresponding to the quantity \( Q_s^* \) is strictly positive because \( Q_s^* > \underline{Q} \). Further, because

\[
(r + v)\overline{F}(Q_s^*) - c \left( 1 + \left[ f(Q_s^*)/\overline{F}(Q_s^*)^2 \right] E \min(M, Q_s^*) \right) \bigg|_{Q_s^*=\overline{F}^{-1}(c/(r+s^*))} = 0
\]

if and only if (17) holds, the optimal subsidy \( s^* \) is the unique solution to (17). ■

**Proof of Theorem 1:** If \( v \leq c - r \), then \( Q_a^* = Q_s^* = Q \) (from Propositions 1 and 3). This implies that \( U_a^* = U_s^* \) and \( W_a^* = W_s^* \). For the remainder of the proof suppose that \( v > c - r \). The structure of proof is first to establish (c), then (b) and then (a).

(c) Because \( (r^2/c)f(Q)Q < (r^3/c^2)E \min(M, Q)f(Q) \), if \( v \leq (r^2/c)f(Q)Q \) then, \( Q_a^* = Q_s^* = Q \) (from Propositions 1 and 3). If \( v \in ((r^2/c)f(Q)Q, (r^3/c^2)E \min(M, Q)f(Q)) \), then \( Q_a^* > Q = Q_s^* \). Finally, suppose that \( v > (r^3/c^2)E \min(M, Q)f(Q) \). Note that

\[
(\partial/\partial Q)U_a(Q_a^*) = (r + v)\overline{F}(Q_a^*) - c \left( 1 + \left[ f(Q_a^*)/\overline{F}(Q_a^*)^2 \right] E \min(M, Q_a^*) \right)
\]

\[
= f(Q_a^*)[rQ_a^* - cE \min(M, Q_a^*)/\overline{F}(Q_a^*)^2]
\]

\[
< \left[ rf(Q_a^*)E \min(M, Q_a^*)/\overline{F}(Q_a^*)^2 \right] \overline{F}(Q_a^*) - c/r
\]

\[
< 0 \tag{39}
\]

where (38) follows from (11), and (39) follows because \( Q_a^* > Q \). Because \( F \) is IFR, \( (\partial/\partial Q)U_a(Q) \) is strictly decreasing in \( Q \), which implies \( Q_a^* > Q_s^* \).

(b) First we establish that

\[
Q_a^* \leq \overline{Q} \tag{40}
\]

If \( v \leq (r^2/c)f(Q)Q \), then \( Q_a^* = Q \leq \overline{Q} \), where the equality follows from Proposition 1 and the inequality follows because \( e \geq 0 \). If \( v > (r^2/c)f(Q)Q \), then (40) follows from the facts that \( U_a(Q) \) is unimodal in \( Q \) (from Lemma 2) and maximized at \( Q = Q_a^* \) and

\[
(\partial/\partial Q)U_a(Q) = v\overline{F}(Q) - c + r[\overline{F}(Q) - \overline{Q}f(Q)]
\]

\[
= -c(e - v)/(e + r) - r\overline{Q}f(Q)
\]

\[
\leq 0,
\]

where the inequality follows because \( v \leq e \). Inequalities \( (r^2/c)f(Q)Q < v \) and \( v \leq e \) imply \( e > 0 \), which implies \( Q < \overline{Q} \). Expected social welfare \( W(Q) \) is strictly increasing in \( Q \) for \( Q \in [\underline{Q}, \overline{Q}] \), so the result follows from Part (c), \( Q \leq Q_s^* \), and (40).
(a) We will show that for \( Q \in [\underline{Q}, \overline{Q}] \),

\[
U_a(Q) \geq U_s(Q).
\]  

Inequality (41) is equivalent to

\[
0 \geq [r - c/F(Q)] \int_0^Q mdF(m),
\]

and this inequality holds because \( Q \in [\underline{Q}, \overline{Q}] \). Then

\[
U_a^* = U_a(Q_a^*) \geq U_a(Q_s^*) \geq U_s(Q_s^*) = U_s^*,
\]  

(42)

If \( v > (r^2/c)f(Q)Q \) holds, then \( Q_a^* > Q_s^* \) (from Part (c)) and because \( Q_a^* \) is unique (from Proposition 1), the first inequality in (42) is strict. \( \blacksquare \)

Proof of Proposition 4: From (4), \( a = c - (r + s)F(Q_b) \). For a given sales subsidy \( s \), there is a one-to-one mapping between the donor’s purchase subsidy \( a \) and the retailer’s acquisition quantity \( Q_b \). Consequently, we can rewrite the donor’s problem as one of choosing the quantity \( Q \in [\underline{Q}, \overline{Q}] \) and sales subsidy \( s \geq 0 \) to maximize

\[
(v - s)E\min(M, Q) - cQ + (r + s)F(Q)Q,
\]

which simplifies to

\[
vE\min(M, Q) - cQ + rF(Q)Q - s \int_0^Q mdF(m).
\]

Because this quantity is strictly decreasing in \( s \), the optimal sales subsidy is \( s_b^* = 0 \). \( \blacksquare \)

The proof of Lemma 1 is in the internet appendix.

Lemma 3 is useful in the proof of Proposition 5.

Lemma 3 The donor’s problem under a purchase subsidy \( \max_{a \geq 0} u_a \) can be written as \( \max_{Q \in [\underline{Q}, \overline{Q}]} u_a(Q) \), where

\[
u_a(Q) = \int_0^{(vQ)/(2b)} (vm)/(2b) dF(m) + \int_{2bQ}^{\overline{Q}} (v + m - 2bQ)dF(m) - cQ
\]  

(43)

is the donor’s expected utility as a function of the induced acquisition quantity \( Q \) that results from a given purchase subsidy.

The proof of Lemma 3 is in the internet appendix.

Proof of Proposition 5: (a) Because \( u_a < 0 \) for \( a > v \) and \( u_a \geq 0 \) for \( a \leq v \), the optimal purchase subsidy \( a^* \leq v \). Therefore, if \( c \geq \mu + v \), then \( a^* \leq v \leq c - \mu \). From Lemma 1, because the optimal purchase subsidy \( a^* \leq c - \mu \), the retailer’s optimal acquisition quantity \( Q_s^* = 0 \) and an optimal purchase subsidy is \( a^* = 0 \). In the remainder of the proof, we suppose that \( c < \mu + v \).
From Lemma 3, the donor’s problem under a purchase subsidy can be written as
\[
\max_{\Theta \in [\Theta, \mu]} (v + m - 4b\Theta) dF(m) - c.
\]
We will show that
\[
(\partial / \partial Q)u_a(Q) > 0 \text{ if and only if } c > \int_v^{\infty} (m - v) dF(m).
\] (44)

If \( c \geq \mu \), then \( Q = 0 \) and
\[
(\partial / \partial Q)u_a(Q) = v + \mu - c > 0.
\] (45)

If \( c < \mu \), then \( Q = z(0)/(2b) \) and
\[
(\partial / \partial Q)u_a(Q) = \int_{2bQ_a}^{\infty} (v + m - 4bQ_a) dF(m) - \int_{2bQ_a}^{\infty} (m - 2bQ_a) dF(m)
= (v - 2bQ_a^*) \bar{F}(2bQ_a),
\] (46)
which is strictly positive if and only if \( c > \int_v^{\infty} (m - v) dF(m) \). This establishes (44).

Because \( \lim_{Q \to \infty} (\partial / \partial Q)u_a(Q) < 0 \), the retailer’s optimal acquisition quantity under the optimal purchase subsidy, \( Q_a^* \), either is the boundary solution \( Q_a^* = Q \) or satisfies the first order condition \( (\partial / \partial Q)u_a(Q_a^*) = 0 \), which is stated explicitly as (25). We will now show that if \( Q_a^* \) satisfies (25), then \( Q_a^* \) is the unique solution to (25). Suppose that \( Q_a^* \) satisfies (25). Then, the optimal purchase subsidy is
\[
a^* = c - \int_{2bQ_a^*}^{\infty} (m - 2bQ_a^*) dF(m)
= \int_{2bQ_a^*}^{\infty} (v + m - 4bQ_a^*) dF(m) - \int_{2bQ_a^*}^{\infty} (m - 2bQ_a^*) dF(m),
= (v - 2bQ_a^*) \bar{F}(2bQ_a^*),
\] (47)
Because \( a^* \geq 0 \), (47) implies
\[
Q_a^* \leq v/(2b),
\]
where the inequality is strict if \( a^* > 0 \). Further,
\[
(\partial^2 / \partial Q^2)u_a(Q) = -2b(v - 2bQ) f(2bQ) - 4b \bar{F}(2bQ),
\]
so \( u_a(Q) \) is strictly concave on \( Q \in [0, v/(2b)] \), and so the retailer’s optimal acquisition quantity under the optimal purchase subsidy \( Q_a^* \) is the unique solution to (25). In sum, one of the following holds: \( Q_a^* = Q \) and \( a^* = 0 \); or \( Q_a^* \) is the unique solution to (25), \( a^* \) is given by (26) and \( a^* > 0 \).

Case 1: \( c \in \left[0, \int_v^{\infty} (m - v) dF(m)\right] \). First observe that \( c \leq \int_v^{\infty} (m - v) dF(m) \) implies \( Q \geq v/(2b) \).
Suppose that \( a^* > 0 \). Then \( c \leq \mu \) implies that \( a^* > c - \mu \) and \( Q^a_\mu > Q \). However, because \( Q^a_\mu \) satisfies (25), \( Q^a_\mu \leq v/(2b) \), which contradicts \( Q^a_\mu > Q \geq v/(2b) \). We conclude that \( a^* = 0 \), which implies that \( Q^a_\mu = Q \).

Case 2: \( c \in \left( \int_v^\infty (m - v) dF(m), \mu + v \right) \). From (44), \((\partial/\partial Q)u_s(Q) > 0\), which implies \( Q^a_\mu > Q \). Therefore, \( Q^a_\mu \) is the unique solution to (25), \( Q^a_\mu \in (Q, v/(2b)) \), \( a^* \) is given by (26) and \( a^* > 0 \).

(b) From part (a), because \( c \in \left( \int_v^\infty (m - v) dF(m), \mu + v \right) \), \( Q^a_\mu \) is the unique solution to (25) and \( a^* \) is given by (26). From (25), \( Q^a_\mu \) is strictly decreasing in \( c \). From (26), \( a^* \) is strictly decreasing in \( Q^a_\mu \). Therefore, \( a^* \) is strictly increasing in \( c \). Because \( \lim_{c \to (\mu + v)^-} Q^a_\mu = 0 \), (26) implies \( \lim_{c \to (\mu + v)^-} a^* = v \).

Lemma 4 is useful in the proof of Proposition 6.

**Lemma 4** The donor’s problem under a purchase subsidy \( \max_{s \geq 0} u_s \) can be written as \( \max_{Q \in [Q, \min(\max(v, z), \overline{\mu})/b]} u_s(Q) \), where

\[
u_s(Q) = \begin{cases} 
\frac{(v - 2bQ + z)/(2b)}{v - 2bQ - \left(c - \int_{Q}^{\overline{\mu}} m dF(m) \right)/F(bQ)} & \text{if } Q < z/b \\
E \min(M, 2bQ - z) + \mu - c & \text{otherwise}
\end{cases}
\]

is the donor’s expected utility as a function of the induced acquisition quantity \( Q \) that results from a given sales subsidy. Further, \( u_s(Q) \) is strictly concave on \( Q \in [Q, \min(\max(v, z), \overline{\mu})/b] \); \( s < z \) if and only if the induced acquisition quantity \( Q < z/b \). If the induced acquisition quantity \( Q < z/b \), then the sales subsidy \( s \) that induces this acquisition quantity is

\[
s = 2bQ - z; 
\]

otherwise, the sales subsidy is

\[
s = 2bQ + \left(c - \int_{Q}^{\overline{\mu}} m dF(m) \right)/F(bQ). 
\]

The proof of Lemma 4 is in the internet appendix.

**Proof of Proposition 6:** From Lemma 4, the donor’s problem is to choose \( Q \in [Q, \min(\max(v, z), \overline{\mu})/b] \) to maximize \( u_s(Q) \).

Case 1: \( c \in [0, \mu - v] \). Then \((\partial/\partial Q)u_s(Q) = v - \mu + c \leq 0 \). Because \( u_s(Q) \) is strictly concave, this implies that \( u_s(Q) \) is decreasing in \( Q \) on \( Q \geq Q \). Therefore, the optimal acquisition quantity \( Q^s_s = Q \), which corresponds to an optimal sales subsidy of \( s^* = 0 \).

Case 2: \( c \geq \mu + v \). Now \( c > \mu \) implies \( z = 0 \) and \( Q = 0 \). Therefore, \( s^* \geq z \) and \((\partial/\partial Q)u_s(Q) = v + \mu - c \leq 0 \). By the same argument in Case 1, \( Q^s_s = Q \) and \( s^* = 0 \).

Case 3: \( c \in (\mu - v, \min(\mu - (v - z)F(z)/2, \mu + v)) \). Then \((\partial/\partial Q)u_s(Q) = v - \mu + c > 0 \). Further, because \( c < \mu - (v - z)F(z)/2 \), from (??), \((\partial/\partial Q)u_s(z/b) < 0 \). So \( Q^s_s < z/b \) and \( Q^s_s \) is the unique
solution to the first order condition

\[
(\partial/\partial Q)u_s(Q_s^*) = (v - 4bQ_s^* + 2z)\overline{F}(2bQ_s^* - z) - \int_0^{2bQ_s^* - z} mdF(m) - \mu + c = 0. \tag{51}
\]

Because \(Q_s^* < z/b\), the optimal sales subsidy \(s^* < z\) (from Lemma 4). From Lemma 1, \(Q_s^* = (s^* + z)/(2b)\). Therefore, the second equality in (51) can be rewritten as (27) and \(s^*\) is the unique solution to (27). Further, \(s^* < z\) implies \(s^* < bQ_s^* < z\), where the second inequality follows from Lemma 4.

Case 4: \(c \in [\mu - (v - z)\overline{F}(z)/2, \mu + v]\). Because \(c \geq \mu - (v - z)\overline{F}(z)/2\), from (29), \((\partial/\partial Q)u_s(z/b) \geq 0\). So \(Q_s^* \geq z/b\) and \(Q_s^*\) is the unique solution to the first order condition \((\partial/\partial Q)u_s(Q_s^*) = 0\), which is equivalent to (28). Equation (29) follows from (50). It remains to show that \(s^* \geq bQ_s^*\). From (29), it is sufficient to show that

\[
c - \int_{bQ_s^*}^{\overline{m}} (m - bQ_s^*)dF(m) \geq 0. \tag{52}
\]

If \(c \geq \mu\), then (52) follows. If \(c < \mu\), then \(z\) satisfies (23) where \(a = 0\). This, along with the fact that \(bQ_s^* \geq z\), implies (52).

**Proof of Theorem 2:** (a) First, we establish the results for \(v \leq z\). Suppose \(c < \mu\). Then \(c = \int_{z}^{\overline{m}} (m - z)\overline{F}(m) \leq \int_{v}^{\overline{m}} (m - v)\overline{F}(m)\), where the inequality holds because \(v \leq z\). Thus, from Proposition 5, if \(v \leq z\), then \(a^* = 0\), which implies (30). From Proposition 6, if \(v > \mu - c\), then \(s^* > 0\) and so for \(v \in (\mu - c, z]\), the inequality in (30) is strict. Suppose instead that \(c \geq \mu\). Then

\[
z = 0. \text{ If } v = 0, \text{ then by Propositions 5 and 6, } a^* = s^* = 0, \text{ which implies } u_s^* = u_a^*. \]

Second, we establish the results for \(v > \nu\). We begin by establishing a technical result that will be useful subsequently: if the donor’s valuation is sufficiently large \(v \geq 4z + c/\overline{F}(2z)\), then

\[
Q_s^* \geq z/b. \tag{53}
\]

If \(c \geq \mu\), then \(z = 0\), so (53) holds. Suppose instead that \(c < \mu\). If \(v > z\), then \(c = \int_{z}^{\overline{m}} (m - z)\overline{F}(m) \geq \int_{v}^{\overline{m}} (m - v)\overline{F}(m)\). Therefore, from Proposition 5, \(Q_a^*\) is the unique solution to (25). If \(v \geq 4z + c/\overline{F}(2z)\), then \(\int_{2z}^{\overline{m}} (v + m - 4z)dF(m) \geq c\), and it is straightforward to show that this implies (53).

Case 1: \(\overline{m} < \infty\). We will establish that (30) holds with strict inequality when \(\nu = \max(4z + c/\overline{F}(2z), 2[\overline{m} + c/\overline{F}(\overline{m}/2)])\). To do so, we will establish that

\[
u_a(Q) > u_a(Q) \text{ for } Q \in [z/b, \min(v, \overline{m})/(2b)] \text{ and } v > 2[\overline{m} + c/\overline{F}(\overline{m}/2)]. \tag{54}
\]
For $Q \in [z/b, \min(v, \overline{m})/(2b))$ and $v > 2[\overline{m} + c/F(\overline{m}/2)]$,
\[
 u_s(Q) - u_a(Q) = \frac{1}{b} \left[ \int_0^{bQ} mf(m) \left( \frac{v}{2} - 2bQ - c - \int_{bQ}^{\overline{m}} mf(m) \right) / F(bQ) \right] \\
+ (v/2 - bQ) \int_{bQ}^{2bQ} (2bQ - m) dF(m) \tag{55} \quad \text{(because $Q \geq z/b$)}
\]
\[
> \frac{1}{b} \int_0^{bQ} mf(m) [v/2 - 2bQ - c/F(bQ)] \tag{56} \quad \text{(because $Q < v/2b$)}
\]
\[
> \frac{1}{b} \int_0^{bQ} mf(m) [v/2 - \overline{m} - c/F(\overline{m}/2)] \tag{57} \quad \text{(because $Q < \overline{m}/2b$)}
\]
\[
> 0. \tag{58} \quad \text{(because $v > 2[\overline{m} + c/F(\overline{m}/2)]$)}
\]

If $v > \overline{m}$, then $Q_a^* \in [z/b, \min(v, \overline{m})/(2b))$ and (54) implies
\[
u_s^* = u_s(Q_s^*) \geq u_s(Q_a^*) > u_a(Q_a^*) = u_a^*.
\]

Case 2: $\overline{m} = \infty$. We will establish that (30) holds with strict inequality when $v$ is sufficiently large. To do so, we will establish that
\[
u_s(Q_a^*) > u_a(Q_a^*) \quad \text{for $v > \max(4z + c/F(2z), \mu - c, \bar{v})$}
\]
for some finite $\bar{v}$. Suppose that $v > \max(4z + c/F(2z), \mu - c)$. Because $Q_a^*$ satisfies (25),
\[
c - \int_{bQ_a^*}^{\overline{m}} mf(m) = (v - 4bQ_a^*)F(2bQ_a^*) - \int_{bQ_a^*}^{2bQ_a^*} mf(m).
\]

Because $Q_a^* \in [z/b, v/(2b))$, (55) holds when $Q = Q_a^*$. Using this and (58) yields
\[
u_s(Q_a^*) - u_a(Q_a^*) = \frac{1}{b} \left[ \int_0^{bQ_a^*} mf(m) \left( \frac{v}{2} - 2bQ_a^* - \left( v - 4bQ_a^* \right) F(2bQ_a^*) - \int_{bQ_a^*}^{2bQ_a^*} mf(m) \right) / F(bQ_a^*) \right] \\
+ (v/2 - bQ_a^*) \int_{bQ_a^*}^{2bQ_a^*} (2bQ_a^* - m)dF(m).
\]

Because $Q_a^*$ satisfies (25), $\lim_{v \to \infty} Q_a^* = v/(4b)$. Therefore,
\[
\lim_{v \to \infty} \left[ u_s(Q_a^*) - u_a(Q_a^*) \right] = \lim_{v \to \infty} \frac{1}{b} \left[ \int_0^{v/4} mf(m) \left( \int_{v/4}^{v/2} mf(m) / F(v/4) \right) \right] \\
+ (v/4) \int_{v/4}^{v/2} (v/2 - m)dF(m) > 0.
\]

Because $Q_a^*$ is continuous in $v$, and $u_s(Q)$ and $u_a(Q)$ are continuous in $Q$, we conclude that there exists $\bar{v} < \infty$ such that (57) holds. Inequality (57) implies that for $v > \max(4z + c/F(2z), \mu - c, \bar{v})$, (56) holds.
(b) First, we establish the results for \( v \leq \underline{v} \). Let
\[
\Gamma(Q) = \int_0^{2bQ-z} m^2/(2b)dF(m) + \int_{2bQ-z}^{z} (3m - 2bQ + z)(m + 2bQ - z)/(8b)dF(m) \\
+ \int_{z}^{\overline{m}} (m - bQ/2)QdF(m) + e [E \min(M, 2bQ - z) + \mu - c]/(2b - c).
\]
Suppose \( c < \mu \). Then because \( z \) is the unique solution to \( \int_{z}^{\overline{m}} (m - z)dF(m) - c = 0 \), \( \int_{z}^{\overline{m}} (m - x)dF(m) \) is decreasing in \( x \), and \( \int_{\mu-c}^{\overline{m}} (m - \mu + c)dF(m) - c = \int_0^{\mu-c}(\mu - c - m)dF(m) > 0 \), \( \mu - c < z \). If \( v \leq z \), then \( a^* = 0 \) (as established in Part (a)), which implies
\[
w_a^* = \Gamma(Q);
\]
if \( v \leq z \), then \( c \leq \int_{v}^{\overline{m}} (m - v)dF(m) < \min(\mu - (v - z)F(z)/2, \mu + v) \), so \( Q^*_s < z/b \) (by Proposition 6), which implies
\[
w_s^* = \Gamma(Q^*_s).
\]
Note that
\[
(\partial/\partial Q)\Gamma(Q) = \int_0^{z} (m/2)dF(m) + \int_{z}^{\overline{m}} (m - z/2)dF(m) + e - c \\
= \int_0^{z} (m/2)dF(m) + (z/2)F(z) + e > 0,
\]
where the second equality follows because \( c = \int_{z}^{\overline{m}} (m - z)dF(m) \). Because \( \Gamma(Q) \) is continuous in \( Q \) and because \( Q^*_s \) is is continuous and strictly increasing in \( v \) on \( v \in (\mu - c, z] \) (by Proposition 6), there exists \( \underline{v} \in (\mu - c, z] \) such that if \( v \in (\mu - c, \underline{v}] \),
\[
\Gamma(Q^*_s) > \Gamma(Q).
\]
This, along with (59) and (60), implies that for \( v \in (\mu - c, \underline{v}] \), (31) holds with strict inequality. If \( v \leq \mu - c \), then \( a^* = s^* = 0 \), and (31) holds with equality. Suppose instead that \( c \geq \mu \). Then \( z = 0 \).
If \( v = 0 \), then by Propositions 5 and 6, \( a^* = s^* = 0 \), and (31) holds with equality.
Second, we establish the results for \( v > \overline{v} \). Because \( Q^*_s \) satisfies (25) when \( v > c - \mu \), \( Q^*_s \) satisfies (28) when \( v \geq 2(\mu - c)/F(z) + z \), and \( \overline{m} < \infty \), \( \lim_{v \to \infty} Q^*_a = \overline{m}/(2b) \) and \( \lim_{v \to \infty} Q^*_s = \overline{m}/b \). Therefore,
\[
\lim_{v \to \infty} [w^*_a - w^*_a] = (E[M^2]/4 + e\mu - \overline{m} - E[M^2]/4)/b > 0,
\]
where the inequality follows because \( e > (\overline{m} - E[M^2]/4)/\mu \). Because \( Q^*_a \) and \( Q^*_s \) are continuous in \( v \), \( w^*_a \) is continuous in \( Q^*_a \), and \( w^*_s \) is continuous in \( Q^*_s \), we conclude that there exists \( \overline{v} < \infty \) such that (31) holds with strict inequality if \( v > \overline{v} \).
Paper Title: Combating Strategic Counterfeiters in Licit and Illicit Supply Chains

Author Information:

1. Soo-Haeng Cho
   - Email: soohaeng@andrew.cmu.edu
   - Affiliation: Assistant Professor of Operations Management, Tepper School of Business, Carnegie Mellon University
   - Address: 5000 Forbes Avenue, Pittsburgh, PA 15213
   - Telephone: 412-268-5744
2. Xin Fang
   - Email: xfang@andrew.cmu.edu
   - Affiliation: Doctoral Student, Tepper School of Business, Carnegie Mellon University
3. Sridhar Tayur
   - Email: STayur@smartops.com
   - Affiliation: Ford Distinguished Research Chair, Professor of Operations Management, Tepper School of Business, Carnegie Mellon University

Presenter: Soo-Haeng Cho
Combating Strategic Counterfeiters in Licit and Illicit Supply Chains

Abstract: Counterfeit products are being produced and consumed in all economies. While clothing and fashion accessories dominated counterfeit supply in the past, today’s counterfeit goods are in a wide range of simple to sophisticated products including items that have an impact on personal health and safety. When a counterfeiter sells his goods as low-price substitutes for brand-name products (i.e., non-deceptive counterfeit), he tends to distribute the goods through illicit supply chains. Conversely, a counterfeiter must infiltrate licit supply chains of brand-name products when he intends to deceive consumers into believing they purchase genuine goods (i.e., deceptive counterfeits). Using a normative model of counterfeiting, we show how a counterfeiter’s decisions of his price, quality, and distribution channel depend on product characteristics, counterfeit costs and risks. We then analyze the effectiveness of operations, marketing, and enforcement strategies to combat strategic counterfeiters. Our analysis highlights that the effectiveness of these strategies depends critically on whether a brand-name company faces non-deceptive, deceptive, or both types of counterfeits. Although it is ideal that strategies which are effective against counterfeiting also benefit consumers, this is not always the case. Therefore, policy-makers should identify the characteristics of counterfeiters and devise anti-counterfeiting strategies accordingly, while minimizing negative consequences to consumers.

Key words: intellectual property, illegal operations, supply chain management
1 Introduction

Intellectual properties such as trademarks, copyrights and patents represent the most valuable assets of many firms. They require significant investment in research and development as well as years of efforts in maintaining high product quality and careful brand management. Famous global brands such as GE, Nike and Nestlé are popular because they offer a guarantee of quality, which is vital to consumers when they make purchasing decisions. For those goods for which the mere display of a particular brand confers prestige on their owners, such as luxury watches and fashion apparel, many consumers purchase branded goods to demonstrate to others that they are consumers of the particular good. These intrinsic values of trademarks or brands create incentives for counterfeiting.

According to the Agreement on Trade-Related Aspects of Intellectual Property Rights (WTO 1994), counterfeit goods mean “any goods, including packaging, bearing without authorization a trademark which is identical to the trademark validly registered in respect of such goods, or which cannot be distinguished in its essential aspects from such a trademark, and which thereby infringes the rights of the owner of the trademark in question under the law of the country of importation.” Since a counterfeit bears the trademark of a branded product, a consumer enjoys the brand image even when he/she purchases the counterfeit.

Counterfeiting business was once called “the world’s fastest growing and most profitable business” (O’Donnell 1985). Nowadays counterfeits have developed into a substantial threat to many industries. A recent OECD study estimates that international trade in counterfeit and pirated goods could amount to up to $250 billion or 1.95% of world trade in 2007, up from $105 billion in 2001 (OECD 2009). If including domestically produced and consumed products or non-tangible pirated digital products, the total magnitude could be several hundred billion dollars more (OECD 2008). The problem is no longer limited to prestigious, easy-to-manufacture products such as designer clothing, branded sportswear, and fashion accessories. It affects nearly all product categories including items that have an impact on personal health and safety such as pharmaceuticals, food, drink, toys, tobacco, medical equipment, and automotive parts (OECD 2008).

Counterfeits are broadly categorized into two types: non-deceptive and deceptive (Grossman and Shapiro 1988a). A non-deceptive counterfeit is the counterfeit a consumer can distinguish from the brand-name product at time of purchase. This type of counterfeits tends to be sold at a substantial discount through an unauthorized sales channel. For example, consumers can easily tell that $1 DVDs and $10 luxury watches sold by street vendors are counterfeit. On the contrary, a deceptive counterfeit is the counterfeit a consumer believes to be authentic at time of purchase even if it is, in fact, counterfeit. In order to deceive consumers, this type of counterfeit goods has
to infiltrate licit supply chains; for example, fake auto parts were found in legitimate repair shops, counterfeit pharmaceutical products at chemists, food products on supermarket shelves (OECD 2008), and pirated software products sold by one of the largest resellers (Bass 2010). Solomon (2009) notes that counterfeit drugs make their way through the licit supply chain via a distributor who moves a product from a low-cost channel to a high-cost channel. Collusion between counterfeiters and licit supply chain members occurs due to a higher profit from selling counterfeits (Green and Smith 2002, Bass 2010). A deceptive counterfeit is usually sold at the same price as its branded product. Although it appears to function properly at time of purchase, it lacks product warranty and often involves health and safety risks of consumers.

In order to stop or at least to reduce the incidence of counterfeits, brand-name companies are spending millions of dollars. They hire full-time employees, invest in new technologies, and redesign their products to make counterfeiting more difficult (Balfour 2005). However, the anti-counterfeiting strategies found to be useful to one product may not work for another or can even unintentionally make counterfeits flourish more in the market. For example, Chinese shoe manufacturers successfully addressed their counterfeiting issues by improving the quality of their products (Qian 2008). This is the outcome of the competition in which high-quality authentic products defeat low-quality non-deceptive counterfeits. However, the same strategy backfired against a Scotch whisky company in the Thailand market (Green and Smith 2002). At the peak of the company’s sales in 1988, 42% of its premium Scotch whisky sales was counterfeit; high quality made the products more popular, increasing aggregate demand, and attracted more counterfeits. In this case, the counterfeits were sold as the genuine products and commanding the same price, i.e., sold as deceptive counterfeits. After the initial attempt to fight counterfeits by improving quality had failed, the company eventually succeeded in radically reducing the incidence of counterfeiting by establishing a system that monitors supply chains: the company focused on identifying members in its supply chain who were selling the counterfeits, facilitating seizure of the counterfeits and punishing counterfeiters.

These contrasting results illustrate a need for anti-counterfeiting strategies that are tailored to specific products. Yet, due to the limited understanding of relations among the characteristics of products, the types of counterfeits, and the effectiveness of anti-counterfeiting strategies, OECD (2008) calls for research that strengthens the analysis of counterfeiting and says:

“Asessing the factors driving production and consumption of counterfeit and pirated products can generate insights into the types of products that are most likely to be infringed, . . . , and lead to more efficient and effective [anti-counterfeiting] strategies.”

This paper attempts to provide such an analysis to help industry and governments identify the types of counterfeiters they may face and design their effective anti-counterfeiting strategies accordingly.
Specifically, we aim to provide insights to the following questions: (Q1) How do the characteristics of a brand-name product such as the importance of brand value in the overall quality of a product and the complexity of a product affect the type of its counterfeits? (Q2) How does the effectiveness of anti-counterfeiting strategies differ across the different types of counterfeits? How can we deter the entry of counterfeitors? (Q3) What is the impact of counterfeits on consumer welfare? Do consumers also benefit from the strategies that are effective in combating counterfeits?

To answer these questions, we develop a normative model of licit and illicit supply chains in which a brand-name company competes with her potential counterfeiters. We consider three possible scenarios in which a brand-name company faces: (Scenario 1) only a non-deceptive counterfeiter, (Scenario 2) only a deceptive counterfeiter, or (Scenario 3) both non-deceptive and deceptive counterfeiters. In each scenario, either type of a counterfeiter decides the functional quality and wholesale price of his counterfeit goods. Depending on the type, the counterfeiter faces different opportunities and risks. The non-deceptive counterfeiter competes directly with a brand-name company for price and quality. Thus the counterfeiter may have to invest in improving the quality of his goods, which will increase the risk of losing the investment in case of getting caught by the authorities. Conversely, the deceptive counterfeiter may not need to invest as much in improving the quality as non-deceptive counterfeits (as long as he can deceive consumers successfully at time of purchase), but he has to infiltrate a licit supply chain via a legitimate distributor who sources both brand-name and counterfeit products. The legitimate distributor then faces a trade-off between a greater profit margin and a risk of getting punished for selling counterfeits. In addition to analyzing three separate scenarios, we analyze a counterfeiter’s decision of choosing between the two types to answer our first research question posed above. After finding optimal decisions of the counterfeiter, we evaluate the following anti-counterfeiting strategies of which the effectiveness depends on the subsequent reaction of the strategic counterfeiter: (i) operations strategies that improve the quality of brand-name products or enhance the level of technological complexity (in order to make it harder for a counterfeiter to market his products or to copy the original products, respectively), (ii) marketing strategies that reduce the price of brand-name products or educate consumers about the adversity of counterfeits, and (iii) enforcement strategies that increase the chances to seize the production of counterfeits or increase a penalty on a licit distributor who sells deceptive counterfeits illegally. We conduct a comprehensive analysis by evaluating each strategy in terms of its impacts on the expected market share and profit of the brand-name company, the expected market share and profit of the counterfeiters, and consumer welfare.

The rest of the paper is organized as follows. In Section 2 we review the related literature and discuss our contributions. In Section 3 we present our model. In Section 4 we find the optimal
decisions of the counterfeiter under each scenario. Section 5 addresses our first research question regarding the effects of product characteristics on counterfeit types. In Section 6 we analyze the impacts of anti-counterfeiting strategies on the firms’ performance to answer our second research question. We provide a welfare analysis in Section 7 to answer our third research question. We conclude this paper in Section 8. All proofs and various model extensions are provided in Appendix.

2 Literature Review

Traditional supply chain management research is focused on licit supply chains in which members of supply chains interact with each other by exchanging goods and services legally. In this era of globalization, supply chains are no longer confined within one country as more and more companies offshore and outsource their operations to less developed countries. However, this has created a frightening phenomenon: an ever-rising flood of counterfeit items coming into markets (Business Week 2005). This paper is intended to shed light on counterfeit problems in both licit and illicit supply chains and to analyze the effectiveness of anti-counterfeiting strategies.

The majority of studies on counterfeits are conceptual and descriptive. They provide frameworks for fighting counterfeiting usually based on case studies. Staake and Fleisch (2008) provide an extensive review of this literature. Marketing researchers have conducted empirical studies on counterfeits. They mainly focus on the demand side of counterfeits and try to answer questions such as ‘why consumers purchase counterfeits?’ and ‘how to educate consumers not to purchase counterfeits?’ Eisend and Schuchert-Guler (2006) review this literature and conclude that further investigation is needed to develop a general framework that integrates the existing results consistently. Recently, using data from Chinese shoe companies, Qian (2008) finds that brand-name companies tend to improve their product quality after the entry of non-deceptive counterfeitors. Our result shows that such anti-counterfeiting strategy is effective in combating non-deceptive counterfeitors but may not necessarily benefit consumers.

There are only a handful of analytical studies that present prescriptive models of counterfeits. Grossman and Shapiro (1988a, 1988b) develop an equilibrium model of trades between brand-name firms in a home country and low-quality producers in a foreign country. To sell their goods as counterfeits in the home market, foreign producers must pass the goods through the home-country border, hence facing the risk of confiscation. Grossman and Shapiro (1988a) analyze the consequences of deceptive counterfeits in a market where consumers cannot observe the quality of a product, and provide a welfare analysis of border inspection policy. They assume that counterfeiters are capable of selling deceptive counterfeits directly to consumers. Grossman and Shapiro (1988b)
present a Cournot competition model between brand-name products and non-deceptive counterfeits given their exogenous quality levels. Because non-deceptive counterfeits can contribute positively to consumer welfare due to their lower price, the authors conclude that policies that discourage foreign counterfeiting need not improve welfare, which is consistent with our finding. Scandizzo (2001) views competition between brand-name firms and non-deceptive counterfeitors as a patent race over time. The author finds that counterfeits improve consumer welfare while reducing firms’ profits, and that the more skewed the income distribution within the economy is towards the poor, the greater the welfare effect and the smaller the profit effect.

There have been growing interests in counterfeit research among operations researchers. Liu et al. (2005) study the decision of an inventory manager who can source both genuine and deceptive counterfeit products and sell them to consumers at one price. Using a multi-item newsvendor model, they analyze the effectiveness of a monitoring and limiting regime on the cheating activity of the manager. Sun et al. (2010) study a global firm’s decision of outsourcing the production of its components to a foreign country. The firm faces a trade-off between lower labor cost and increased risk of imitation by a foreign firm. The foreign firm incurs a fixed cost of imitation which increases with the complexity of a component. The authors find the optimal strategy in choosing the range of components to transfer. Zhang et al. (2010) analyze the case when a brand-name firm faces non-deceptive counterfeits. They show that a non-deceptive counterfeit lowers the price and profit of the brand-name product, and a brand-name firm has more incentive to improve her own quality rather than reducing that of a counterfeit. They also analyze a situation in which two brand-name products complete, which we do not consider in this paper.

We draw on and contribute to this stream of research by addressing the following important issues in counterfeiting problems:

1. Strategic counterfeiters: The common assumption used in the literature is that the quality as well as the type of a counterfeit (deceptive or non-deceptive) is fixed a priori. For example, Grossman and Shapiro (1988a, 1988b) assume that foreign producers always choose the lowest quality because they lack capital, resource, and technology for quality improvement and that there are no entry costs of counterfeiters. Today, thanks to outsourcing and offshoring of numerous global firms, counterfeiters benefit greatly from increasingly easy access to modern production facilities (Staake and Fleisch 2008). Schmidle (2010) note that today’s counterfeiters come in varying levels of quality depending on their intended markets, and diversify their products and distribution channels to manage the risks involved in their criminal activities. In our model, a counterfeiter decides the type, functional quality, and wholesale price of his products by considering a trade-off between the benefit from stealing brand value and the risk of confiscation. Our analysis
shows that the effectiveness of anti-counterfeiting strategies depends critically on the strategic response of a counterfeiter to those strategies.

(2) Licit and illicit supply chains: The previous analytical papers assume that a counterfeiter is capable of selling his counterfeits directly to consumers regardless of his type (non-deceptive or deceptive). Although this is quite possible for non-deceptive counterfeits, a deceptive counterfeiter has to infiltrate a licit supply chain in order to deceive consumers into believing they are purchasing authentic products: today very few consumers would believe the counterfeits sold by street vendors or unknown websites are authentic. In contrast to Grossman and Shapiro (1988a), we make a more realistic assumption that a deceptive counterfeiter sells his counterfeits through a licit distributor who sources both brand-name and counterfeit products.

(3) Product characteristics: Although clothing and fashion accessories dominated counterfeit supply in the past, today’s counterfeit goods are in a wide range of simple to sophisticated products. Our model captures two major product characteristics that differ among product categories. First, we decompose the quality of a product into functional quality and brand value, and analyze how the importance of brand value in a product drives different results. Second, we take technological complexity into account in modeling a counterfeiter’s decisions. The more complex the technology of a brand-name product is, the higher investment is required for a counterfeiter to imitate the product.

(4) Evaluation of anti-counterfeiting strategies: We evaluate the six anti-counterfeiting strategies used in practice mentioned earlier by examining their impacts on consumer welfare as well as the expected market share and profit of a brand-name company and her potential counterfeiters under the three different scenarios. Our comprehensive analysis complements the previous findings (discussed above) of Grossman and Shapiro (1988a, 1988b) and Zhang et al. (2010).

Lastly, we note that a research question similar to counterfeiting arises in the literature of parallel importing (or gray market) and software piracy. Parallel importing is the practice of purchasing products in a lower-priced region and shipping them illegally to a higher priced region (e.g., Ahmadi and Yang 2000). Both counterfeits and parallel imports deal with intra-brand competition and illicit supply chain issues. However, the products distributed through parallel importing differ fundamentally from counterfeits in that the parallel imported goods are authentic but sold at a lower price, whereas counterfeits are not authentic, possess lower quality, and are sold at a lower price for non-deceptive counterfeits or at the same price for deceptive counterfeits. Piracy differs from counterfeiting in that piracy refers to infringement of copyright, whereas counterfeits include any goods bearing a trademark without authorization (WTO 1994). However, companies often protect their products under either of the intellectual property rights as they share several common
characteristics. In our model, software piracy can be viewed as a special case of counterfeiting, in which counterfeit products have almost the same functional quality as authentic ones but their cost of development and production is very low.

3 Model

We consider a market served by a brand-name company (‘she’) and her potential counterfeiter(s) (‘he’). We use subscript $i = B$ to denote the brand-name company, $i = N$ to denote the non-deceptive counterfeiter, and $i = D$ to denote the deceptive counterfeiter (when the same symbol is used for all products). A consumer in this market purchases at most one unit of a product. In making a purchasing decision of product $i$, a consumer considers his/her utility $u_i = \theta \phi_i - p_i$, where $\theta$ represents his/her taste, $\phi_i$ represents the quality of the product a consumer perceives at time of purchase, and $p_i$ represents the retail price of the product. All consumers prefer high quality for a given price but a consumer with a higher $\theta$ is more willing to pay to obtain a high-quality product. We assume that $\theta$ is uniformly distributed over $[0, 1]$ and the size of the market is one. A consumer purchases a product only if the utility from purchasing the product is nonnegative in which case he/she selects a product that provides the highest utility. This is the standard vertical (quality) differentiation model (e.g., page 296 in Tirole (1988)) which is also used by Qian (2008) and Zhang et al. (2010). We next provide the detailed components of our model that capture the unique aspects of counterfeiting.

Depending on the counterfeit type, the quality of product $i$ a consumer perceives at time of purchase, $\phi_i$, may differ from its real quality $q_i$. For the non-deceptive counterfeit as well as the brand-name product, consumers know what product they are purchasing, so the perceived quality of either product is the same as its real quality, i.e., $\phi_B = q_B$ and $\phi_N = q_N$. However, for the deceptive counterfeit, consumers cannot distinguish it from the brand-name product and perceive both products as the same, hence $\phi_D = q_B$. This assumption is reasonable as long as the market share of the deceptive counterfeit is relatively small because then consumers either do not know the existence of counterfeits or think the likelihood of purchasing counterfeits is negligible (Staake and Fleisch 2008). However, when consumers’ expectations about such likelihood are not negligible, they may take those expectations into account when purchasing products. The analysis of this case is available from authors upon request. In the remainder of this paper, unless mentioned specifically as the perceived quality, the quality refers to the real quality.

Following the literature, we assume that the quality of the brand-name product is superior to that of the counterfeit, i.e., $q_B > q_N$ and $q_B > q_D$. Since both non-deceptive and deceptive
counterfeits bear the trademark of the brand-name product, a consumer enjoys the brand image even when he/she purchases the counterfeit. Thus we may represent the quality of the counterfeit as \( q_i = f_i + \beta q_B \) \((i = N \text{ or } D)\), where \( f_i > 0 \) is the functional quality of the counterfeit \( i \) and \( \beta q_B \) \((\text{where } \beta > 0)\) is the brand value that the counterfeit steals from the brand-name product\(^1\).

The parameter \( \beta \) captures the following two factors. First, \( \beta \) captures a fraction of the brand value in the quality of the brand-name product, \( q_B \). For example, this fraction may be high for luxury goods because a brand plays a significant role when consumers purchase such products, whereas it may be low for fast moving consumer goods (which are sold quickly at relatively low cost) because a brand is less of a concern to consumers for such goods. Second, \( \beta \) captures a discount factor of the original brand value for the counterfeit because the counterfeit draws only a part of the brand value from the brand-name product. If both products function equally well and sell at the same price, everyone would prefer the brand-name product to the counterfeit. Since both non-deceptive and deceptive counterfeits bear the trademark of the same product, we assume \( \beta \) is the same for both types of counterfeits. The marginal costs of counterfeits are assumed constant and normalized to zero.

We consider three possible scenarios in which a brand-name company faces: (Scenario 1) only a non-deceptive counterfeiter, (Scenario 2) only a deceptive counterfeiter, or (Scenario 3) both non-deceptive and deceptive counterfeiters. We observe in practice that counterfeiters always enter a market following a brand-name company, often after a brand-name product becomes popular. Thus we assume that counterfeiters observe the price \( p_B \) and quality \( q_B \) of the brand-name product before making their decisions. In all scenarios, either type of counterfeiter \( i (= N \text{ or } D) \) makes two decisions to maximize his expected profit \( \pi_i \): a functional quality \( f_i \) and a wholesale price \( w_i \) to a distributor. However, different types of counterfeiters use different distribution channels to sell their goods. The non-deceptive counterfeiter \( (i = N) \) distributes his goods through an illicit distributor, who decides the retail price of the non-deceptive counterfeit to consumers, \( p_N \). On the other hand, the deceptive counterfeiter \( (i = D) \) has to break into a licit supply chain by distributing his goods through a licit distributor, who then sells both brand-name products and deceptive counterfeits to consumers at the same price \( p_B \). In this case, the licit distributor determines the proportion \( s \in [0, 1] \) of the deceptive counterfeit among all products he sells to consumers. Figure 1 illustrates different supply chain structures in the three scenarios. We present the details of our model in each scenario next.

In Scenario 1, the brand-name company competes with the non-deceptive counterfeiter. Con-

\(^1\)Essentially, we assume that a product has two attributes: functionality and brand value as in the multi-attribute model in marketing (e.g., see Lilien et al. 1992).
Consumers choose between two products having the same brand name but different qualities and prices. The competition between the non-deceptive counterfeiter and the brand-name company is then analogous to a duopoly in a vertically differentiated market. The non-deceptive counterfeiter and the illicit distributor make their decisions in three sequential stages as follows. In stage 1, the non-deceptive counterfeiter chooses the functional quality $f_N \in [f, \bar{f}]$ and makes an initial investment to develop and produce goods having $f_N$. The upper bound $\bar{f} (>0)$ may represent the functional quality of the brand-name product because the brand-name company usually has more resources and more advanced technology than the counterfeiter. We assume $\bar{f} < (1-\beta)q_B$ such that $q_B > q_N$. The lower bound $f (\geq 0)$ may represent the minimum level of quality at which a product functions or appears to function properly. To produce counterfeits having $f_N$, the counterfeiter needs to invest $t \cdot f_N$ in acquiring technology and setting up production facilities. The parameter $t (>0)$ represents the level of technological complexity of a product. This reflects the fact that more investment is required to make products with greater quality and complexity. After the investment takes place, however, there are some chances that the investment will be confiscated because it is illegal to produce counterfeits. Suppose this occurs with a probability $\gamma \in (0,1)$. The parameter $\gamma$ characterizes the monitoring efforts of the government and the brand-name company on counterfeit production. The potential loss of the investment is a major risk to the counterfeiter that deters him from making large investments to improve the functional quality of his products (OECD 2008). If the counterfeiter’s investment is confiscated, the counterfeiter cannot sell his goods to the market. Otherwise, the game proceeds to stage 2 in which the non-deceptive counterfeiter decides...
his wholesale price $w_N$. Since the non-deceptive counterfeiter distributes his products through the illicit supply chain, he sets his wholesale price $w_N$ to the illicit distributor. For simplicity, we group all distributors/retailers in the illicit supply chain into a single entity. In stage 3, the illicit distributor decides the retail price of the non-deceptive counterfeit to consumers, $p_N$. The illicit distributor has to pay a penalty of $l_N$ if getting caught by the authorities with probability $\alpha_N$.

In Scenario 2, the brand-name company confronts the deceptive counterfeiter who distributes his goods through the licit distributor. In this scenario, consumers perceive both products in the market identical. As in Scenario 1, the deceptive counterfeiter first determines his functional quality $f_D \in [\underline{f}, \overline{f}]$ and then his wholesale price $w_D$, while facing the risk of getting his investment on $f_D$ confiscated. However, in contrast to Scenario 1, the deceptive counterfeiter distributes his goods to the licit distributor, who then sells the counterfeits to consumers at the same price $p_B$ as the brand-name products. The licit distributor decides how many units to source from the deceptive counterfeiter at the wholesale price $w_D$ and from the brand-name company at the wholesale price $w_B$. We model the risk of the licit distributor selling deceptive counterfeits as the product of a likelihood of getting caught $\alpha_D$ and a penalty $l_D$. Since this likelihood increases as the licit distributor sells more deceptive counterfeits, we assume $\alpha_D$ is equal to the fraction of deceptive counterfeits, $s$. Because of this risk, we later show that a strategic counterfeiter chooses a lower price $w_D$ than $w_B$. Therefore, in deciding the fraction of deceptive counterfeits $s$, the licit distributor faces a trade-off between a greater profit margin and a risk of getting punished from selling counterfeits. For ease of computation, we make the following two assumptions. First, we assume the licit distributor does not make a profit from selling brand-name products, i.e., $w_B = p_B$, while it makes a positive profit from selling deceptive counterfeits. Relaxing this assumption does not alter the structural properties of our results. Second, we normalize $l_N = 0$ while having $l_D = l > 0$. In practice, a loss of an illicit distributor from potential seizure is much smaller than that of a licit distributor. Illicit distributors are usually street vendors or internet sites. Since their potential loss from seizure is small, they tend to close their stores temporarily when they get caught and then reopen the same stores or open new ones later. At Shanghai in China, for example, vendors in the Xiang Yang market, which were once famous for its high-quality counterfeits but closed due to the government’s massive campaigns in 2006, relocated to the Yantai Xinyang market that is now famous among tourists (Naumann 2009). In contrast, the punishment on the licit distributor for illegal distribution of deceptive counterfeits is very severe. For example, the Chinese court

\[^2\text{Alternatively, we may set } \alpha_D = se \text{ where } e \text{ represents the monitoring efforts of industry and governments on the illegal distribution of deceptive counterfeits. We can show that the distributor sells less counterfeits as } e \text{ increases. The subsequent analysis remains unchanged.}\]
sentenced the distributor of fake pills to 17 years in prison, the nation’s longest term for the crime (Bennett 2010) and the U.S. court sentenced a distributor who sold counterfeit networking cards to the military to 51 months in prison, the maximum term recommended by federal prosecutors (McKinley 2010). In our model, as the penalty \( l \) increases, the licit distributor will be more reluctant to take the risk of selling deceptive counterfeits; hence, for the counterfeiter, \( l \) represents the level of difficulty in infiltrating the licit supply chain.

In Scenario 3, the brand-name company faces both types of counterfeiters in the market. We consider a setting where the illicit and licit supply chains operate as in Scenarios 1 and 2, respectively (see Figure 1). However, the non-deceptive counterfeiter and the deceptive counterfeiter make their decisions simultaneously by taking the other party’s decision into account; and similarly, the illicit distributor and the licit distributor make their decisions simultaneously. We assume that the confiscation of non-deceptive counterfeits is independent of that of deceptive counterfeits, and that whether or not the confiscation of non-deceptive or deceptive counterfeits has occurred is public information. We can easily show that either counterfeiter makes the same decisions even if the confiscation of the other party’s counterfeits is not known to her.

4 Equilibrium Analysis

For each of the three scenarios, we analyze the model using backward induction to derive subgame-perfect Nash equilibrium. In stage 3, we find the optimal decision of the illicit (or licit) distributor in the illicit (or licit respectively) supply chain. In stage 2, we compute the optimal wholesale price of the non-deceptive (or deceptive) counterfeiter who distributes his goods through the illicit (or licit respectively) distributor. Lastly, in stage 1, we find the optimal functional quality of the non-deceptive (or deceptive) counterfeiter. We use superscript \((k)\) to denote the equilibrium under Scenario \(k\) (= 1, 2 or 3).

4.1 Scenario 1: Non-Deceptive Counterfeits

Suppose non-deceptive counterfeits are present in the market. Consumers are aware that the quality of the non-deceptive counterfeit \(q_N\) is lower than that of the brand-name product \(q_B\). There are three segments of consumers: (i) consumers who value the quality of a product highly and purchase the brand-name product, (ii) consumers who value the quality less and purchase the counterfeiter, and (iii) consumers who value the quality the least and do not purchase any product. The consumer who is indifferent between purchasing the brand-name product and the counterfeit has the following
taste $\tilde{\theta}$:

$$\tilde{\theta} q_N - p_N = \tilde{\theta} q_B - p_B \implies \tilde{\theta} = \frac{p_B - p_N}{q_B - q_N} = \frac{p_B - p_N}{(1 - \beta)q_B - f_N}.$$  (1)

The consumer who is indifferent between purchasing the counterfeit and none has the following taste $\hat{\theta}$:

$$\hat{\theta} q_N - p_N = 0 \implies \hat{\theta} = \frac{p_N}{q_N} = \frac{p_N}{f_N + \beta q_B}.$$  (2)

Let $m_i (\in [0, 1])$ denote the market share of product $i (= B$ or $N)$ when non-deceptive counterfeits are available in the market, and let $m_0$ represent the proportion of consumers who do not purchase any product, so that $m_B + m_N + m_0 = 1$. From (1) and (2), we obtain

$$m_B = 1 - \tilde{\theta} = 1 - \frac{p_B - p_N}{(1 - \beta)q_B - f_N}$$
and
$$m_N = \hat{\theta} - \hat{\theta} = \frac{p_B - p_N}{(1 - \beta)q_B - f_N} - \frac{p_N}{f_N + \beta q_B}.$$  (3)

Note that $m_B$ and $m_N$ in (3) are the market shares of two products, provided that non-deceptive counterfeits are available to consumers without being confiscated. Let $Em_i$ denote the expected market share of product $i$ that takes into account the likelihood that counterfeits do not reach the market.

$$Em_B = (1 - \gamma)m_B + \gamma \left( 1 - \frac{p_B}{q_B} \right)$$
and
$$Em_N = (1 - \gamma)m_N.$$  (4)

We also use the notation of $m_i$ and $Em_i$ for $i = B$ or $D$ in Scenario 2 and for $i = B$, $N$ or $D$ in Scenario 3.

In stage 3, given the wholesale price $w_N$ and the functional quality $f_N$ of the non-deceptive counterfeit, the illicit distributor determines the retail price to consumers, $p_N$, by solving the following problem:

$$\max_{p_N} (p_N - w_N)m_N = (p_N - w_N) \left\{ \frac{p_B - p_N}{(1 - \beta)q_B - f_N} - \frac{p_N}{f_N + \beta q_B} \right\}.$$  (5)

By noting that the profit of the illicit distributor in (5) is concave in $p_N$, one can easily obtain her optimal retail price $p_N^{(1)} (w_N, f_N) = \frac{(\beta q_B + f_N)p_B + q_Bw_N}{2q_B}$.

In stage 2, given the functional quality $f_N$, the non-deceptive counterfeiter determines his wholesale price $w_N$. By anticipating the best response of the illicit distributor, the non-deceptive counterfeiter chooses his optimal wholesale price that maximizes his expected profit given by:

$$\pi_N (w_N, f_N) = (1 - \gamma) \left\{ w_N \left( \frac{p_B - p_N}{q_B - q_N} - \frac{p_N}{q_N} \right) - tf_N \right\} - \gamma tf_N.$$  (6)

In (6), $(1 - \gamma)$ represents the likelihood that the counterfeiter is able to sell his goods without getting caught and the next term in the bracket represents the profit of the counterfeiter in that case. Note that it is possible that the confiscation occurs after the counterfeiter sells some units. In
that case, \((1 - \gamma)\) can be interpreted as the fraction of sales the counterfeiter has generated before he gets caught. In (6), the initial investment is considered a sunk cost but it may have a residual value. Also, the counterfeiter may have to pay a fine when he gets caught. In Appendix, we discuss various forms of penalties for counterfeiting. Since \(\pi_N\) is concave in \(w_N\) given \(f_N\), we can easily obtain the optimal wholesale price \(w_N^{(1)}\) and the corresponding expected profit of the non-deceptive counterfeiter \(\pi_N^{(1)}\), respectively, as follows:

\[
w_N^{(1)}(f_N) = \frac{p_B(f_N + \beta q_B)}{2q_B} \quad \text{and} \quad \pi_N^{(1)}(f_N) = \frac{p_B^2(1 - \gamma)(f_N + \beta q_B)}{8q_B(1 - \beta q_B - f_N)} - tf_N.
\]

By substituting \(w_N^{(1)}\) into \(p_N^{(1)}\), one can verify the following well-known results: (i) \(p_N^{(1)} < p_B\): the illicit distributor charges a lower price than that of the brand-name product because \(q_N < q_B\); and (ii) \(w_N^{(1)}(f_N) = 2\{p_N^{(1)}(f_N) - w_N^{(1)}(f_N)\}\) for any \(f_N\): the classic double marginalization result.

In stage 1, the non-deceptive counterfeiter decides the functional quality \(f_N\) by considering his optimal wholesale price in stage 2 and the best response of the illicit distributor in stage 3. The counterfeiter solves \(\max_{f_N \in [\underline{f}, \overline{f}]} \pi_N^{(1)}(f_N)\). The following lemma shows that the non-deceptive counterfeiter will choose either the upper bound \(\overline{f}\) or the lower bound \(\underline{f}\) of the functional quality.

**Lemma 1** In Scenario 1, the optimal functional quality of non-deceptive counterfeits \(f_N^{(1)}\) is \(\overline{f}\) if \(\pi_N^{(1)}(\underline{f}) \geq \pi_N^{(1)}(\overline{f})\); and otherwise, \(f_N^{(1)}\) is \(\underline{f}\).

The non-deceptive counterfeiter needs to consider the following trade-off in deciding his functional quality. A higher level of functional quality will draw more consumers but require more investment, which increases a potential loss from seizure. We show in the proof that the marginal benefit of functional quality is increasing in \(f_N\), while the marginal cost is constant. Therefore, it is optimal for the non-deceptive counterfeiter to choose the highest level of quality if the benefit is larger than the risk at any feasible \(f_N\) or to choose the lowest level otherwise. We discuss practical examples that match these results in Section 5.

### 4.2 Scenario 2: Deceptive Counterfeits

Suppose deceptive counterfeits are present in the market. Both brand-name products and deceptive counterfeits are sold at price \(p_B\) and their quality perceived by consumers is \(q_B\). By following the analysis similar to (1) and (2) in Scenario 1, we obtain the market share of the brand-name product and that of the deceptive counterfeit respectively as follows:

\[
m_B = (1 - s) \left(1 - \frac{p_B}{q_B}\right) \quad \text{and} \quad m_D = s \left(1 - \frac{p_B}{q_B}\right).
\]
In (8), \(1 - \frac{p_B}{q_B}\) represents the number of consumers who perceive the quality of products as \(q_B\) and purchase the products for price \(p_B\). Among those consumers, \(s\) percents receive deceptive counterfeits unknowingly. Considering the likelihood \(\gamma\) that deceptive counterfeits will be confiscated, we compute the expected market shares of the two products as follows:

\[
Em_B = \{(1 - \gamma)(1 - s) + \gamma\} \left(1 - \frac{p_B}{q_B}\right) \quad \text{and} \quad Em_D = (1 - \gamma)s \left(1 - \frac{p_B}{q_B}\right),
\]

(9)

In stage 3, the licit distributor solves the following problem to determine \(s\):

\[
\max_s s(1 - s)(p_B - w_D) \left(1 - \frac{p_B}{q_B}\right) - sl.
\]

(10)

In (10), \((1 - s)\) represents the likelihood that the distributor will not be detected for selling counterfeits and \(s(p_B - w_D) \left(1 - \frac{p_B}{q_B}\right)\) represents the distributor’s profit in that case. Recall that the distributor’s profit margin from selling brand-name products is assumed zero. The term ‘\(- sl\)’ in (10) represents the expected loss from potential seizure. From the concavity of the distributor’s expected profit, we can determine the optimal \(s_D^{(2)}(w_D, f_D) = \frac{1}{2} - \frac{2(p_B - w_D)(q_B - p_B)}{2(q_B - p_B)}\). This suggests that the fraction of counterfeits the distributor sells in equilibrium is always less than 50%.

In stage 2, the deceptive counterfeiter decides his wholesale price \(w_D\) that maximizes his expected profit given by:

\[
\pi_D(w_D, f_D) = (1 - \gamma) \left[w_D \left\{\frac{1}{2} - \frac{lq_B}{2(p_B - w_D)(q_B - p_B)}\right\} \left(1 - \frac{p_B}{q_B}\right) - tf_D\right] - \gamma tf_D.
\]

(11)

By solving \(\max_{w_D} \pi_D(w_D, f_D)\) for fixed \(f_D\), we obtain the optimal price \(w_D^{(2)}\) and the corresponding expected profit of the deceptive counterfeiter \(\pi_D^{(2)}\) as follows:

\[
w_D^{(2)}(f_D) = p_B - \sqrt{\frac{l p_B}{1 - \frac{p_B}{q_B}}} \quad \text{and} \quad \pi_D^{(2)}(f_D) = \frac{1}{2}(1 - \gamma) \left(\sqrt{p_B \left(1 - \frac{p_B}{q_B}\right) - tf_D}\right) - tf_D.
\]

(12)

Note from (12) that \(\pi_D^{(2)}(f_D)\) increases with \(p_B \left(1 - \frac{p_B}{q_B}\right)\), the revenue of the brand-name company without counterfeits. This is because the deceptive counterfeit gets a free ride on the brand name of the genuine product. As the risk of the licit distributor selling counterfeits increases with \(l\), the deceptive counterfeiter has to reduce his price \(w_D^{(2)}\) and as a result his expected profit \(\pi_D^{(2)}\) decreases.

In stage 1, the deceptive counterfeiter decides the level of the functional quality \(f_D\) by solving \(\max_{f_D \in [f, \hat{f}]} \pi_D^{(2)}(f_D)\). By observing that \(\frac{\partial \pi_D^{(2)}}{\partial f_D} = -t < 0\), we obtain the following result:

**Lemma 2** In Scenario 2, the deceptive counterfeiter will always choose the lower bound \(\hat{f}\) of the functional quality; i.e., \(f_D^{(2)} = \hat{f}\).
This result is intuitive. Since consumers cannot tell the difference between the deceptive counterfeit and the brand-name product, there is no incentive for the deceptive counterfeiter to invest in improving the functional quality beyond the minimum level $f$ at which he can deceive consumers. This result is consistent with the observation of the practice: the functional quality of deceptive counterfeits such as fake drugs and fake mechanical parts is usually so low that they often threaten the health and safety of consumers (Staake and Fleisch 2008).

4.3 Scenario 3: Both Non-Deceptive and Deceptive Counterfeits

Suppose both types of counterfeits coexist with brand-name products in the market. From our analysis of Scenarios 1 and 2, we can easily obtain the market shares of each product as follows:

$$m_B = (1 - s) m_{B+D}, \quad m_D = s m_{B+D}, \quad \text{and} \quad m_N = \frac{p_B - p_N}{(1 - \beta)q_B - f_N} - \frac{p_N}{f_N + \beta q_B},$$

where we define $m_{B+D} \equiv 1 - \frac{p_B - p_N}{(1 - \beta)q_B - f_N}$. Note that $m_N$ in (13) is the same as $m_N$ in Scenario 1 because consumers cannot distinguish deceptive counterfeits from brand-name products. Therefore, non-deceptive counterfeits compete with brand-name products as if deceptive counterfeits did not exist in the market. Hence, in equilibrium, the non-deceptive counterfeiter and the illicit distributor make the same decisions as in Scenario 1, i.e., $p_N^{(3)} = p_N^{(1)}$, $w_N^{(3)} = w_N^{(1)}$, and $f_N^{(3)} = f_N^{(1)}$. We next investigate how the coexistence of non-deceptive counterfeits affects the decisions of the deceptive counterfeiter and the licit distributor.

In stage 3, the licit distributor can be in one of the following two states: (i) non-deceptive counterfeits have been confiscated, and (ii) non-deceptive counterfeits have not been confiscated. (Recall that stage 3 of the licit distributor is reached only when deceptive counterfeits are available for sales without being confiscated.) The likelihood that the distributor is in state (i) is $\gamma$ and that of state (ii) is $1 - \gamma$. In state (i), the licit distributor makes the same decision as in Scenario 2 because non-deceptive counterfeits are not brought into the market. In state (ii), by considering the market share of each product in (13), the licit distributor determines the proportion between deceptive counterfeits and brand-name products. One can show the optimal $s^{(3,i)}$ in state (i) is $\frac{1}{2} - \frac{q_B}{2(p_B - w_D)(q_B - p_B)}$ ($= s^{(2)}$) and $s^{(3,ii)}$ in state (ii) is $\frac{1}{2} - \frac{1}{2(p_B - w_D)m_{B+D}}$. Similarly to Scenarios 1 and 2, we can obtain the expected market shares of the three products $Em_N$, $Em_D$ and $Em_B$ in Scenario 3.

In stage 2, the deceptive counterfeiter chooses his wholesale price to maximize his expected profit given by

$$\pi_D(w_D, f_D) = (1 - \gamma) \left[ w_D \left\{ \gamma s^{(3,i)} \left( 1 - \frac{p_B}{q_B} \right) + (1 - \gamma) s^{(3,ii)} m_{B+D} \right\} - tf_D \right] - \gamma tf_D.$$

(14)
By conducting the analysis similar to Scenario 2, we obtain the optimal wholesale price \( w_D^{(3)}(f_D) \) and the corresponding expected profit of the deceptive counterfeiter \( \pi_D^{(3)}(f_D) \) as follows:

\[
 w_D^{(3)}(f_D) = p_B - \frac{tp_B}{\sqrt{(1-\gamma)m_D + \gamma(1 - \frac{p_B}{q_B})}} \quad \text{and} \quad (15)
\]

\[
 \pi_D^{(3)}(f_D) = \frac{1}{2}(1 - \gamma)\left\{ \sqrt{p_B \left\{ (1-\gamma)m_D + \gamma(1 - \frac{p_B}{q_B}) \right\} - \sqrt{I}} \right\}^2 - tf_D. \quad (16)
\]

The following lemma compares the equilibrium outcomes under Scenario 3 with those under Scenario 2.

**Lemma 3** In equilibrium, \( f_D^{(3)} = f_D^{(2)} = f \), \( w_D^{(3)} < w_D^{(2)} \), \( s^{(3)} \equiv \gamma s^{(3, i)} + (1-\gamma)s^{(3, ii)} < s^{(2)} \), \( E_m^{(3)} < E_m^{(2)} \), and \( E_m^{(3)} < E_m^{(2)} \).

In both Scenarios 2 and 3, the deceptive counterfeiter finds it optimal to choose the lowest functional quality because consumers cannot tell the difference between the deceptive counterfeit and the brand-name product. However, as we discussed earlier, the presence of non-deceptive counterfeits reduces the aggregate demand \( m_B + D \) in Scenario 3. This induces the deceptive counterfeiter to charge a lower wholesale price \( w_D^{(3)} \) in Scenario 3 than in Scenario 2. Despite the lower wholesale price \( w_D^{(3)} \), the licit distributor procures a smaller proportion \( s^{(3)} \) of deceptive counterfeits on average, due to the lower aggregate demand in Scenario 3. Thus the expected market share of deceptive counterfeits, \( E_m^{(3)} \), is lower in Scenario 3. The brand-name company also suffers from the lower aggregate demand in Scenario 3 but benefits from the lower \( s^{(3)} \) chosen by the licit distributor. Our analysis shows that the former effect dominates the latter, resulting in \( E_m^{(3)} < E_m^{(2)} \).

## 5 Effects of Product Characteristics on Counterfeit Types

In this section we examine how the characteristics of a brand-name product affect the type of its counterfeits. To answer this question, we assume that a counterfeiter can choose whether to sell his products as non-deceptive counterfeits through an illicit distributor as in Scenario 1 or as deceptive counterfeits through a licit distributor as in Scenario 2. In the following we shall represent the counterfeiter’s optimal decision in the threshold form of \( \beta \) that represents the importance of brand value in a product. At the end of this section, we shall show the effect of \( \beta \) is also similar in Scenario 3 when both types of counterfeits may coexist in the market.

The following theorem shows that three categories of counterfeits emerge in equilibrium; see Figure 2 for illustration.
Figure 2: Three Categories of Counterfeits in Equilibrium when: (a) $\beta^{*} < \beta^{**}$ and (b) $\beta^{*} \geq \beta^{**}$.

**Theorem 1.** There exist three positive real numbers $\beta^{*}$, $\beta^{**}$ and $\beta^{***}$ (of which the closed-form expressions are presented in the proof) that lead to the following results:

(i) Suppose $\beta^{*} < \beta^{**}$. Then it is optimal for the counterfeiter to choose deceptive counterfeits having functional quality $\tilde{f}$ when $\beta \leq \beta^{*}$, to choose non-deceptive counterfeits having functional quality $\tilde{f}$ when $\beta^{*} < \beta \leq \beta^{**}$, and to choose non-deceptive counterfeits having functional quality $\tilde{f}$ when $\beta > \beta^{**}$.

(ii) Suppose $\beta^{*} \geq \beta^{**}$. Then it is optimal for the counterfeiter to choose deceptive counterfeits having functional quality $\tilde{f}$ when $\beta \leq \beta^{***}$, and to choose non-deceptive counterfeits having functional quality $\tilde{f}$ when $\beta > \beta^{***}$.

The first category of counterfeits appears when $\beta \leq \beta^{*}$ in case (i) or $\beta \leq \beta^{***}$ in case (ii). Counterfeits in this category are deceptive and have a low functional quality. They are visually identical to brand-name products, so they are likely to pass as genuine goods if not carefully examined. Because of the low functional quality, these counterfeits may result in a substantial financial loss to consumers or even endanger their health and safety. Consequently, both counterfeiter and distributor often face considerable punishments if they get caught. Typical examples in this category are food, beverage, agricultural products, pharmaceuticals, and automotive spare parts.

The second category appears when $\beta^{*} < \beta \leq \beta^{**}$ in case (i). Counterfeits in this category are non-deceptive and still possess a low functional quality. Counterfeits in this category are easy to tell from brand-name products because of their lower price and quality. Their functional quality is just enough for consumers to use them but their durability, stability or performance is substandard. Examples in this category include brand-name costumes, footwear, personal accessories and some luxury goods. Consumers who purchase such counterfeits are those who want to enjoy the snob appeal of brands but do not want to pay the high price of genuine articles.

The third category appears when $\beta > \beta^{**}$ in case (i) or $\beta > \beta^{***}$ in case (ii). Counterfeits in this category are non-deceptive but have a high functional quality. The functional quality of
these counterfeits is high enough to make the consumption or usage of counterfeit products just like brand-name products. Consumers can still recognize that they are counterfeits from the place of sales because the counterfeiter does not intend to cheat consumers and sells them through an unauthorized channel. Counterfeiters in this category usually face the least pressure from local enforcement agencies and some are likely to turn into licit competitors once intellectual property rights become more strictly enforced (Staake and Fleisch 2008). Examples include consumer electronics and electronic components.

The above results are consistent with the empirical findings of Staake and Fleisch (2008). They collected data for ten product categories about functional quality, visual quality, conflict with law, potential danger for users, and other four characteristics, and identify the five categories of counterfeiters called Disaggregators, Imitators, Fraudsters, Desperados and Smugglers. The first category in our paper corresponds to Fraudsters and Desperados (which are similar to Fraudsters but take a more extreme position with respect to endangering the well-being of consumers), the second category to Disaggregators, and the third category to Imitators. In this paper we do not consider Smugglers who sell counterfeits (e.g., cigarettes) mainly to evade taxes. Our normative model and predictions can be applied to any product category including those studied in Staake and Fleisch (2008).

While the analysis so far in this section assumes that only one type of counterfeits can exist in the market as in Scenarios 1 and 2, similar insights can be obtained in Scenario 3 when both types of counterfeits can exist in the market\(^3\). The following corollary presents the impact of \(\beta\) on the expected market shares of non-deceptive and deceptive counterfeits in Scenario 3.

**Corollary 1** As the counterfeit draws more value from the brand-name product with a higher \(\beta\) in Scenario 3, (a) the functional quality of non-deceptive counterfeits \(f_{\text{N}}^{(3)}\) is nondecreasing, whereas that of deceptive counterfeits \(f_{\text{D}}^{(3)}\) is unchanged; and (b) the expected market share of non-deceptive counterfeits \(Em_{\text{N}}^{(3)}\) is increasing, whereas that of deceptive counterfeits \(Em_{\text{D}}^{(3)}\) is decreasing. Moreover, if we assume a counterfeiter does not enter the market when expecting a negative profit, there exist real numbers \(\beta_N\) and \(\beta_D\) such that if \(\beta < \beta_N\), \(Em_{\text{N}}^{(3)} = 0\) and if \(\beta > \beta_D\), \(Em_{\text{D}}^{(3)} = 0\).

We can derive the following insights from Corollary 1 that is in line with Theorem 1. When a counterfeit product can draw only an insignificant brand value from the brand-name product

---

\(^3\)Scenario 3 represents a situation where two independent counterfeiters choose to sell different types of counterfeits within the same product category. This may happen due to various idiosyncratic reasons we do not consider in our model. For example, counterfeiters may incur different variable costs of production or different indirect costs of creating and managing licit or illicit distribution channels. The incorporation of these two costs into our model does not change the structural results.
While requiring significant investments to develop a high-quality product, low-quality deceptive counterfeits are more likely to be observed in the market. For example, the brand value of fast-moving consumer goods is not significant to consumers but building high quality of such goods requires significant investments; hence, we observe in practice that fake food and medicine are sold as deceptive counterfeits through licit supply chains (OECD 2008, Staake and Fleisch 2008). On the other hand, when a counterfeit product can steal a significant brand value from the brand-name product (i.e., high $\beta$), high-quality non-deceptive counterfeits are more likely to be observed. In fact, some counterfeit electronic devices such as cell phones include appealing features which are not included even in authentic products - this is called Shan-Zhai phenomenon in China (Barboza 2009, Schmidle 2010). According to Gartner, Shan-Zhai phones account for more than 20 percent of sales in China, the world’s largest mobile phone market (Barboza 2009).

6 Effectiveness of Anti-counterfeiting Strategies

In this section we investigate the effectiveness of the following anti-counterfeiting strategies: (i) operations strategies that improve the quality of the brand-name product $q_B$ or enhance the level of technological complexity $t$; (ii) marketing strategies that reduce the price of the brand-name product $p_B$ or reduce the fraction of the brand value that the counterfeit steals from the brand-name product $\beta$; and (iii) enforcement strategies that increase the chances to seize the production of counterfeits $\gamma$ or increase the level of punishment on the licit distributor for distributing counterfeits $l$. Depending on the specific goal of a brand-name company or a government, the effectiveness of these strategies may be measured in terms of an increase in the expected market share or profit of the brand-name company or a decrease in the expected market share or profit of a counterfeiter. Therefore we conduct a comprehensive analysis by examining how each strategy affects the expected market share and profit of the brand-name company as well as those of her counterfeitors.

We summarize our results in Table 1, which highlights that anti-counterfeiting strategies can result in different outcomes, depending on whether the company faces only the non-deceptive counterfeiter as in Scenario 1, only the deceptive counterfeiter as in Scenario 2, or both types of counterfeiters as in Scenario 3.

**Theorem 2.** Under each scenario $k$ (= 1, 2 or 3), the effects of anti-counterfeiting strategies on the expected market share $E_{m_i}^{(k)}$ and the expected profit $\pi_i^{(k)}$ of product $i$ (= N, D or B) are as shown in Table 1.

Table 1. Effectiveness of Anti-Counterfeiting Strategies
Scenario 1

<table>
<thead>
<tr>
<th>Operations</th>
<th>$q_B \uparrow$</th>
<th>$t \uparrow$</th>
<th>$p_B \downarrow$</th>
<th>$\beta \downarrow$</th>
<th>$\gamma \uparrow$</th>
<th>$l \uparrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_m(1)$</td>
<td>$\downarrow$</td>
<td>$\downarrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
</tr>
<tr>
<td>$\pi_N(1)$</td>
<td>$\downarrow$</td>
<td>$\downarrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
</tr>
<tr>
<td>$\pi_B(1)$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
</tr>
</tbody>
</table>

Scenario 2

<table>
<thead>
<tr>
<th>Operations</th>
<th>$q_B \uparrow$</th>
<th>$t \uparrow$</th>
<th>$p_B \downarrow$</th>
<th>$\beta \downarrow$</th>
<th>$\gamma \uparrow$</th>
<th>$l \uparrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_m(2)$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\downarrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
</tr>
<tr>
<td>$\pi_D(2)$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\downarrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
</tr>
<tr>
<td>$\pi_B(2)$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\downarrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
</tr>
</tbody>
</table>

Scenario 3

<table>
<thead>
<tr>
<th>Operations</th>
<th>$q_B \uparrow$</th>
<th>$t \uparrow$</th>
<th>$p_B \downarrow$</th>
<th>$\beta \downarrow$</th>
<th>$\gamma \uparrow$</th>
<th>$l \uparrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_m(3)$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\downarrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
</tr>
<tr>
<td>$\pi_N(3)$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\downarrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
</tr>
<tr>
<td>$\pi_D(3)$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\downarrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
</tr>
<tr>
<td>$\pi_B(3)$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\downarrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
<td>$\uparrow$</td>
</tr>
</tbody>
</table>

Note: $\uparrow$ (\$\downarrow$) denotes strictly increasing (decreasing), $\uparrow$ (\$\downarrow$) denotes non-decreasing (non-increasing), $\uparrow$ denotes non-monotonic, and $\cdot$ denotes no change. The same notation is used in Table 2. For the non-monotonic cases, we provide in the proof the conditions under which the effect of a parameter is increasing or decreasing.

In the following, we first discuss the effect of each anti-counterfeiting strategy under Scenarios 1 and 2 when only one type of counterfeits may exist in the market; and then we extend our discussion to the general case when both types of counterfeits may exist in the market under Scenario 3.

Table 1 shows that improving the quality $q_B$ will increase the expected market share and profit of the brand-name company under all scenarios. This strategy will also reduce the market share and expected profit of the non-deceptive counterfeiter under Scenario 1. When the brand-name company improves her quality, the quality of the non-deceptive counterfeiter is also increased because the counterfeit steals some value from the brand-name product. However, as the counterfeit steals only a part of the brand value with $\beta < 1$, the difference in quality between two competing products becomes larger, so the non-deceptive counterfeiter loses some consumers to the brand-name company. In contrast, the same strategy will increase the expected market share and profit of the deceptive counterfeiter under Scenario 2. When the brand-name company faces deceptive counterfeits, improving the quality makes her own products more appealing to consumers. However, the deceptive counterfeiter can have a free ride on the improved quality and obtain a larger market share. This explains the initial failure of the Scotch whisky company (Green and Smith 2002) that we mentioned earlier.

---

\textsuperscript{4} Zhang et al. (2010) also obtain $\frac{\partial \pi_D(1)}{\partial q_B} > 0$ when the quality of the brand-name product is independent of that of the non-deceptive counterfeit, i.e., $\beta = 0$, and the quality level of the non-deceptive counterfeit $q_N$ is fixed. In such a case, they further show that increasing $q_B$ is, ceteris paribus, more effective than decreasing $q_N$. 

---

21
More complex technology requires more investments to develop and produce counterfeits, hence increasing the loss when the investments get confiscated. Under Scenario 1, with a higher $t$, the non-deceptive counterfeiter will either lower his functional quality from $\tilde{f}$ to $f$ or maintain the quality level (see the proof). In the former case, he will lose his market share to the brand-name company, whereas in the latter case there will be no changes in market shares. Under Scenario 2, the deceptive counterfeiter always chooses $f$ (see Lemma 2), hence there will be no changes in market shares. Under both scenarios, the expected profits of the counterfeiter will always decrease due to the increased risk of confiscation. This strategy is often used in practice. For example, a large beverage company in Korea has implemented the strategy to enhance her products with unique tastes and elements so that the counterfeit drinks, which use recycled bottles, cannot easily imitate the original products (Choi 2009).

Table 1 suggests that by reducing the price $p_B$, the brand-name company will gain more market share in Scenario 1 by inducing some consumers to switch from non-deceptive counterfeits to branded goods. As a result, the expected profit of the non-deceptive counterfeiter will decrease. We further find that the larger $\beta$ is, the faster the expected profit of the non-deceptive counterfeiter will decrease. This is because as the non-deceptive counterfeiter steals more brand value, the brand-name company relies more on the price to compete with the counterfeiter (as the quality levels of two products are not so distinguished). However, the brand-name company has to forego her profit margin, so her expected profit may decrease as well. On the contrary, reducing $p_B$ can be ineffective in dealing with the deceptive counterfeiter in Scenario 2. If the brand-name company sets her price without regard to the presence of the deceptive counterfeiter, she would choose the optimal monopoly price of $\frac{q_B + c_B}{2}$ that maximizes her expected profit $(p_B - c_B) \left(1 - \frac{p_B}{q_B}\right)$ where $c_B (\geq 0)$ is her marginal cost. From (12), it is easy to see that $\pi_D^{(2)}$ is unimodal and maximized at $p_B = \frac{q_B}{2}$. Thus, if the brand-name company currently charges $p_B$ higher than $\frac{q_B}{2}$ (e.g., $\frac{q_B + c_B}{2}$), reducing $p_B$ will increase the expected profit of the deceptive counterfeiter. Similarly, the expected market share of deceptive counterfeits is non-monotonic in $p_B$ (i.e., $\frac{\partial \pi_M^{(2)}}{\partial p_B} = \frac{(1-\gamma)\delta}{\partial p_B} \left(1-\frac{p_B}{q_B}\right) < 0$). To understand this counterintuitive result, note that there are two effects of lowering $p_B$: (i) it increases the aggregate demand for brand-name goods and deceptive counterfeits (i.e., $\frac{\partial}{\partial p_B} \left(1 - \frac{p_B}{q_B}\right) < 0$), and (ii) it reduces the distributor’s margin from selling deceptive counterfeits (i.e., $\frac{\partial}{\partial p_B} \left(p_B - w_D^{(2)}\right) = \frac{\partial}{\partial p_B} \sqrt{\frac{q_B - p_B}{1 - \frac{p_B}{q_B}}} > 0$ from (12)). Because of the latter effect, the deceptive counterfeiter may not be able to get a free ride on the reduced $p_B$, as he does on the improved $q_B$; i.e., $\frac{\partial s^{(2)}}{\partial p_B}$ can be either positive or negative. Interestingly, however, we find that the expected market share of brand-name goods always increases as $p_B$ decreases, implying that the benefit from a higher aggregate demand outweighs a possible increase of the fraction $s^{(2)}$ of the
deceptive counterfeits. For the same reason as in Scenario 1, the expected profit of the brand-name company in Scenario 2 can increase or decrease with $p_B$. As an example of using this strategy in practice, East African Breweries launched a cut-price beer, called “Senator Keg”, to help reduce demand for illicit alcohol, which is cheap but is frequently contaminated. The company reduced the cost of the beer by negotiating a tax waiver with the government and by distributing it in kegs rather than in bottles (The Economist 2010).

Instead of competing for the price, the brand-name company may conduct marketing campaigns to educate consumers about the adversity of counterfeit goods. For example, a handbag producer may advertise that counterfeits use cheap dye and materials which may contain some harmful chemicals; and an electronic manufacturer may emphasize the fact that counterfeit electronics lack in safety features. Such campaigns help reduce the brand value the counterfeit steals from the brand-name product, i.e., reduce $\beta$. Enhancing consumers’ awareness about the legal consequences of infringing intellectual property could also be beneficial. Table 1 reveals that marketing campaigns that reduce $\beta$ work well against non-deceptive counterfeits in Scenario 1, but have no effect on deceptive counterfeits in Scenario 2 because consumers are not aware that they are purchasing deceptive counterfeits.

The most direct enforcement strategy is to raid counterfeit factories. Table 1 shows that increasing the chances to seize the production of counterfeits $\gamma$ is the only strategy that reduces the expected market shares and profits of both types of counterfeiters in Scenarios 1 and 2. There are two effects of increasing $\gamma$: (1) both types of counterfeits are less likely to reach the market, and (2) the functional quality of the non-deceptive counterfeit is non-increasing because the non-deceptive counterfeiter would be more reluctant to invest in quality improvement due to the increased risk of counterfeiting. The latter effect is similar to that of $t$. Despite the seemingly positive effects of this enforcement strategy, Schmidle (2010) argues that some governments are not putting in enough effort. There could be several possible explanations for the negligence of the governments. Counterfeit production increases the utilization of local factories especially when the economy is down (Clifford 2010), and it also helps small factories build their capabilities which will eventually result in strong legitimate businesses (Schmidle 2010). As we shall discuss later, consumers can also benefit from having low-price alternatives that bear brand labels.

The brand-name company and the government can combat counterfeits by putting more effort in monitoring their licit supply chains. Table 1 shows that raising the level of punishment $l$ on the

---

5 As a demand-side measure, increasing a penalty on illegal purchase of non-deceptive counterfeits may reduce consumers’ utilities from purchasing those goods and thereby reduce the market share of non-deceptive counterfeits. However, due to a low detection rate, this measure has not been effective (OECD 2008).
licit distributor is effective in combating the deceptive counterfeiter in Scenario 2. As discussed in
Introduction, this is the strategy executed successfully by the Scotch whisky company. Also, many
pharmaceutical companies consider deploying radio frequency identification (RFID) to deter fake
drugs from entering their licit supply chains (e.g., visit http://www.bridge-project.eu/). Recall
that we do not consider the punishment of the illicit distributor in Scenario 1 because industry and
governments put less effort in punishing small street vendors for selling non-deceptive counterfeits
due to limited success in the past (e.g., Naumann 2009).

Now let us discuss the effects of the anti-counterfeiting strategies under Scenario 3 in which
both types of counterfeits may coexist. Notice from Table 1 that their effects on the expected
market share and profit of the non-deceptive counterfeiter are the same as those under Scenario
1. As discussed in Section 4.3, this is because the non-deceptive counterfeiter does not compete
directly with the deceptive counterfeiter. In contrast, Table 1 shows that the expected market
share and profit of the deceptive counterfeiter tend to increase more under Scenario 3 than under
Scenario 2; in other words, the anti-counterfeiting strategies (except $l$ which applies only to the
deceptive counterfeiter) are less effective against the deceptive counterfeiter due to the presence of
the non-deceptive counterfeiter in Scenario 3. The reason is that those strategies that reduce the
market share of the non-deceptive counterfeiter successfully will in turn expand the target market
for the deceptive counterfeiter (i.e., increase $m_{B+D}$). For example, in Scenario 2, an increase of
 technological complexity $t$ does not affect the market share of the deceptive counterfeiter but still
reduces his expected profit due to the increased counterfeiting cost and risk; whereas in Scenario
3 the same strategy can increase his expected market share and profit. While the marketing
campaigns that reduce $\beta$ have no effect on the deceptive counterfeiter in Scenario 2, they benefit
the deceptive counterfeiter in Scenario 3 by inducing some consumers to switch from non-deceptive
counterfeits to brand-name goods and also to deceptive counterfeits inadvertently. With a higher
aggregate demand $m_{B+D}$, the licit distributor is more willing to take the risk by increasing the
fraction of deceptive counterfeits $s$. Finally, it is easy to see from Table 1 that the effects of the
anti-counterfeiting strategies on the brand-name company under Scenario 3 simply combine those
under Scenarios 1 and 2.

7 Consumer Welfare

In this section we first investigate how the anti-counterfeiting strategies we considered in the pre-
vious section affect consumer welfare. Ideally, we hope to see that anti-counterfeiting strategies
which are effective against counterfeiting also benefit consumers. However, our subsequent analysis
shows that this is not always the case.

When only brand-name products exist in the market, we can define consumer welfare $CS_B$ (or aggregate consumer surplus) as $CS_B = \int_{\frac{p_B}{q_B}}^{1} (\theta q_B - p_B) d\theta$. Similarly, we can define $CS_N$, $CS_D$, and $CS_{N+D}$ as consumer welfare when only non-deceptive counterfeits, only deceptive counterfeits, and both types of counterfeits coexist with brand-name products, respectively, as follows:

\[
CS_N = \int_{\frac{q_B}{p_B}}^{\tilde{\theta}} (\theta q_N - p_N) d\theta + \int_{\frac{p_B}{q_B}}^{1} (\theta q_B - p_B) d\theta, \tag{17}
\]

\[
CS_D = s^{(2)} \int_{\frac{p_B}{q_B}}^{1} (\theta q_D - p_B) d\theta + (1 - s^{(2)}) \int_{\frac{p_B}{q_B}}^{1} (\theta q_B - p_B) d\theta, \tag{18}
\]

\[
CS_{N+D} = \int_{\frac{q_B}{p_B}}^{\tilde{\theta}} (\theta q_N - p_N) d\theta + s^{(3,ii)} \int_{\frac{q_D}{p_B}}^{1} (\theta q_D - p_B) d\theta + (1 - s^{(3,ii)}) \int_{\frac{p_B}{q_B}}^{1} (\theta q_B - p_B) d\theta. \tag{19}
\]

In (17), the first term represents the surplus of those consumers who purchase the non-deceptive counterfeit and the second term represents the surplus of those consumers who purchase the brand-name product; $\tilde{\theta}$ and $\tilde{\theta}$ are defined in (2) and (1), respectively. In (18), the first term represents the surplus of those consumers who are cheated and receive the deceptive counterfeit although they pay the price of the brand-name product, and the second term represents the surplus of those consumers who purchase and receive the brand-name product. The first term in (19) is the same as that in (17), and the second and third terms in (19) correspond to the first and second terms in (18) but with different fractions $s^{(3,ii)}$ and $s^{(2)}$ of deceptive counterfeits in equilibrium. Considering the chances that non-deceptive and/or deceptive counterfeits do not reach the market due to seizure, we can further define $CS^{(1)}$, $CS^{(2)}$ and $CS^{(3)}$ as the expected consumer welfare in Scenarios 1, 2 and 3, respectively, as follows:

\[
CS^{(1)} = (1 - \gamma)CS_N + \gamma CS_B, \tag{20}
\]

\[
CS^{(2)} = (1 - \gamma)CS_D + \gamma CS_B, \tag{21}
\]

\[
CS^{(3)} = (1 - \gamma)^2 CS_{N+D} + \gamma (1 - \gamma)(CS_N + CS_D) + \gamma^2 CS_B. \tag{22}
\]

One can easily show that $CS_D < CS_B < CS_N$, hence $CS^{(2)} < CS_B < CS^{(1)}$. This is because the non-deceptive counterfeiter in Scenario 1 offers a low price alternative to those consumers who like to enjoy the brand value of the brand-name product but do not appreciate its high quality or cannot afford its high price, whereas some consumers are cheated to receive low-quality deceptive counterfeits in Scenario 2. Due to the opposite effects of non-deceptive and deceptive counterfeits, $CS^{(3)}$ can be either higher or lower than $CS_B$.

\footnote{We do not consider the socio-economic effects of counterfeiting on corruption, criminal activities, employment, environment, innovation, and tax revenues. If taking these indirect or long-term effects into account, non-deceptive}
The next theorem highlights that consumers do not necessarily benefit from several anti-counterfeiting strategies.

**Theorem 3.** Under each scenario $k (= 1, 2$ or $3)$, the effects of anti-counterfeiting strategies on expected consumer welfare $CS^{(k)}$ are summarized in Table 2.

Table 2. Effects of Anti-Counterfeiting Strategies on Expected Consumer Welfare

| Operations | $q_B$ | $t$ | $CS^{(1)}$ | $CS^{(2)}$ | $CS^{(3)}$
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>↑</td>
<td>↑</td>
<td>↓</td>
<td>↓</td>
<td>↓</td>
</tr>
<tr>
<td>$p_B$</td>
<td>↓</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>$\beta$</td>
<td>↓</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>Enforcement</td>
<td>$\gamma$</td>
<td>↑</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td></td>
<td>$l$</td>
<td>↑</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
</tr>
</tbody>
</table>

Let us discuss the impact of each anti-counterfeiting strategy on consumer welfare. While improving $q_B$ always increases the expected market share and profit of the brand-name company as shown in Table 1, this strategy does not always benefit consumers in all three scenarios. Under Scenarios 1 and 3, this strategy can lead the non-deceptive counterfeiter to lower his functional quality as well as his wholesale price to compete against high-quality brand-name products. While this strategy will reduce the market share of non-deceptive counterfeits (see Table 1), those consumers who purchase non-deceptive counterfeits can suffer from the lower quality, resulting in a lower total surplus. When deceptive counterfeits are in the market under Scenarios 2 and 3, there are two opposing effects of improving $q_B$: consumers benefit from high quality and more consumers buy products, but at the same time increasingly more consumers are deceived to buy counterfeits. The latter effect is due to the licit distributor who increases the fraction $s$ of deceptive counterfeits as the aggregate demand becomes high. When the quality of deceptive counterfeits is much lower than that of brand-name products, the latter effect dominates the former effect and expected consumer welfare decreases (see the proof).

We have shown in Table 1 that more complex technology with a higher $t$ can be an entry barrier to both types of counterfeiters. However, Table 2 shows that more complex technology could decrease consumer welfare when non-deceptive counterfeiters exist in the market under Scenario 1 and 3. As discussed earlier, it may lead the non-deceptive counterfeiter to lower his functional quality and therefore those consumers who purchase the non-deceptive counterfeit can suffer from its lower quality. A welfare loss is more significant in Scenario 3 than in Scenario 1 because a lower quality counterfeit may also decrease consumer welfare. Unfortunately, however, consumers are unlikely to take those effects into consideration when making purchasing decisions.
of the non-deceptive counterfeit leads some consumers to switch from non-deceptive counterfeits to brand-name products but some of those consumers will receive deceptive counterfeits unknowingly. In Scenario 2, this strategy has no impact on consumer welfare, as the deceptive counterfeiter always chooses the lowest functional quality.

Table 2 reveals that the impact of reducing the price $p_B$ is similar to that of improving the quality $q_B$: both strategies do not always benefit consumers. In contrast, the marketing campaigns that reduce $\beta$ hurt consumer welfare under all scenarios. Although this strategy is effective in reducing the sales of non-deceptive counterfeits (see Table 1), it makes those consumers who purchase non-deceptive counterfeits knowingly or deceptive counterfeits unknowingly enjoy the counterfeit brand less, resulting in a welfare loss.

The direct enforcement strategy affects consumer welfare differently, depending on the type of counterfeits that are available in the market. As discussed in the previous section, with an increase of enforcement effort $\gamma$, both types of counterfeits are less likely to reach the market and the non-deceptive counterfeiter may reduce his functional quality. Under Scenario 1, consumers suffer from less availability of non-deceptive counterfeits, cheaper substitutes for brand-name goods, as well as from their lower quality level. Under Scenario 2, consumer welfare is increased because a less number of consumers will be fooled with deceptive counterfeits. Depending on which force (negative for non-deceptive or positive for deceptive) is dominant, the effect on expected consumer welfare can be either positive or negative when both types of counterfeits exist under Scenario 3.

More severe penalty $l$ on the distribution of deceptive counterfeits increases the risk of the licit distributor who sources deceptive counterfeits illegally. Thus it always reduces the amount of deceptive counterfeits that break successfully into the licit supply chain, so a less number of consumers will receive deceptive counterfeits unknowingly under Scenario 2 and 3.

8 Concluding Remarks

Today counterfeit products are being produced and consumed in virtually all economies (OECD 2008). While easy-to-manufacture goods dominated counterfeit supply until a decade ago, there has been an alarming expansion of product categories being infringed. As a result of outsourcing and offshoring, counterfeiters have easy access to modern technology and equipment, and they are capable of producing high-quality replicas. Consumers are not easily deceived by fake goods that are sold by vendors in open markets and unknown internet sites. These changing business conditions require industry and governments to enhance their understanding of the current and potential counterfeiters they may face and to develop strategies to limit their activities.
To aid the efforts of industry and governments to combat counterfeiting, we have developed a normative model of counterfeiting. Our model captures the recent changes in counterfeiting supply and demand that are not addressed in the previous literature. For example, the previous literature focuses on the pricing decision of a counterfeiter, assuming that his type between non-deceptive and deceptive is known, the quality level of his goods is fixed, and he is capable of selling his goods, even deceptive ones, directly to consumers. In contrast, our model takes into account the strategic decisions of a counterfeiter regarding his type, price, and functional quality; and the fundamental differences between non-deceptive and deceptive counterfeits in consumers’ awareness, distribution channels, and penalty on illegal distribution. We also consider a realistic situation where both types of counterfeits may coexist in the market. Modeling these factors explicitly enables us to evaluate several anti-counterfeiting strategies and draw practical insights.

Our model predicts that there are three categories of counterfeits in equilibrium: low-quality deceptive, low-quality non-deceptive, and high-quality non-deceptive counterfeits. This result is consistent with empirical findings of Staake and Fleisch (2008). As the brand value of a product is more important to consumers, a counterfeiter tends to produce non-deceptive and high-quality products. On the other hand, as a product becomes more complex, low-quality counterfeits become more prevalent in the market. Our analysis highlights the importance of implementing anti-counterfeiting strategies that are tailored to specific products and markets. Unfortunately, there is no anti-counterfeiting strategy that is always effective against counterfeiting and at the same time benefit consumers. More specifically,

(a) A strategy that may work against one type of counterfeits may not work well against the other type. For example, improving the quality of the brand-name good works well against non-deceptive counterfeits but not against deceptive counterfeits.

(b) A strategy that may work against one type of counterfeits may not work against even the same type when the other type of counterfeits is also present in the market. For example, putting more effort on seizure of counterfeits is effective against deceptive counterfeits but may become ineffective as non-deceptive counterfeits become more dominant in the market.

(c) A strategy that is effective in terms of one measure may not be effective in terms of the other measures. For example, reducing the price of the brand-name good helps regain the market share of the brand-name company but it may reduce her expected profit, while increasing the expected profit of the deceptive counterfeiter.

(d) A strategy that is effective against counterfeiting does not necessarily improve consumer welfare. For example, enhancing technological complexity can reduce the expected profit of either type of a counterfeiter by raising the counterfeiting cost and risk, but in response to the increased cost and
risk, the non-deceptive counterfeiter may lower the quality of his goods and consequently reduce consumer welfare.

On the surface, these results do not appear so encouraging. However, having observed the rapid growth of counterfeit markets in the last few decades, we all understand that combating counterfeiters is extremely challenging. The main message is that industry and governments should be cautious in developing their strategies to minimize potential negative consequences. Tables 1 and 2 we presented earlier can be useful guidelines to understand trade-offs among the different goals of countering counterfeits.

References


Appendix

A. Proofs

We use (A1)-(A3) to indicate the following assumptions we have made earlier:

(A1) $q_B > q_N = f_N + \beta q_B$ and $q_B > q_D = f_D + \beta q_B$;

(A2) $1 - \frac{p_B}{q_B} > 0$ and $1 - \frac{p_B - p_N}{q_B - q_N} > 0$ so that $m_B > 0$ in Scenarios 1 and 2, respectively;

(A3) $p_B = w_B$.

Proof of Lemma 1: From (7), we obtain $\frac{\partial^2 \pi_N}{\partial q_N^2} = \frac{(1-\gamma)p_B^2}{4(1-\beta)q_B - f_N}$, which is positive by (A1). Thus $\pi_N$ is convex in $f_N$ and hence the result follows. □

Proof of Lemma 3: From (16), we have $\frac{\partial f_D}{\partial w_D} = -t < 0$. Thus, $f_D^{(3)} = f$. Next, let us compare $w_D^{(3)}$ in (15) and $w_D^{(2)}$ (12) given that $f_D^{(3)} = f_D^{(2)} = f$. For $w_D^{(3)} < w_D^{(2)}$, it suffices to show that $m_{B+D} < 1 - \frac{p_B}{q_B}$ in equilibrium. By substituting $p_N^{(1)} = \frac{3(\beta q_B + f_N)p_B}{4q_B}$ into $m_{B+D} = 1 - \frac{p_B - p_N}{(1-\beta)q_B - f_N}$, we obtain

$$m_{B+D} = 1 - \frac{p_B}{q_B} - \frac{(\beta q_B + f_N)p_B}{4q_B(1 - \beta)q_B - f_N},$$

which is smaller than $1 - \frac{p_B}{q_B}$ by (A1).

By substituting $w_D^{(2)}$ in (12) into $s^{(2)}$ given by (15) and (12) into $s^{(3)}$ given by (16), we obtain $s^{(2)} = \frac{1}{2} - \frac{1}{2} \sqrt{\frac{t}{f_D^{(1)}}}$ in equilibrium. By substituting $w_D^{(3)}$ in (15) into $s^{(3)} = \gamma s^{(3,ii)} + (1 - \gamma)s^{(3,ii)}$, we obtain after simplification $s^{(3)} = \frac{1}{2} - \frac{1}{2} \sqrt{\frac{t}{f_D^{(1)}} \frac{(1-\gamma)m_{B+D} + (1-\gamma)\gamma}{1-\gamma} + \gamma}$.

We now compare $s^{(3)}$ with $s^{(2)}$ as follows: $s^{(3)} \leq \frac{1}{2} - \frac{1}{2} \sqrt{\frac{t}{f_D^{(1)}} \frac{(1-\gamma)m_{B+D} + (1-\gamma)\gamma}{1-\gamma} + \gamma} < \frac{1}{2} - \frac{1}{2} \sqrt{\frac{t}{f_D^{(1)}}} = s^{(2)},$ where the first inequality holds because

$$\frac{1}{1-\gamma} \left( \frac{1}{m_{B+D}} + \frac{\gamma}{1-\gamma} \right) \geq \frac{1}{1-\gamma} \left( \frac{1}{m_{B+D} + \gamma(1-\gamma)} \right) > \frac{1}{2} - \frac{1}{2} \sqrt{\frac{t}{f_D^{(1)}}}$$

for any $\gamma \in [0, 1]$, which can be verified by checking $\frac{\partial^2}{\partial \gamma^2} \left( \frac{1}{1-\gamma} \left( \frac{1}{m_{B+D}} + \frac{\gamma}{1-\gamma} \right) - \frac{1}{1-\gamma} \left( \frac{1}{m_{B+D} + \gamma(1-\gamma)} \right) \right) < 0$, $\frac{1}{2} - \frac{1}{2} \sqrt{\frac{t}{f_D^{(1)}}}$.

The second inequality follows from $m_{B+D} < 1 - \frac{p_B}{q_B}$.

The proofs for $\bar{E}m^{(3)}_D < \bar{E}m^{(2)}_D$ and $\bar{E}m^{(3)}_B < \bar{E}m^{(2)}_B$ are similar to the above proof, and hence are omitted. □

Proof of Theorem 1: Define $\beta^*$, $\beta^{**}$ and $\beta^{***}$ as follows: (i) $\beta^*$ solves $\pi_N^{(1)}(f) = \pi_D^{(2)}(f)$, (ii) $\beta^{**}$ solves $\pi_N^{(1)}(f) = \pi_N^{(1)}(f)$, and (iii) $\beta^{***}$ solves $\pi_N^{(1)}(f) = \pi_D^{(2)}(f)$. To prove the existence and uniqueness of $\beta^*$, $\beta^{**}$ and $\beta^{***}$, it suffices to show that $\frac{\partial f_D^{(2)}}{\partial \beta} = 0$ and $\frac{\partial f_D^{(1)}}{\partial \beta} > 0$ (see Figure 3 for illustration). Since $\pi_N^{(1)}(f) = \pi_N^{(1)}(f)$, we have $\frac{\partial f_D^{(2)}}{\partial \beta} = 0$. From (7),

$$\frac{\partial f_D^{(1)}}{\partial \beta} = \frac{p_B q_B (1-\gamma)}{\pi_B f_N^{(1)}} + \frac{p_B q_B (1-\gamma) f_N^{(1)} + p_B q_B}{\pi_B (1-\beta) q_B - f_N^{(1)}} > 0$$

for any $f_N$. To prove $\frac{\partial f_D^{(1)}}{\partial \beta} > \frac{\partial f_D^{(1)}}{\partial \beta}$, we
Figure 3: (a) Case (i) \( \beta^* < \beta^{**} \) and (b) Case (ii) \( \beta^* \geq \beta^{**} \). (Note: for the purpose of illustration, \( \pi_N^{(1)}(f) \) and \( \pi_N^{(1)}(f) \) are drawn linearly although they are not, in general, linear in \( \beta \).)

Define

\[
\Delta \equiv \pi_N^{(1)}(f) - \pi_N^{(1)}(f) = \frac{(1 - \gamma)p_B^2}{8q_B} \left( \frac{f + \beta q_B}{(1 - \beta)q_B - f} - \frac{f + \beta q_B}{(1 - \beta)q_B - f} \right) - t(f - f). \tag{24}
\]

Then, \( \frac{\partial \Delta}{\partial \beta} = \frac{(1 - \gamma)p_B^2 q_B(f - f)}{4(1 - \beta)q_B - f} \left\{ \frac{(1 - \beta)q_B - f}{2} \right\} \). Since \( q_B - q_N = (1 - \beta)q_B - f_N > 0 \) for all \( f_N \) by (A1), \( (1 - \beta)q_B - (f + f)/2 > 0 \). Therefore, \( \frac{\partial \Delta}{\partial \beta} = \frac{\partial \pi_N(f)}{\partial f} - \frac{\partial \pi_N(f)}{\partial \beta} > 0 \).

From (7) and (12), we can also find the closed-form expressions of \( \beta^* \), \( \beta^{**} \) and \( \beta^{***} \) as follows:

\[
\beta^* = \frac{8Gq_B(f - f) - p_B^2 f}{8Gq_B^2 + q_B p_B^2} \quad \text{where} \quad G \equiv \frac{1}{2} \left\{ \sqrt{\frac{p_B^2}{q_B^2} - \sqrt{f - f}} \right\}^2, \tag{25}
\]

\[
\beta^{**} = \frac{2q_B - f - f - \sqrt{(f - f)^2 + \frac{1 - \gamma}{2t} p_B^2}}{2q_B}, \quad \text{and} \tag{26}
\]

\[
\beta^{***} = \frac{A(q_B - f) - f}{(A + 1)q_B} \quad \text{where} \quad A \equiv \frac{8q_B}{p_B^2 (1 - \gamma)} \left\{ (1 - \gamma)G + t(f - f) \right\}. \tag{27}
\]

Remark: As shown in the proof of Lemma 1, the benefit of functional quality of the non-deceptive counterfeit is increasing and convex. If the cost of functional quality is non-linear and increasing faster than the benefit, then an interior optimum \( f_N^{(i)} \in (f, f) \) may exist. Specifically, suppose the investment of \( t \cdot g(f_N) \) is required to build counterfeit goods having functional quality \( f_N \). The necessary condition for such \( f_N^{(i)} \) to exist is \( \frac{\partial^2 g}{\partial f_N} < 0 \) for some \( f_N \in (f, f) \), which is simplified to \( \frac{\partial^2 g}{\partial f_N^2} > \frac{(1 - \gamma)q_B^2}{4t(1 - \beta)q_B - f_N} \). This condition is not intuitive as it depends on the values of \( p_B, q_B, \beta, t \) and \( \gamma \) as well as the non-linear functional form of \( g(\cdot) \). In this case, the category of counterfeits in Theorem 1 looks similar to case (ii) in which the counterfeiter chooses low-quality deceptive counterfeits if \( \beta \) is below some threshold or chooses non-deceptive counterfeits having \( f_N^{(i)} \) otherwise.
Proof of Corollary 1: (a) Suppose $f^{(1)}_N = \bar{f}$ for some $\beta = \beta_1$. This means $\Delta \geq 0$ from (24). Because $\frac{\partial \Delta}{\partial \beta} > 0$ from the proof of Theorem 1, $f^{(1)}_N = \bar{f}$ for any $\beta \geq \beta_1$. Conversely, if $f^{(1)}_N = f$, i.e., $\Delta < 0$ for some $\beta = \beta_2$, then $f^{(1)}_N = \bar{f}$ or $\bar{f}$ for any $\beta \geq \beta_2$. Therefore, $\frac{\partial f^{(1)}_N}{\partial \beta} \geq 0$. Since $f^{(1)}_N = f^{(3)}_N$, $\frac{\partial f^{(3)}_N}{\partial \beta} \geq 0$. By Lemma 3, $f^{(1)}_D = f^{(3)}_D = f$, hence $\frac{\partial f^{(3)}_D}{\partial \beta} = 0$.

(b) We obtain the expected market shares of the three products in Scenario 3 as follows:

\[
E m_N = (1 - \gamma) \left\{ \frac{p_B - m p_B}{(1 - \beta)q_B - f_N} - \frac{p_B}{f_N + \beta q_B} \right\},
\]

\[
E m_D = (1 - \gamma) \left\{ \gamma s^{(3,ii)} \left( 1 - \frac{p_B}{q_B} \right) + (1 - \gamma) s^{(3,ii)} m_{B+D} \right\}, \quad \text{and}
\]

\[
E m_B = (1 - \gamma) \left\{ \gamma \left( 1 - s^{(3,ii)} \right) \left( 1 - \frac{p_B}{q_B} \right) + (1 - \gamma) \left( 1 - s^{(3,ii)} \right) m_{B+D} \right\}
+ \gamma \left\{ (1 - \gamma) \left( 1 - \frac{p_B - m p_B}{(1 - \beta)q_B - f_N} \right) + \gamma \left( 1 - \frac{p_B}{q_B} \right) \right\}.
\]

Since $p^{(3)}_N = p^{(1)}_N$, we substitute $p^{(1)}_N$ into $E m^{(3)}_N$ in (28) and $E m^{(3)}_D$ in (29). After simplification, we get $E m^{(3)}_N = \frac{1}{2} \left\{ (1 - \gamma) p_B \right\} \frac{\gamma}{\{1 - \beta)q_B - f^{(3)}_N\}^2} = \frac{1}{2} \frac{\gamma}{\{1 - \beta)q_B - f^{(3)}_N\}^2}$ and $E m^{(3)}_D = (1 - \gamma) \left\{ \gamma (1 - \frac{p_B}{q_B}) + (1 - \gamma) m_{B+D} \right\} \left\{ \frac{1}{2} - \frac{1}{2} \frac{\gamma}{\{1 - \beta)q_B - f^{(3)}_N\}^2} \right\}$. Then, $\frac{\partial E m^{(3)}_N}{\partial \beta} = \frac{1}{2} \frac{\gamma}{\{1 - \beta)q_B - f^{(3)}_N\}^2}$ and $\frac{\partial E m^{(3)}_D}{\partial \beta} = \frac{1}{2} \frac{\gamma}{\{1 - \beta)q_B - f^{(3)}_N\}^2}$. Since $\frac{\partial f^{(3)}_D}{\partial \beta} \geq 0$ from Part (a), $\frac{\partial E m^{(3)}_N}{\partial \beta} > 0$. We can also compute $\frac{\partial E m^{(3)}_D}{\partial \beta} = \frac{1}{2} \frac{\gamma}{\{1 - \beta)q_B - f^{(3)}_N\}^2}$ by observing $E m^{(3)}_N$, one can easily see $\frac{\partial E m^{(3)}_D}{\partial \beta} > 0$. Thus it suffices to show $\frac{\partial m_{B+D}}{\partial \beta} < 0$. From (23), $\frac{\partial m_{B+D}}{\partial \beta} = \frac{p_B(-q_B - \frac{\partial f^{(3)}_N}{\partial \beta})}{4\{1 - \beta)q_B - f^{(3)}_N\}^2}$, which is negative due to $\frac{\partial f^{(3)}_N}{\partial \beta} \geq 0$ from Part (a).

To prove the existence of $\beta_N$, it suffices to show that $\pi^{(3)}_N$ is increasing in $\beta$. Suppose that $\beta$ is increased from $\beta_L$ to $\beta_H$. By part (a), the optimal functional quality of non-deceptive counterfeits is increased from $f^{(3)}_N(\beta_L)$ to $f^{(3)}_N(\beta_H)$ accordingly. Since $\pi^{(3)}_N = \pi^{(1)}_N$, we can represent $\pi^{(3)}_N$ using (7) as follows:

\[
\pi^{(3)}_N(\beta_N) = \frac{p_B^2 \{1 - \gamma(1 - \gamma) \{f^{(3)}_N(\beta_H) + \beta_H q_B\} \}}{8q_B \{1 - \beta_H)q_B - f^{(3)}_N(\beta_H)\}} - tf^{(3)}_N(\beta_H)
\]

\[
\geq \frac{p_B^2 \{1 - \gamma(1 - \gamma) \{f^{(3)}_N(\beta_L) + \beta_L q_B\} \}}{8q_B \{1 - \beta_H)q_B - f^{(3)}_N(\beta_L)\}} - tf^{(3)}_N(\beta_L)
\]

\[
> \frac{p_B^2 \{1 - \gamma(1 - \gamma) \{f^{(3)}_N(\beta_L) + \beta_L q_B\} \}}{8q_B \{1 - \beta_H)q_B - f^{(3)}_N(\beta_L)\}} - tf^{(3)}_N(\beta_H) = \pi^{(3)}_N(\beta_L),
\]

where the first inequality follows the optimality of $f^{(3)}_N(\beta_H)$ and the second inequality follows the fact that $\beta_L < \beta_H$. Next, to prove the existence of $\beta_D$, we show $\pi^{(3)}_D$ is decreasing in $\beta$. By differentiating $\pi^{(3)}_D(f^{(3)}_D)$ in (16) with $\beta$, we obtain $\frac{\partial \pi^{(3)}_D(f^{(3)}_D)}{\partial \beta} = \frac{1}{4} \frac{1 - \gamma}{\{1 - \gamma) \{f^{(3)}_N(\beta_H) + \beta_H q_B\} \}} - tf^{(3)}_N(\beta_H)$.
Since $\frac{\partial m_{B,D}}{\partial q} < 0$ as shown above, we have $\frac{\partial \pi^{(3)}_N(f^{(3)}_D)}{\partial q} < 0$.

We can also obtain the closed-form expressions of $\beta_N$ and $\beta_D$. Since $\beta_N$ solves
\[ \frac{\partial^2 (1-\gamma)(f^{(3)}_N + \beta_N q_B)}{\partial q_B} \]
$t f_N^{(3)} = 0$, we get $\beta_N = \frac{8q_B(q_B-f^{(3)}_N)+f^{(3)}_N}{p_B^{(1)}(1-\gamma)q_B^2+f^{(3)}_N}$. Similarly, $\beta_D$ solves $\frac{1}{2}(1-\gamma)\times$
\[ \sqrt{p_B \left((1-\gamma)m_{B,D} + \gamma(1-p_B/q_B)\right) - v^l} - t f_D^{(3)} = 0, \text{ so } \beta_D = 1 - \frac{p_B}{4q_B(1-m_{B,D}+\frac{3p_B}{4q_B})} - f^{(3)}_D/q_B. \]

**Proof of Theorem 2:** Since we have obtained the closed-form expressions of all equilibrium outcomes, we can also obtain the closed expressions of $E m^{(k)}$ and $\pi^{(k)}_i$ in equilibrium, and compute their first derivatives $\frac{\partial E m^{(k)}_i}{\partial x}$ and $\frac{\partial \pi^{(k)}_i}{\partial x}$ with respect to each parameter $x (= q_B, t, p_B, \beta, \gamma$ and $l$) for each pair of $i (= N, D$ or $B$) and $k (= 1, 2$ or $3$). To determine the signs of $\frac{\partial E m^{(k)}_i}{\partial x}$ and $\frac{\partial \pi^{(k)}_i}{\partial x}$, we will need to use the following results about the signs of $\frac{\partial f^{(k)}_N}{\partial x}$ and $\frac{\partial f^{(k)}_D}{\partial x}$. From Lemmas 2 and 3, we know $f^{(2)}_D = f^{(3)}_D = f$, hence $\frac{\partial f^{(2)}_D}{\partial x} = \frac{\partial f^{(3)}_D}{\partial x} = 0$ for all $x$. Using the method described in the proof of Corollary 2(a), we can also show $\frac{\partial f^{(1)}_D}{\partial x} < 0$ for $x = q_B$, $t$ or $\gamma$, $\frac{\partial f^{(1)}_D}{\partial x} > 0$ for $x = p_B$ or $\beta$, and $\frac{\partial f^{(1)}_D}{\partial x} = 0$. Since $f^{(1)}_N = f^{(3)}_N$, $\frac{\partial f^{(3)}_N}{\partial x} = \frac{\partial f^{(1)}_N}{\partial x}$ for every parameter $x$.

There are 84 derivatives for us to evaluate. Due to the limited space, we omit detailed proofs (which are available upon the request to the authors.) \(\square\)

**Proof of Theorem 3:** Below we present the proof for the effect of $q_B$ on consumer welfare $CS^{(k)}$ for $k = 1, 2$ and $3$. The proofs for other parameters are similar and available upon request to the authors. Suppose the brand-name company has improved her quality $q_B$. Then;

(a) If $f^{(1)}_N$ is unchanged, then $CS^{(1)}$ increases. If $f^{(1)}_N$ is decreased from $\bar{f}$ to $f$ at $q_B^l$ (i.e., $f^{(1)}_N = \bar{f}$ for $q_B \leq q_B^l$ and $f^{(1)}_N = f$ for $q_B > q_B^l$), $CS^{(1)}$ decreases when $q_B$ is increased from $q_B^l$ to $q_B^l + \varepsilon$ for small $\varepsilon > 0$.

(b) If $q_B - q_B^l > \frac{1}{2} \frac{1-\gamma}{(1-\gamma)(p_B^{(1)}s_2^{(1)} + \frac{1}{2} (1 - \gamma) q_B^{(2)} \frac{\partial q_B}{\partial x})} (\geq 0)$, $CS^{(2)}$ decreases. Otherwise, $CS^{(2)}$ increases.

(c) If $f^{(1)}_N$ is unchanged, then $CS^{(3)}$ increases if and only if $(1 - \gamma) \frac{\partial CS^{(3)}_{B,D}}{\partial q_B} + \gamma (1 - \gamma) (\frac{\partial CS^{(3)}_{B,L}}{\partial q_B} + \frac{\partial CS^{(3)}_{B,B}}{\partial q_B}) > 0$. If $f^{(3)}_N$ is decreased from $\bar{f}$ to $f$ at $q_B^l$ (i.e., $f^{(3)}_N = \bar{f}$ for $q_B \leq q_B^l$ and $f^{(3)}_N = f$ for $q_B > q_B^l$), $CS^{(3)}$ decreases from $q_B^l$ to $q_B^l + \varepsilon$ for small $\varepsilon > 0$.

**The proof of (a):** We first prove that $CS^{(1)}$ is increasing in $q_B$ for any given $f_N$. From the definition of $CS_B$, it is easy to see that $CS_B$ is increasing in $q_B$. Hence, it suffices to show that $CS_N$ is also increasing in $q_B$. Suppose that $q_B$ is increased from $q_{BL}$ to $q_{BH}$. Then $q_N$ is also increased from $q_{NL} \equiv f_N + \beta q_{BL}$ to $q_{NH} \equiv f_N + \beta q_{BH}$ given $f_N$; $\bar{\theta}$ is decreased from $\bar{\theta}_L \equiv \frac{3p_B}{4q_B} \to \bar{\theta}_H \equiv \frac{3p_B}{4q_B}$; and $\bar{\theta}$ is decreased from $\bar{\theta}_L \equiv \frac{3p_B}{4q_B}$ to $\bar{\theta}_H \equiv \frac{3p_B}{4q_B}$. Using $p^{(1)}_N = \frac{3p_B q_N}{4q_B}$, (17) can be rewritten as

\[ CS_{q_B}(q_{BH}) = \int_{\bar{\theta}_H}^{\bar{\theta}_H} \left( \theta - \frac{3p_B}{4q_B} \right) q_{NH} d\theta + \int_{\bar{\theta}_H}^{1} (\theta q_{BH} - p_B) d\theta \]
for any given i.e., consumers with Following the method similar to the proof of (a), we can show these consumers prefer the non-deceptive counterfeit having functional quality where \( s \). From (22), In equilibrium, \( CS(2) \) in (21) becomes

\[
CS(2) = (1 - \gamma) \left\{ s^{(2)} \int_{\bar{q}_B}^{q_B} \left( \theta - \frac{3\beta B}{4\beta B} \right) q_B d\theta + \gamma \int_{\bar{q}_B}^{q_B} (\theta q_B - p_B) d\theta \right\} + \gamma \int_{\bar{q}_B}^{q_B} (\theta q_B - p_B) d\theta
\]

where \( s^{(2)} = \frac{1}{2} - \frac{1}{4} \sqrt{\frac{t}{p_B(1 - \frac{t}{\bar{q}_B})}} \) is derived in the proof of Lemma 3. Then,

\[
\frac{\partial CS(2)}{\partial q_B} = (1 - \gamma) \left\{ \frac{p_B^2}{q_B^2} s^{(2)} \left( q_B - q^{(2)}_D \right) + \frac{1}{2} \left( 1 - \frac{p_B^2}{q_B^2} \right) \frac{\partial s^{(2)}}{\partial q_B} \left( q_B - q^{(2)}_D \right) \right\} + \gamma \left( 1 - \frac{p_B^2}{q_B^2} \right) \left( 1 - s^{(2)} \right) \]

\[
= -(1 - \gamma) \left\{ \frac{p_B^2}{q_B^2} s^{(2)} + \frac{1}{2} \left( 1 - \frac{p_B^2}{q_B^2} \right) \frac{\partial s^{(2)}}{\partial q_B} \right\} \left( q_B - q^{(2)}_D \right) + \frac{1}{2} \left( 1 - \frac{p_B^2}{q_B^2} \right) \left( 1 - (1 - \gamma)s^{(2)} \right).
\]

Using \( \frac{\partial s^{(2)}}{\partial q_B} = \frac{1}{2(2q_B - p_B)^2} \int_{\bar{q}_B}^{q_B} \frac{q_B}{p_B(2q_B - p_B)} \), we have the following result: \( \frac{\partial CS(2)}{\partial q_B} < 0 \) if and only if \( q_B > q^{(2)}_D \),

\[
\lambda \left[ \frac{n}{(1 - \gamma) \left\{ \frac{t}{q_B^3} s^{(2)} + \frac{1}{2} \left( 1 - 2 \frac{p_B^2}{q_B^2} \right) \frac{\partial s^{(2)}}{\partial q_B} \right\} \left( 1 - (1 - \gamma)s^{(2)} \right) \right] > 0.
\]

The proof of (c): From (22), \( \frac{\partial CS(3)}{\partial q_B} = (1 - \gamma)^2 \frac{\partial CS_{N+D}}{\partial q_B} + \gamma (1 - \gamma) \left( \frac{\partial CS_N}{\partial q_B} + \frac{\partial CS_D}{\partial q_B} \right) + \gamma^2 \frac{\partial CS_B}{\partial q_B} \). Following the method similar to the proof of (a), we can show \( CS(3) \) is decreased when \( f_N^{(1)} \) is
changed from $\overline{f}$ to $f$ for any given $q_B$. In the case when $f_N^{(1)}$ is unchanged, we know from parts (a) and (b) that $\frac{\partial CS_N}{\partial q_B} > 0$ and $\frac{\partial CS_B}{\partial q_B} > 0$ but $\frac{\partial CS_D}{\partial q_B}$ can be either positive or negative. It can also be shown that $\frac{\partial CS_N + D}{\partial q_B}$ can be either positive or negative. Thus, one has to evaluate $(1 - \gamma)^2 \frac{\partial CS_N + D}{\partial q_B} + \gamma(1 - \gamma)(\frac{\partial CS_N}{\partial q_B} + \frac{\partial CS_D}{\partial q_B}) + \gamma^2 \frac{\partial CS_B}{\partial q_B}$ to determine the sign of $\frac{\partial CS}{\partial q_B}$. By substituting $\frac{\partial CS_N + D}{\partial q_B}$, $\frac{\partial CS_N}{\partial q_B}$, $\frac{\partial CS_D}{\partial q_B}$, and $\frac{\partial CS_B}{\partial q_B}$ into $\frac{\partial CS}{\partial q_B}$, we can find the condition under which $\frac{\partial CS}{\partial q_B} < 0$ (which is very complex and hence omitted here). □

B. Various Forms of Penalties for Counterfeiting

In the base model, we have assumed that the investment of the counterfeiter is a sunk cost. His expected profit is represented as $\pi_i(w_i, f_i) = -tf_i + (1 - \gamma)v_i(w_i, f_i) + \gamma \cdot 0$, where $v_i$ denotes the profit of the counterfeiter $i (= N$ or $D$) when he does not get caught by the authorities. In some cases, the investment may not be a sunk cost or the counterfeiter may have to pay an additional fine when he gets caught. In the following, we discuss various forms of punishments on the counterfeiter and present the corresponding expected profit of the counterfeiter:

- The counterfeiter pays an additional fine $h$ when he gets caught: $\pi_i(w_i, f_i) = -tf_i + (1 - \gamma)v_i(w_i, f_i) + \gamma(-h)$.

- The investment has residual or salvage value (assuming no depreciation) but gets confiscated when the counterfeiter gets caught: $\pi_i(w_i, f_i) = (1 - \gamma)v_i(w_i, f_i) + \gamma(-tf_i)$.

- The combination of the above two: $\pi_i(w_i, f_i) = (1 - \gamma)v_i(w_i, f_i) + \gamma(-tf_i - h)$.

Since $\pi_i(w_i, f_i)$ depends on $w_i$ through $v_i(w_i, f_i)$ in all cases, the pricing decision of the counterfeiter in the base model will not change. Moreover, it is easy to verify that the second derivative of $\pi_i(w_i, f_i)$ with respect to $f_i$ is identical in all cases and that the first derivative differs only by a constant term. This suggests that the decision of the counterfeiter regarding the functional quality in the base model will not change, either. Therefore, our insights obtained previously still hold under the various forms of punishments.
Equipment Procurement Strategy with Reservation and Forecast Revision: A Heuristic Approach

Feryal Erhun, Chen Peng
Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, USA.
ferhun@stanford.edu, chenpeng@stanford.edu

Erik F. Hertzler
Staff Engineer, Intel Corporation, Chandler, AZ 85226, USA.
erik.f.hertzler@intel.com

Karl G. Kempf
Decision Technologies Group, Intel Corporation, Chandler, AZ 85226, USA.
karl.g.kempf@intel.com

Presenter: Chen Peng
In capital-intensive industries, such as the semiconductor, electronic, automotive, and pharmaceutical industries, capital expenditures often constitute about one quarter of total revenue and roughly two thirds of manufacturing costs. Furthermore, a business environment with volatile market demand, high penalties for unmet demand, long capacity delivery leadtimes, and rigid capital procurement contracts often leads firms in these industries to over-purchase capital equipment. To reduce this inefficiency, which can amount to losses of hundreds of millions of dollars per annum, we propose a dual-mode equipment procurement model. In this model, the equipment supplier provides two delivery modes to the firm: a base mode that is less expensive but slower and a flexible mode that is faster but more expensive. The combination of these two modes provides the firm the much-needed flexibility to mitigate demand risk at a potentially lower cost.

The model consists of three layers: a strategic contract negotiation layer, in which the firm chooses the best combination of leadtime and price for each mode from the supply contract menu; a tactical reservation layer, in which the firm reserves total equipment procurement quantities from the two supply modes by paying the supplier a reservation fee up front before the planning horizon starts; and an operational execution layer, in which the firm acquires the latest demand information in each period and orders equipment from the two supply modes, given that the total order quantities until that point cannot exceed previously determined reservation levels. With a comprehensive numerical analysis, we quantify the value of the added flexibility of dual-mode equipment procurement for the firm. Although motivated by and validated at Intel, the dual-mode equipment procurement model is versatile enough to be adapted by firms in other capital-intensive industries.

Key words: capital-intensive industries, equipment procurement, capacity expansion, forecast revision, dual-mode procurement, dual sourcing, options contracts, heuristic.
1. The Motivation

Capacity planning is a complex balancing act, especially in the semiconductor industry. This industry is one of the most capital-intensive industries in the world; a single piece of semiconductor manufacturing equipment commonly costs tens of millions of dollars. For example, over the last five years Intel has spent on average 5.3 billion USD annually on capital additions to its property, plant, and equipment (Intel Corporation 2009). Compounding the problem of high costs are the long leadtimes and the volatile consumer market. The order-to-production cycle for semiconductor manufacturing equipment can take up to 16 months, which exacerbates the difficulty of forecasting demand accurately. This paper addresses the challenges of capacity planning in the semiconductor industry and describes Intel’s efforts to address these challenges by continuously improving the set of rules for its engagement with its suppliers.

In the semiconductor industry, the marginal cost of unmet demand is considered to be significantly higher than the marginal cost of idle capacity (Fleckenstein 2004). Thus, despite the astounding costs, semiconductor firms often err on the side of having excess capacity and keep some equipment idle in order to not lose customer goodwill and loyalty. As an example, Figure 1 shows Intel’s historical capacity purchases for six consecutive process technologies (denoted by $T_{-5}$ through $T_0$). The shaded areas in the figure display the difference between purchased capacity and realized demand at the peak of the demand curve. Consistently, all six generations suffered from excess capacity, the value of which is estimated to be several hundred million USD in capital depreciation per technology. Clearly, there are opportunities to improve this process.

Traditionally, an easy way for firms with strong bargaining power to partially mitigate this capacity risk and avoid purchasing too much excess equipment has been through soft orders. To secure procurement contracts, suppliers have allowed firms to over-order capacity at the beginning of the planning horizon and to cancel some of the excess equipment without paying high penalties when more precise demand information becomes available, a process similar to the “phantom ordering” common in personal computer and electronics industry (Lee et al. 1997, Cohen et al. 2003). As a result, equipment suppliers have carried most of the demand risk. However, this type of
relationship is no longer sustainable. To keep pace with the technology requirements of maintaining Moore’s Law (Moore 1965), the cost of capital equipment in the semiconductor industry has been steadily rising with no end in sight. Building a fab now costs in excess of 5 billion USD, up from 6 million USD in 1970 and around 2 billion USD in 2001 (Kanellos 2003). Hence, soft orders are costly to suppliers. In addition, the recent trend of supply consolidation (Armbrust 2009) has further increased suppliers’ bargaining power and as a result, suppliers are becoming reluctant to carry demand risk through soft orders. Thus, the challenge for a semiconductor manufacturing firm is to derive innovative ways to order the right amount of equipment at the right time and price in advance of the demand realization to minimize unnecessary equipment purchases while maintaining a high level of service. Further, this goal should be achieved without pushing all of the risk onto suppliers.

The industry leader Intel and its suppliers are always working to refine the set of rules for their engagement. At Intel, the continuous improvement of capital acquisition and installation processes is an ongoing corporate priority. To be able to respond rapidly to changing customer demand for products while minimizing unnecessary capital expenditures is vitally important to both the semiconductor market and the firm itself. From Intel’s perspective, demand uncertainties due to extremely long procurement leadtimes add too much idle capacity to the system and jeopardize the agility of the supply chain. Intel prefers to have tighter control of its capital supply chain by shortening equipment procurement leadtimes and improving the accuracy of the demand forecast.
used in capacity planning. In return for this flexibility, Intel is willing to take on some risk from its suppliers and is considering risk-sharing mechanisms. In this application paper, we propose a dual-mode equipment procurement (DMEP) model, which effectively addresses the above concerns by incorporating a fast supply mode (in addition to the regular procurement mode), a forecast revision mechanism, and a capacity reservation procedure. In designing DMEP, our goal is to guide Intel during the phases of equipment procurement. In particular, we are interested in the following questions: How can Intel quickly evaluate different flexible options during contract negotiations? Under what circumstances does the flexible mode create value for the firm? How much of the total capacity should be reserved through the flexible mode? When should this capacity be exercised?

The rest of the paper is organized as follows: in Section 2, we briefly introduce the dual-mode equipment procurement model, which consists of three layers: a strategic contract negotiation layer, a tactical reservation layer, and an operational execution layer. Although motivated by Intel’s continuous improvement efforts on equipment procurement, this model is versatile enough to be adapted by firms in other capital-intensive industries, such as the pharmaceutical and automotive industries. Section 3 provides a brief literature review for each layer of the model. In particular, Peng and Erhun (2010) show that the execution layer of this model becomes intractable even in the simplest settings. Hence, we construct a heuristic approach for this problem in Section 4. In Section 5, we implement a detailed numerical analysis to quantify the value of the added flexibility that the dual-mode equipment procurement model provides. Section 6 concludes the paper and provides managerial insights. All proofs are presented in the appendix. Throughout the paper, we use the terms increasing and decreasing in the weak sense; i.e., including equalities.

2. The Business Problem

As discussed in Section 1, Intel’s goal is to continuously improve its equipment procurement strategies to support customer demand without over-purchasing costly capacity. Such a strategy should also embody fairness to its suppliers through a risk-sharing mechanism. To achieve this goal, we introduce the dual-mode equipment procurement (DMEP) model.
DMEP enables a firm to procure equipment from one supplier using two supply modes with complementary leadtimes and prices: a base mode (which is the regular procurement mode) that is less expensive but has a longer procurement leadtime $L_b$ and a flexible mode that is more expensive but has a shorter procurement leadtime $L_f$. The introduction of the flexible supply mode allows Intel to learn more about demand before committing to capacity. In return, to share the risk with its suppliers, Intel makes an up-front payment to secure certain base and flexible capacity levels ahead of time and faces prohibitive costs to cancel an order once it has been placed. As such, DMEP captures different phases of the relationship between Intel and its suppliers and is composed of three stages: contract negotiation, capacity reservation, and procurement execution (Figure 2).

Figure 2  The Dual-Mode Equipment Procurement Model

During the contract negotiation stage, several years before the adoption of a new process and the production of the necessary equipment, Intel and its supplier agree on the parameters of the supply modes. That is, they negotiate the leadtimes and prices (a reservation price and an execution price) of the base and flexible modes. During the reservation stage, several quarters before the planning horizon starts, Intel reserves the total equipment procurement quantities $B^T$ and $F^T$ from the base
and flexible supply modes at unit reservation prices $R_b$ and $R_f$, respectively. These reservation quantities act as an upper bound on the total order volumes from the two supply modes in all subsequent periods. During the execution stage, in each period $n$, Intel orders specific amounts of equipment $B_{n+L_b}$ and $F_{n+L_f}$ (the subscript denotes the period when the ordered equipment arrives) from the two supply modes at unit execution prices $c_b$ and $c_f$, respectively, given the latest demand forecast as well as the realized demand information from the market.

With its three stages, DMEP evaluates various trade-offs inherent in this setting. First, the contract negotiation stage balances accurate demand information with sourcing from expensive modes. Procuring from the flexible mode allows Intel to react to the market condition later with better demand information; however, the flexible mode is costlier than the base mode. Furthermore, Intel pays more for shorter procurement leadtimes, and choosing the right leadtime-cost combination is crucial. Secondly, the reservation stage balances flexibility (by reserving adequate capacity) with high reservation costs. By having a large reservation pool up front, Intel can guarantee the punctual delivery of its future orders; however, the reservation payment implies a high opportunity cost if the equipment is not ordered. Finally, the execution stage balances capacity holding costs with demand backordering costs. Early equipment procurement reduces the potential risk of having unsatisfied demand; however, it also leads to a higher capacity holding cost and a high opportunity cost if the equipment remains idle.

The goal of DMEP is to choose the appropriate leadtime-cost combinations, reservation levels, and execution quantities to maximize the expected total profit for Intel across the entire planning horizon. It is important that DMEP achieves its goals without pushing all of the risk onto the supplier. As such, the reservation payment enables the supplier to obtain capital to prepare its own production capacity and to help the supplier provide Intel with the guaranteed flexibility of ordering equipment when needed. That is, the reservation stage allows Intel to share risk with its supplier. Through the availability of a shorter leadtime mode, DMEP allows Intel to have tighter control of its capital supply chain. Through up front payments and commitments, DMEP is more fair to Intel’s suppliers.
3. Literature Review

The contract negotiation stage of DMEP builds on the literature on supply contracts. We refer the reader to Cachon (2003). The contract that we study is closely related to the one analyzed in Yazlali and Erhun (2010). The authors investigate a dual-supply contract with minimum order quantity and maximum capacity restrictions. Their contract provides supply chain partners an enhanced mechanism to share and manage demand uncertainty in the context of inventory management. We adopt a similar structure here for capacity procurement. The negotiation of price-leadtime combinations is also related to the literature on pricing and leadtime quotation. This literature often assumes that the buyer’s order quantity is a function of the price and leadtime quoted by the supplier and thus focuses on the supplier’s optimal price-leadtime decision (Palaka et al. 1998; Liu et al. 2007). We, on the other hand, take the buyer’s perspective and provide the buyer a decision-support tool that it can use while negotiating the procurement price and leadtime with its supplier.

The reservation stage addresses a capacity expansion problem. The capacity expansion problem under single-sourcing has been extensively studied in the literature; see Van Mieghem (2003) for a review of this literature. Wu et al. (2005) provides a thorough review of the literature on capacity planning problems in the high-tech industry. DMEP is also closely related to the literature on procurement and options contracts (e.g., Bassok and Anupindi 1997; Bassok et al. 1999; Vaidyanathan et al. 2005). In particular, the reservation stage of DMEP builds on the paper by Vaidyanathan et al. (2005), which focuses on the capacity contracts at Intel and discusses capacity options to better enable the factory ramps. Thus, we extend and merge these two streams of research by studying capacity expansion and option contracts in a dual-mode setting.

The execution stage of DMEP is closely related to the dual-sourcing problem, except that here the firm is sourcing from two service modes of the same supplier, rather than from two different suppliers. The dual-sourcing problem has been studied extensively in the context of inventory since the early 1960s (Daniel 1963; Fukuda 1964; Whittemore and Saunders 1977; Scheller-Wolf and Tayur 1998; Feng et al. 2006; Yazlali and Erhun 2007, 2009). Dual-source inventory management
has also been commonly adopted as an operational risk hedging strategy by firms in different industries for many years; e.g., Mattel (Johnson 2005) and HP (Billington and Johnson 2002). Despite the existence of literature on the dual-sourcing inventory problem, research concerning the dual-sourcing capacity procurement problem is scarce, with the exception of two recent papers (Chao et al. 2009; Peng and Erhun 2010). Peng and Erhun (2010) show that capacity behaves very differently from inventory when backordering is allowed and that classical inventory management policies do not apply to capacity planning problems. The authors demonstrate that the capacity problem is more complicated than its inventory counterpart, even in simple settings. Hence, the authors call for heuristic solutions for this complex problem.

By utilizing the concepts of capacity expansion and dual-sourcing, we build a heuristic solution for DMEP. Using an open-loop algorithm, the heuristic considers the essential trade-offs and balances the solution accuracy with the computing efficiency. We describe this heuristic in detail in the next section.

4. The Dual-Mode Equipment Procurement Heuristic

As discussed in Section 3, theoretically, the equipment procurement decision that the firm faces fits in the scope of a dual-source capacity expansion problem and can be formulated as a dynamic programming model. In practice, however, this may not be the best approach to take for two reasons. First, this is a complex problem and the dynamic programming model is rather difficult to solve. Even for the simplest setting where the leadtimes are consecutive with $L_b = 1$ and $L_f = 0$, Peng and Erhun (2010) demonstrate that the dual-source capacity expansion problem with backorders does not have a well-behaved optimal policy. The general case with nonconsecutive leadtimes is even more complex. This complexity not only jeopardizes the possibility of finding any structural policy, but also sharply reduces the computational efficiency due to the famous curse of dimensionality. Secondly, although the dynamic programming model may help us identify the optimal equipment procurement strategy, it comes with strict assumptions, such as fixed and known distributions of all uncertain factors and consecutive leadtimes, which can hardly be justified in
practice. Therefore, to avoid the above restrictions, we propose an open-loop simulation model as a heuristic approach. The goal of this approach is to provide Intel a fast and accurate decision-support tool with what-if capabilities that will guide the firm in answering the three questions we pose in Section 1.

To map the structure presented in Figure 2, the DMEP heuristic algorithm consists of the same three layers: the outermost is the contract negotiation layer, which identifies the indifference curves of leadtime and price combinations so that the firm can pick the best alternative from the contract menu; the middle is the reservation layer, which calculates the optimal equipment reservation quantities $B_T$ and $F_T$ for the two supply modes; and the innermost is the execution layer, which determines the equipment order quantities $B_{n+L_h}$ and $F_{n+L_f}$ from the two supply modes in period $n$. Before elaborating on each of these three layers in more detail, we first consider the demand forecast revision process, which is one of the main drivers of the problem.

4.1 The Forecast Revision Mechanism

One of the major factors that complicate the equipment procurement decision is demand volatility. If demand in each period were deterministic, it would be trivial to choose the optimal equipment order quantity for each period and match the production capacity with demand. When demand is highly uncertain, however, we need a detailed demand model which can describe not only the variance associated with each period’s demand, but also the process by which the firm continuously adjusts its anticipation of the future demand distribution based on the available information. In this section, we describe such a forecast revision mechanism.

Let demand in period $n$ be $D_n$ with a mean of $\mu_n$ and variance of $\sigma_n$. The firm has a forecast for $D_n$ in each period before $n$ and revises its forecast based on two evolution processes. As time progresses, the forecast of $\mu_n$ is updated; we call this process mean forecast evolution. As time progresses, the variance $\sigma_n$ associated with the forecast is also updated; we call this process forecast variance evolution. Specifically, we use $\mu^m_n$ and $\sigma^m_n$ to represent the mean and variance of the forecasts made in period $m$ for the corresponding values in period $n$ ($m < n$). We assume
Figure 3 Graphical Demonstration of the Forecast Revision Mechanism

(a) Forecast for the anchor period’s demand made in all previous periods
(b) Forecasts for all future periods’ demand made in the anchor period

that the mean forecast evolution follows a Markovian process (Hausman 1969, Heath and Jackson 1994) and that the variance forecast evolution is deterministic and uniquely determined by the forecasting leadtime. These two assumptions greatly simplify the problem by reducing the amount of information we need to carry for each period when we determine the capacity reservation levels at the beginning of the planning horizon. These assumptions are also largely consistent with the historical data collected at Intel.

Figure 3 graphically describes the demand forecast updating process, where period 0 is chosen as an anchor period and does not necessarily correspond to the starting period of the demand ramp. Figure 3(a) shows how forecasts for period-0 demand $\mu_0^m$ and $\sigma_0^m$ ($m \leq 0$) evolve as we approach period 0. The solid curve represents the mean forecast evolution path, and the interval between the dashed curves denotes the variance range, which shrinks as the forecasting leadtime decreases. That is, the forecast accuracy for a certain period’s demand improves as one moves closer to that period in time. Figure 3(b) depicts the forecasts $\mu_0^n$ and $\sigma_0^n$ ($n \geq 0$) made in period 0 for all future periods’ demand: the solid curve represents the mean forecast and the dashed interval denotes the variance range, which diverges as the forecasting leadtime increases. That is, the forecasting accuracy decreases as one forecasts further into the future.

Next, we explain the DMEP heuristic in greater detail. We first discuss the execution layer, then
the reservation layer and the contract negotiation layer, since the former is a fundamental building block for the others.

4.2 The Execution Problem

The execution module is the core of the DMEP heuristic. Given the reservation quantities $B^T$ and $F^T$, the execution module characterizes how the firm should place the base and the flexible orders in each period. At the beginning of each period, the firm obtains the latest demand realization and forecast updates for all future periods. Based on this information, all previously placed base and flexible orders, and the backorder quantity from the preceding period, the open-loop execution algorithm calculates the myopic optimal base and flexible order quantities for the remaining periods.

Figure 4 displays the timeline of the execution problem. The selling season of the product is $N$ periods and starts in period 1. To prepare for the demand ramp, the firm starts placing orders $L_b$ periods before the first demand realization and stops placing orders $L_f$ periods before the end of the selling season. Therefore, the length of the planning horizon is $N + L_b - L_f$ periods. The firm orders only from the base mode in periods $1 - L_b, \ldots, -L_f$; it orders from both modes in periods $1 - L_f, \ldots, N - L_b$; and it orders only from the flexible mode in periods $N - L_b + 1, \ldots, N - L_f$. In periods $N - L_f + 1$ to $N$, the firm simply satisfies the demand with the existing capacity.

Specifically, in period $n$ of the planning horizon ($n = 1 - L_b, 2 - L_b, \ldots, N - L_f$), the execution module solves the following stochastic optimization problem:

$$\max_{\vec{B}_1, \ldots, \vec{B}_N; \vec{F}_1, \ldots, \vec{F}_N} \mathbb{E}_{d_1, \ldots, d_N} \left[ \sum_{i=1}^{N} \delta^i \left( p_i s_i - c_b B_i - c_f F_i - c_h k_i \right) - \delta^{N+1} c_u d_{rem} \right]$$  \hspace{1cm} (1)
subject to: For each realized sample path \( d_{1\vee(n+1)}, \ldots, d_N \)

\[
s_i = \min\{k_i, d_i + d_i^{rem}\} \quad \text{for } i = 1, \ldots, N
\]

(2)

\[
s_i \geq \psi d_i \quad \text{for } i = 1 \vee (n + L_f), \ldots, N
\]

(3)

\[
k_i = k_{i-1} + B_i + F_i, \quad \text{for } i = 1, \ldots, N \text{ with } k_0 = 0
\]

(4)

\[
d_i^{rem} = (d_{i-1} + d_i^{rem} - k_{i-1})^+, \quad \text{for } i = 1, \ldots, N \text{ with } d_1^{rem} = 0
\]

(5)

\[
\sum_{i=1}^{N} B_i \leq B^T; \quad \sum_{i=1}^{N} F_i \leq F^T
\]

(6)

\[
\bar{B}_{1, \ldots, N} \geq 0; \quad \bar{F}_{1, \ldots, N} \geq 0
\]

(7)

\[
\bar{B}_{1, \ldots, n-1+L_b} = \bar{B}_{1, \ldots, n-1+L_b}; \quad \bar{F}_{1, \ldots, n-1+L_f} = \bar{F}_{1, \ldots, n-1+L_f}
\]

(8)

We define \( x \vee y := \max(x, y) \) and use the notation \( \bar{X}_{1,2,\ldots,m} \) to represent the vector \( [X_1, X_2, \ldots, X_m]^T \). Additionally, \( p_i \) is the profit margin for period \( i \); \( c_b \) the base mode execution price; \( c_f \) the flexible mode execution price; \( c_h \) the unit holding cost; \( c_u \) the unit penalty cost for unmet demand after the planning horizon ends; \( d_i \) the incoming demand in period \( i \); \( d_i^{rem} \) the unsatisfied demand remaining from period \( i - 1 \); \( s_i \) the actual sales quantity during period \( i \); \( k_i \) the cumulative capacity position at period \( i \); \( \psi \) the service level constraint; and \( \delta \) the discount factor. \( B_i^* \) and \( F_i^* \) represent the optimal decisions that were already executed in previous periods. We note that the subscripts used for order quantities denote the period when the ordered equipment arrives.

The objective function maximizes the expected profit by considering the profit margin, equipment procurement costs, and inventory-related costs. The expectation is taken over all future demand and the execution cost is calculated when the orders arrive. We note that as our goal is to understand the strategic and tactical level capacity decisions, we suppress the operational level inventory problem and assume that production in a period will not exceed demand. Although restrictive, this assumption is prevalent in the literature (see Chao et al. 2009) and is in line with the goals of DMEP. Constraint (2) guarantees that sales in a given period cannot be larger than the demand or the supply; also, the firm should satisfy the service level constraint in each period (constraint
Constraints (4) and (5) enable state transitions for the capacity and backorders, respectively. Constraints (6) and (7) guarantee that the base and flexible orders are within their reservation limits and nonnegative. Finally, constraint (8) freezes the already executed orders before finding the optimal order quantities. One crucial feature of this rolling-horizon algorithm is that, in period \( n \), only the orders \( B_{n+L_b}^* \) and \( F_{n+L_f}^* \) (if \( n + L_f \geq 1 \)) will actually be placed, and the rest of the decisions will be postponed to future periods.

To reflect the industry practice that we are investigating, we make several modeling assumptions. First, we analyze a setting where the unmet demand during a certain period is backordered (instead of lost, which is assumed in most of the capacity planning literature). Instead of defining a unit penalty cost per period, we penalize the backordered demand in two ways: (i) at the end of the planning horizon, any remaining unsatisfied demand incurs a terminal penalty; and (ii) the profit margin of the product is decreasing over time, which implies that there is a loss of revenue associated with backordering. Therefore, satisfying the demand earlier is always preferable if it is possible. Secondly, the on-hand equipment capacity incurs a unit holding cost, which may include the opportunity cost of investment or costs such as a utility fee, maintenance expenditure, or cost of floor space. Thirdly, investment in the equipment capacity is irreversible; i.e., capacity contraction is not allowed. Fourthly, the firm is risk-neutral and maximizes its expected profit. Finally, the raw material inventory is always sufficient for the production process, and we only concentrate on the equipment procurement decisions.

To illustrate the firm’s decision-making process more clearly, we present the execution heuristic algorithm using a concise tabular format in Figure 5. We assume that there are 6 periods in the selling season, the base mode leadtime is 4 periods, and the flexible mode leadtime is 2 periods. The horizontal axis in the table represents the entire selling season from period 1 to period 6. Each horizontal group below the top line contains the demand information and decision profiles corresponding to a decision period, which is labeled on the vertical planning horizon axis.

At the beginning of the planning horizon, period –3, the firm obtains the latest demand forecast information (\( \mu \)’s and \( \sigma \)’s) for period 1 to period 6. The algorithm runs the aforementioned stochastic
Table 1

<table>
<thead>
<tr>
<th>selling season</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>planning horizon</td>
<td>-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B1*</td>
<td>B2</td>
<td>B3</td>
<td>B4</td>
<td>B5</td>
<td>B6</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>F2</td>
<td>F3</td>
<td>F4</td>
<td>F5</td>
<td>F6</td>
</tr>
<tr>
<td></td>
<td>-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>B2*</td>
<td>B3</td>
<td>B4</td>
<td>B5</td>
<td>B6</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>F2</td>
<td>F3</td>
<td>F4</td>
<td>F5</td>
<td>F6</td>
</tr>
<tr>
<td></td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>B2</td>
<td>B3*</td>
<td>B4</td>
<td>B5</td>
<td>B6</td>
</tr>
<tr>
<td></td>
<td>F1*</td>
<td>F2</td>
<td>F3</td>
<td>F4</td>
<td>F5</td>
<td>F6</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>B2</td>
<td>B3</td>
<td>B4*</td>
<td>B5</td>
<td>B6</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>F2*</td>
<td>F3</td>
<td>F4</td>
<td>F5</td>
<td>F6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>B2</td>
<td>B3</td>
<td>B4</td>
<td>B5*</td>
<td>B6</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>F2</td>
<td>F3*</td>
<td>F4</td>
<td>F5</td>
<td>F6</td>
</tr>
</tbody>
</table>

Figure 5 A Tabular Demonstration of the Execution Heuristic. \( L_b = 4, L_f = 2 \); \( \sigma_n \) denotes the forecast variance corresponding to a forecasting leadtime of \( n \) periods; * current period decision; - previous decisions; grey decisions not to be executed; and \( Dmd \) realized demand.
consideration when calculating the total expected profit.

We solve the stochastic program above using Monte Carlo sampling. We are able to consider a large number of samples, which improves the accuracy of the solution, because the problem is easy to execute:

**Proposition 1.** Assuming that demand $d_n$ is discrete and takes finitely many values, the above stochastic programming model can be converted into a linear programming model.

The execution algorithm provides a handy roadmap that indicates when and how much to order from the two supply modes. It captures the evolution of demand information and enables last-minute decision-making. As opposed to the restricted focus of some rule-of-thumb approaches such as executing flexible orders only during the peak period, the necessity of flexible mode is evaluated in each period of the planning horizon. On the flip side, the execution algorithm is myopic in the sense that it finds the “optimal” ordering scheme based on the current best information only, without considering the opportunity to make contingent decisions based on the actual realized demand at each stage. Fortunately, this disadvantage of myopia is greatly mitigated by the rolling-horizon nature of the algorithm.

### 4.3 The Reservation Problem

As the equipment suppliers gain more bargaining power due to the trend of supply consolidation, simply letting the supplier bear the major procurement risk is no longer sustainable. The reservation procedure of the DMEP heuristic, therefore, functions as a mechanism for risk-sharing between the firm and its equipment supplier. The firm, by paying an up front reservation fee, shares the risk of capacity building and installation with the supplier and enjoys the guaranteed delivery of equipment in return.

Determining how much capacity to reserve from the two supply modes, especially the flexible mode, is based on a trade-off between the reservation cost and the potential benefits from the guaranteed flexibility. Specifically, the optimal reservation quantities $B^T$ and $F^T$ are determined according to a scenario analysis of the potential future demand profiles. Intuitively, if the future
demand scenario involves no uncertainty, then flexibility has no value; it is never optimal to order from the expensive flexible mode. In contrast, if the future demand scenario is highly uncertain and the demand mean forecast is very likely to be modified during the updating process, then flexibility has a high value and we should expect a higher flexible reservation level. In general, the reservation quantities maximize the expected horizon-wide profit over all possible demand scenarios.

More precisely, the reservation algorithm determines the optimal $B_T$ and $F_T$ based on a Monte Carlo simulation performed on the mean forecast evolution trajectories. It generates mean forecast evolution paths according to the forecast revision mechanism introduced in Section 4.1. Assuming that the forecast for the demand in different periods evolves independently and possesses the Markovian property, we then have

$$P(\mu_1^m, \cdots, \mu_N^m | \mu_1^{m-1}, \mu_1^{m-2}, \cdots, \mu_N^{m-1}, \mu_1^{m-2}, \cdots) = P(\mu_1^m, \cdots, \mu_N^m | \mu_1^{m-1}, \cdots, \mu_N^{m-1}), \text{ for } m < 1 \quad (9)$$

where $P(\cdot)$ is the probability mass function if $\mu$ takes discrete values and the probability density function if $\mu$ takes continuous values. Therefore, at the beginning of the planning horizon, given the initial demand forecast profile for the entire horizon and $P(\cdot)$, the algorithm enumerates a large number of possible mean forecast evolution paths. For each path it calls the execution module to calculate the specific order quantities as well as the expected horizon-wide profit. The algorithm then chooses the reservation quantities $B_T$ and $F_T$ that maximize the average total profit across the entire planning horizon.

Equation (10) formulates the reservation problem mathematically. We first denote the optimal value function of the period-$n$ execution problem (1)-(8) as $J_n(B_T, F_T, \bar{\mu}_n, \bar{\sigma}_n)$ by decomposing the demand information $d$ into its two components $\mu$ and $\sigma$. We choose $B_T$ and $F_T$ to maximize the expected horizon-wide profit $J_{N-L_f}$ since period $N-L_f$ is the last period during which a decision can be made. The stochastic optimization is given as

$$\max_{B_T \geq 0, F_T \geq 0} \mathbb{E}_M \left[ J_{N-L_f}(B_T, F_T, M, \Sigma | \bar{\mu}_{1-L_f}) \right] - R_B B_T - R_f F_T \quad (10)$$
where the expectation is taken with respect to the mean forecast evolution space $M$:

$$M = \{\bar{\mu}_n : \bar{\mu}_n = (\mu_{1\vee(n+1)}, \ldots, \mu_N^n), \ n = 1 - L_b, \ldots, N - L_f \}.$$ 

As we expressed in equation (9), the mean forecast profile $\bar{\mu}_n$ follows a Markov process with initial state $\bar{\mu}_{1-L_b}$ and transition probability space $P(\cdot)$, such that a particular realization of $M$ occurs with probability

$$P(M|\bar{\mu}_{1-L_b}) = P(\bar{\mu}_{2-L_b}|\bar{\mu}_{1-L_b})P(\bar{\mu}_{3-L_b}|\bar{\mu}_{2-L_b}) \cdots P(\bar{\mu}_{N-L_f}|\bar{\mu}_{N-1-L_f}).$$

Finally, the variance forecast evolution space is

$$\Sigma = \{\bar{\sigma}_n : \bar{\sigma}_n = (\sigma_{1\vee(n+1)}, \ldots, \sigma_N^n), \ n = 1 - L_b, \ldots, N - L_f \}; \quad (11)$$

$\Sigma$ is actually deterministic based on our assumption that the forecasting variance is uniquely decided by the forecasting leadtime. Proposition 2 below establishes the concavity and coerciveness of the reservation problem and thus the existence of finite optimal reservation quantities for the two supply modes. Hence, the reservation problem can be solved using either an optimization software or a simple search algorithm.

**Proposition 2.** The objective function in (10) is concave and coercive in $(B^T, F^T)$.

The reservation algorithm helps the firm determine the optimal amount of equipment to reserve for both supply modes. It builds on the philosophy of scenario analysis and chooses the reservation quantities that guarantee the maximum expected return to the firm. By adjusting the mean forecast transition probability $P(\cdot)$, we can easily create different demand scenarios; hence the reservation algorithm is applicable to a wide range of business settings.

### 4.4 Contract Negotiation Problem

During the contract negotiation stage, the firm and the equipment supplier determine the leadtimes $(L_b, L_f)$ of the two delivery modes as well as the unit reservation $(R_b, R_f)$ and execution $(c_b, c_f)$ prices associated with these leadtimes. Our algorithm involves a simple sensitivity analysis: for
different \((\text{leadtime}, \text{price})\) combinations, we run the reservation and execution heuristic and obtain the corresponding expected horizon-wide profits. The decision-maker can then choose the \((\text{leadtime}, \text{price})\) pair from the contract menu that leads to the highest expected return for the firm.

The strength of this part of the heuristic is that it helps the firm make the optimal strategic level decision by considering potential tactical and operational level contingencies. Thus, the three stages of DMEP constitute a stable decision-support pyramid, in which the decisions are made in a top-down sequence while the underlying algorithm follows an embedded bottom-up order.

5. DMEP as a Decision-Support Tool

In this section, we revisit the questions that we asked in Section 1 and illustrate our approach to them with numerical examples. We first provide the parameter values that we use for the numerical examples. We then explore the value of DMEP as a decision-support tool with an emphasis on these three questions. We implemented the DMEP heuristic using the convex optimization tool CVX (http://cvxr.com/cvx/) in Matlab.

5.1 Parameter Values for the Numerical Examples

The parameter values we use in our examples are based upon the semiconductor business environment. The values presented here are either publicly available or have been normalized to protect the firm. For all numerical samples, the selling season is \(N = 6\) quarters. The profit margin (including the equipment cost) of a single chip is \(P_0 = 33.75\) before the first quarter of the life cycle starts; the profit margin \(P_t\) in quarter \(t\) satisfies an exponential decreasing formula: \(P_t = P_0 e^{-\alpha t}\) (Leachman 2007) where the coefficient \(\alpha\) is equal to 0.23, roughly implying that the margin decreases by 50% per year. The unit penalty cost for unmet demand at the end of the selling season is \(C_p = 50\). The service level constraint is 95%. Each piece of equipment can process 12,000 wafers per quarter, and each wafer can be further sawed into 1,425 chips. The discount factor per quarter is \(\delta = 0.96\).

In this setting, the leadtime and price for the base mode are usually fixed. However, suppliers offer a menu of contracts for the flexible mode, which commonly includes the available flexible leadtime options as well as the price associated with each leadtime. The shorter the flexible leadtime, the
higher the flexible price. Using this fact, we set the base leadtime $L_b$ to 4 quarters and the total equipment price associated with the base mode $P^b_e$ to $25$ million. We impose a minimum base reservation price of $R_b = $10,000 to eliminate the trivial case of infinite base reservation quantity. The base execution price is thus given as $C_b = P^b_e - R_b$. We choose equipment with a long leadtime as these equipments are the bottleneck in capacity planning and are usually the ones which are very expensive. Hence they are difficult to manage and therefore the target product of DMEP. The price associated with the flexible mode is determined by two parameters: the equipment price increase ratio $\theta$ and the flexible reservation price ratio $\lambda$. Namely, the total equipment price of the flexible mode is $P^f_e = \theta P^b_e$ where $\theta \geq 1$ since a faster mode implies a shorter preparation leadtime for the supplier and hence a higher supply cost. The reservation price of the flexible mode is $R_f = \lambda P^f_e$. The remaining $(1 - \lambda)$ portion together with a fast shipment premium of $50,000$ is paid as the flexible execution price; that is, $C_f = (1 - \lambda)P^f_e + 50,000$.

For the forecast revision process, we make three additional assumptions. First, demand in each period is assumed to be truncated normal. Secondly, we model the mean forecast evolution process with a mean-adjustment factor $\beta$ that can be tailored to each firm’s own forecasting history. We assume that in each period the mean forecast $\mu_n$ for period $n$’s demand may increase by $\beta\%$, decrease by $\beta\%$, or remain the same, with equal probability of $1/3$. The mean forecasts for demand in different periods evolve independently. Finally, we overload $\sigma^m_n$ to represent the coefficient of variation and assume it is increasing in the forecast leadtime; i.e., $\sigma^m_n = \gamma \times 0.01(n - m)$. We note that $\beta$ controls the solid path in Figure 3 and $\gamma$ controls the width of the dashed variance interval.

Table 1 displays the initial demand forecast. The initial forecast is symmetric with periods 3 and 4 being peaks. Using this initial forecast, we study three different forecast scenarios. In the first scenario, the initial forecast is not adjusted and the realized demand in each period matches the corresponding mean forecast. The only randomness in this system is the variance coefficient $\gamma$. Since the mean forecast evolution process is degenerative and hence $\beta = 0$ for all periods in this case, we call this scenario the stationary demand scenario. The second scenario investigates demand forecast shocks: during period $-1$ (1), the forecast for the mean demand in period 1 (3) is
adjusted upwards. The third scenario corresponds to a situation with demand realization shock: the initial mean forecast profile $\vec{\mu}_{1, \ldots, N}$ is not updated in periods $-2$, $-1$ and $0$; in period 1, however, the actual realized demand is 14,000, which is much higher than the previous mean forecast.

### Table 1 Nonstationary Demand Forecast Scenarios (unit: wafer-start-per-week)

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>period 1</th>
<th>period 2</th>
<th>period 3</th>
<th>period 4</th>
<th>period 5</th>
<th>period 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial forecast</td>
<td>$\vec{\mu}_{1, \ldots, 6}$</td>
<td>4,788</td>
<td>9,577</td>
<td>14,365</td>
<td>14,365</td>
<td>9,577</td>
</tr>
<tr>
<td>Stationary demand</td>
<td>$\vec{\mu}_{1, \ldots, 6}$</td>
<td>4,788</td>
<td>9,577</td>
<td>14,365</td>
<td>14,365</td>
<td>9,577</td>
</tr>
<tr>
<td>Forecast shock</td>
<td>$\vec{\mu}_{1, \ldots, 6}$</td>
<td>7,000</td>
<td>9,577</td>
<td>14,365</td>
<td>14,365</td>
<td>9,577</td>
</tr>
<tr>
<td></td>
<td>$\vec{\mu}_{1, \ldots, 6}$</td>
<td>6,850</td>
<td>9,577</td>
<td>19,320</td>
<td>14,365</td>
<td>9,577</td>
</tr>
<tr>
<td>Realization shock</td>
<td>$\vec{\mu}_{1, \ldots, 6}$</td>
<td>14,000</td>
<td>9,577</td>
<td>14,365</td>
<td>14,365</td>
<td>9,577</td>
</tr>
</tbody>
</table>

### 5.2 The Value of DMEP as a Decision-Support Tool

**Under what circumstances does the flexible mode create value for the firm?**

To capture the settings under which the flexible mode will create value for Intel, we first investigate the optimal equipment procurement decisions given the reservation quantities $B_T$ and $F_T$. Figure 6 illustrates the settings under which the flexible mode is utilized as a risk-hedging channel by summarizing our findings for the scenarios discussed in Table 1. The horizontal axis denotes the ratio between the flexible execution price $C_f$ and the base execution price $C_b$; the vertical axis denotes the size of the forecast shocks and/or the size of the demand realization shocks. The entire plane can be divided into three different regions based on the value of the two coordinates: If $C_f/C_b \leq 1$, then the firm procures equipment only from the flexible mode, since it is not only faster but also cheaper. If $C_f/C_b > 1$ and the forecast and realization shocks are small enough, then it is optimal to procure only from the base mode. That is, in the stationary demand scenario or when shocks are small, cost is the only critical parameter: the firm prefers to single source using the less expensive source as long as there is available reservation quantity. When there are forecast and realization shocks, however, this threshold policy ceases to apply. If $C_f/C_b > 1$ and the forecast or realization shocks surpass a certain threshold value, then both the base mode and the flexible mode are used. Now the flexible source adds value with its shorter leadtime even when it is more...
Forecast or realization shock

Base and Flexible
Flexible only
Base only

Figure 6 Impacts of Price and Demand Property on the Ordering Policy (illustration)

expensive. The firm can wait until the last minute to learn more about demand before placing orders and thus maintain an agile environment.

At the tactical capacity reservation level, Figure 7 demonstrates that as the procurement risk increases (i.e., either the mean evolution jump size $\beta$ or demand variance coefficient $\gamma$ increases), the firm relies on the flexible mode more heavily. The flexible mode is especially valuable in dealing with demand realization shocks since the firm’s contingencies are very limited in that case. Finally, Figure 8 demonstrates that as we gradually increase the service level constraint from 87.5% to 99%, the value of the flexible mode increases. Note that the expected profit associated with a 99% service level is more than 170 million dollars less than that associated with a 95% service level. This is by no means a small compromise, and this trade-off should be considered when making service level decisions.

How much of the total capacity should be reserved through the flexible mode? When should this capacity be exercised?

We note that the conditions for dual-mode procurement correspond to the business environment in which Intel operates; i.e., the forecast and realization shocks, the forecast uncertainty, and service levels are all high. Thus, dual-mode procurement should be seriously considered in this industry, which has traditionally used single-mode procurement. The notion that flexibility is only necessary for peak periods is also a misconception. Actually, the flexible mode may be optimally deployed whenever there is a forecast shock or a realization shock, both of which occur frequently during the
ramp-up stage of the product life cycle, instead of just at the peak. Although the value of flexible mode decreases as its total price (i.e., \( \theta \)) and/or reservation price (i.e., \( \lambda \)) increase (Figure 9), the firm continues to reserve more than 6% of its total capacity through the flexible mode even when the flexible mode is 60% more expensive (Figure 9(a)) or when the firm has to pay 25% up front (Figure 9(b)).

How can Intel quickly evaluate different flexible options during contract negotiations?

One efficient way for Intel to select offers from the contract menu is to compare the position of different leadtime and price combinations for the flexible mode on an iso-profit graph, where the flexible price is adjusted by two parameters: the price increase ratio \( \theta \) and the reservation price ratio \( \lambda \). Figure 10(a) demonstrates the iso-profit curves under a fixed \( \theta \) with value 1.3. The \( (L_f, \lambda) \)
pairs on each of the solid lines lead to the same expected total profit under the optimal reservation decision, while lines towards the lower-left corner correspond to higher profits than those towards the upper-right corner. Intel can utilize these curves in two ways. The curves demonstrate the dominance between different contract options: for example, contract $A$ with $L_f = 1$, $\lambda = 15\%$, and an expected profit of $11.65$ billion should be preferred to contract $B$ with $L_f = 3$, $\lambda = 9\%$, and an expected profit of $11.52$ billion. Alternatively, each curve quantifies the maximum reservation price Intel should be willing to pay for added flexibility; for example, Intel can pay up to 20\% of the total price up front and decrease the leadtime to 1 quarter while still keeping its profits at the same level as in contract $B$. Similarly, Figure 10(b) demonstrates the iso-profit curves under a fixed $\lambda$ with value 15\%. With the assistance of such iso-profit graphs, Intel will know the bottom-line impact of different alternatives while negotiating with its supplier and will make informed tradeoffs between flexibility and cost.

6. Conclusion

Capital equipment purchasing is a crucial yet difficult task for many semiconductor, electronic, automotive, and pharmaceutical firms. In this paper, we have proposed a dual-mode equipment procurement model (DMEP) to guide firms through this complex task. DMEP serves three roles. On the strategic level, it provides decision support to contract negotiation by comparing alternatives with different levels of flexibility and costs. On the tactical level, it guides capacity reservation
decisions by characterizing the amount of capacity that should be reserved from different procurement modes. On the operational level, it quantifies procurement amounts by considering the latest demand information as well as the installed capacity. By incorporating interactions between these three levels of decision making, DMEP enables a flexible supply chain that adapts effectively to changing demand conditions. It helps firms better manage their equipment procurement process, eliminate excess capacity, and thus lower their costs. It benefits suppliers by enabling a risk-sharing mechanism through up-front capacity reservation and the elimination of soft orders.

DMEP formalizes and extends the approach that Intel has used in the past to price and exercise capacity options with reduced leadtimes for a few types of fabrication equipment (Vaidyanathan et al. 2005). It also complements recent improvements in demand forecasting methodologies at Intel (Wu et al. 2010). The dual-mode procurement approach described here is being used at the strategic and tactical levels today at Intel for all types of fabrication equipment, and will soon be used at the operational level. Implementation of the approach has leveraged model structure and details to provide the types of sensitivity analysis needed by Intel decision-makers to understand and take advantage of the subtleties of this improvement in the capital equipment acquisition process. The annual savings on capital procurement at Intel due to implementing DMEP are estimated to exceed tens of millions of dollars.
Despite its many advantages, DMEP also has several limitations. First, we investigate a situation in which the firm procures only one type of equipment from its supplier. This simplification enables us to demonstrate the dynamics of the algorithm without introducing complexity. The DMEP model can be generalized to a multi-equipment scenario in which (1) the firm considers ordering from the base and flexible modes for all types of equipment with different leadtimes in each period and (2) the available capacity in each period is constrained by the lowest capacity among all the types of equipment. As expected, as the number of types of equipment increases, the interactions and the complexity of the problem also increase. Having said that, firms should consider DMEP only for equipment on the critical path of capacity planning as the rest (in Intel’s case the types that have shorter leadtimes and are cheaper) will not impose additional constraints on the system. Yet, alternative formulations for multi-equipment procurement would be valuable. These alternative formulations should consider the multi-tool problem as a portfolio of tools, suppliers, and potentially approaches other than DMEP. Secondly, DMEP is constructed to help with the equipment procurement decision in capital-intensive industries. However, the logic of DMEP can be transplanted to the inventory management of raw material and components. Especially for the equally complex dual-source inventory control problem with non-consecutive leadtimes, DMEP can certainly be applied with minor modifications, such as adapting the state-transition equation for the cumulative capacity to denote depletable inventory. However, the model should be validated and the parameters need to be adjusted for this inventory setting. Finally, we formulate the procurement problem as a linear program. As such, DMEP is better suited to providing the fraction of orders from each mode rather than the actual procurement quantities. However, if the intention is to use DMEP to obtain the specific order quantities, generalizing the DMEP model to an integer program would be preferable.

This paper takes an initial, yet important, step toward formulating the complex dual-mode equipment procurement problem in a practical and insightful way. In capital-intensive industries, capital expenditures often constitute about one quarter of the total revenue and roughly two thirds of the manufacturing costs. Given the millions of dollars that are at stake, we believe that equipment
procurement problems will attract more attention from academia. Fortunately, there is plenty of room for interesting future work around this topic.

Appendix A: Proofs

Proof of Proposition 1: We claim that the equivalent linear program is given in the following form (for ease of exhibition, we keep the expectation in the objective function instead of explicitly using the discrete probability distribution of the demand; we also do not display the decision variables $\vec{s}_1, \ldots, N$, $d^+_{2, \ldots, N+1}$, and $d^-_{2, \ldots, N+1}$ under the maximization operator):

$$\max_{\vec{B}_1, \ldots, N; \vec{F}_1, \ldots, N} \mathbb{E} d_{1V(n+1)}, \ldots, d_N \left[ \sum_{i=1}^{N} \delta^i \{ p_i s_i - c_i B_i - c_f F_i - c_h k_i \} - \delta_{N+1} c_u d^+_{N+1} \right]$$

subject to

$$s_i \leq k_i \text{ for } i = 1, \ldots, N$$

$$s_i \leq d_i + d^{rem,+}_i \text{ for } i = 1, \ldots, N$$

$$d^{rem,+}_i - d^{rem,-}_i = d_{i-1} + d^{rem,+}_{i-1} - k_{i-1}, \text{ with } d^{rem,+}_1 = d^{rem,-}_N = 0$$

$$d^{rem,+}_i \geq 0; \quad d^{rem,-}_i \geq 0 \text{ for } i = 1, \ldots, N+1$$

and constraints (3), (4), (6) − (8)

Comparing the above linear program with the original stochastic program, we observe that there are two major differences: (i) We have replaced the original min(·, ·) operator in constraint (2) with the two inequality constraints (12) and (13). To justify this transformation, we only need to show that at the optimal solution, either (12) or (13) will be binding. This condition is equivalent to the argument that in the optimal solution the firm has no incentive to deliberately withhold its production and backorder some demand into the next period, which is obvious since the profit margin is decreasing in time. (ii) We introduce nonnegative variables $d^{rem,+}_i$ and $d^{rem,-}_i$ in constraint (14) to remove the original $(·)^+$ operator in constraint (5); this is a common technique.

Proof of Proposition 2: From Proposition 1 we know that the objective function of the execution problem (2) is linear in decisions $\vec{B}$ and $\vec{F}$, and therefore trivially concave in $(\vec{B}, \vec{F}, B^T, F^T)$. Also notice that the constraint set is a convex set. Hence, applying the concavity preservation theorem under maximization (or convexity preservation under minimization) (Heyman and Sobel 1984, p. 525), we know that the value function of the execution problem $J^o(B^T, F^T, \hat{\mu}_n, \hat{\sigma}_n)$ (and thus the objective function of the reservation problem) is concave in $(B^T, F^T)$. To see coerciveness, notice that the objective function value of (10) goes to negative infinity as $B^T$ or $F^T$ tends to infinity, given that demand is finite.
References


[27] A.S. Whittemore and S.C. Saunders. Optimal inventory under stochastic demand with two supply 


[31] Ö. Yazlali and F. Erhun. Dual-supply inventory problem with capacity limits on order sizes and unre-

[32] Ö. Yazlali and F. Erhun. Managing demand uncertainty with dual supply contracts on capacity and 
inventory: A heuristic approach. Working paper, Department of Management Science and Engineering, 
Stanford University, Stanford, CA, 2010.
01 September 2010

Professor Steven Graves
Abraham J. Siegel Professor of Management Science,
Professor of Mechanical Engineering and Engineering Systems
Massachusetts Institute of Technology, 77 Massachusetts Avenue, E62-579
Cambridge, MA 02139-4307

Dear Professor Graves,

I am writing this letter on behalf of Intel Corporation to verify that the paper titled “Equipment Procurement Strategy with Reservation and Forecast Revision: A Heuristic Approach” has indeed been originated from an important business problem that we face at Intel Corporation. In search of constructing new set of rules for our engagement with our equipment suppliers to manage our capacity planning more effectively, we collaborated with the co-authors of the paper to develop the dual-mode equipment procurement (DMEP) model.

DMEP formalizes and extends the approach that Intel has used in the past and complements recent improvements in demand forecasting methodologies at Intel. The dual-mode procurement approach described in the paper is being used at the strategic and tactical levels today at Intel for all types of fabrication equipment, and will soon be used at the operational level. Implementation of the approach has leveraged model structure and details to provide the types of sensitivity analysis needed by Intel decision-makers to understand and take advantage of the subtleties of this improvement in the capital equipment acquisition process. The annual savings on capital procurement at Intel due to implementing DMEP are estimated to exceed tens of millions of dollars.

Yours truly,

[Signature]

Robert E. Bruck
Vice President, Technology and Manufacturing Group
General Manager, Technology Manufacturing Engineering
INTEL CORPORATION
Paper Title:
Inventory Write-downs, Sales Growth, and Ordering Policy: An Empirical Investigation

Authors:
Chad Larson, Danko Turcic, and Fuqiang Zhang

Affiliation:
Olin Business School, Washington University in St. Louis

Contact Information:
Fuqiang Zhang
Simon Hall 236
Olin Business School
Washington University in St. Louis
St. Louis, MO 63130
Tel: 314-435-9580
Email: fzhang22@wustl.edu
Inventory Write-downs, Sales Growth, and Ordering Policy: An Empirical Investigation

Abstract

If the market value of inventory declines below its original cost, for reasons such as damage or obsolescence, under U.S. Generally Accepted Accounting Principles a company should write it down to the new market value and recognize a loss. This paper uses publicly available data to conduct an empirical investigation of 290 firms, which experienced a first-time inventory write-down of more than 1% of average total assets between the years 2002 and 2004. The average inventory write-down in our sample is $13.2 million, which represents 3.7% of a firm’s average total assets. We show that these write-downs are associated with a severe negative impact on firms’ operating performance: The mean firm in our sample experiences a -15.4% return on assets in the year of an inventory write-down and a -21.6% market-adjusted return. We also examine the relationships between sales growth, purchasing policies, and inventory write-downs. Empirical evidence suggests that extreme sales growth firms are significantly more likely to experience a future inventory write-down than moderately growing firms. We find that, on average, all growing firms purchase more inventory than they sell, i.e., responding to growth, they tend to build up stock. The extreme sales growth firms, however, purchase less inventory than their moderately growing counterparts, indicating that these firms may be aware of the heightened risk of write-downs, which thus far has not been explicitly considered in the inventory literature. In addition, we find that extreme sales growth firms with write-downs build up more stock than extreme sales growth firms without write-downs. However, we do not find evidence of exuberant inventory purchases among these firms: In fact, inventory policy of an average extreme growth firm with write-downs is statistically indistinguishable from that of a firm experiencing only moderate or no sales growth. Future research may explore whether inventory policies used by moderately growing firms are inappropriate for extreme growth firms. This may lead to useful heuristics that would function well in a non-stationary demand environment with stochastic holding cost.

Key words: inventory; write-downs; obsolescence; sales growth; ordering policy
1 Introduction

According to the U.S. Federal Reserve, in September 2010, the total value of manufacturing and trade inventories reached $1.4 trillion dollars.\(^1\) On average, inventory represents approximately 15\% of total firm assets for public U.S. firms.\(^2\) Because such a significant fraction of firms’ assets are invested in inventories, managing them is important and there is empirical evidence that companies continuously seek to improve their inventory management processes and to reduce inventory levels: for example, between 1992 and 2010, the average inventory to sales ratio has steadily declined from 1.56 to 1.27.\(^3\)

The central concern in inventory management is to find the right trade-off between the cost of holding inventory and the cost of shortages. Since managing these costs commonly depends on the demand pattern, such an issue is significantly more complicated when demand is stochastic and its distribution changes over time. Wagner (2002, p.224) observes “...most demand environments change over time, usually gradually but sometimes abruptly ...”; Graves (1999, p.50) attributes this phenomenon to increasingly shorter life cycles of products that cause demand to become non-stationary.

Frequently, in cases when demand is stochastic and non-stationary (e.g., due to growth or seasonality), little can be said about an optimal inventory policy and managers must rely on approximations, heuristics, or base-stock models with appropriate “fudge factors,” designed to balance conflicting pressures.\(^4\) Pressures for high inventories include reduction of backorders and stockouts (Zipkin, 2000, §6), institutional reasons that lead rational inventory managers to amplify growth in demand (Gilbert, 2005; Kahn, 1987), and shortage gaming, known to occur when distributors are perceived to ration supply (Lee et al., 1997). Reasons for low inventories include high holding cost and risk of incurring obsolescence cost (Song and Zipkin, 1993, 1996).

The primary contributor to obsolescence costs, one the theoretical inventory literature has largely not considered, are inventory write-downs. Under U.S. Generally Accepted Accounting Principles (GAAP), inventories are recorded at cost. However, if the market value of inventory declines below its original cost, for reasons such as damage or obsolescence, a company must write down the inventory to the new market value to recognize this loss. Best Buy’s financial statements, for example, include the following disclosure: “…merchandise inventories are recorded at the lower of average cost or market. …” Any inventory write-down must be reflected as an expense (part of cost of goods sold) on the income statement. Thus, if the value of inventory declines, a company incurs a financial loss. These costs can be significant: For example, there is the well-publicized $2.25 billion inventory write-down at Cisco Systems, which led to a decline in the company’s market value

---

\(^1\)See Inventories: Total Business (BUSINV) at http://research.stlouisfed.org/fred2/series/BUSINV

\(^2\)This value is computed for the firms in the intersection of the COMPUSTAT annual and CRSP files for all non-financial and non-utility firms over the 1980-2008 period.

\(^3\)See Inventory to Sales Ratio: Total Business (ISRATIO) at http://research.stlouisfed.org/fred2/series/ISRATIO

\(^4\)For example, under a widely-used inventory heuristic, the base-stock inventory level moves in parallel with the mean lead-time demand (Zipkin, 2000, §9.2.2). The heuristic performs well if the demand distribution changes slowly, but leads to inventory shortages (or excesses) if the distribution changes suddenly.
from $430 billion in March of 2000 to $108 billion in March of 2001; more recently, the digital video recorder (DVR) supplier TiVo posted a net loss of $17.7 million in its fiscal second quarter of 2007, which the company accredited to an inventory write-down.

Despite its prevalence and potential significant impact, there is a lack of research empirically investigating inventory write-downs in the literature. Many important questions remain open and could be addressed by empirical analysis. For example, how do inventory write-downs affect firms' operating performance? Is there any link between write-downs and demand growth? If yes, how do firms manage inventory when facing growing demand and inventory write-down risks (e.g., is their behavior consistent with the theoretical predictions)? What is the relationship between inventory write-downs and firms' inventory purchasing policies?

To shed some light on the above questions, we use publicly available data to conduct a descriptive investigation of inventory write-downs. Using results from existing analytical inventory and accounting literatures, we propose several hypotheses that relate sales growth, inventory purchases, and inventory write-downs. We then test these relationships based on a sample of 290 firms experiencing a first-time inventory write-down of more than 1% of average total assets between calendar years 2002 and 2004.

We report four main findings. First, we provide large sample descriptive statistics of the magnitude and impact of inventory write-downs. We document an average inventory write-down in our sample of $13.2 million representing 3.7% of a firm's average total assets. We find that the financial consequences of these write-downs are severe. Firms experiencing write-downs have a -15.4% return on assets and a -21.5% market-adjusted return in the period of the write-down.

Second, using sales data as a proxy for unobservable demand data, we establish a link between firms' sales growth and inventory write-downs. Empirical evidence suggests that firms with high sales growth are more likely to experience a subsequent inventory write-down. This leads us to conjecture that high sales growth is more prone to slowdown, which has been confirmed in our analysis. Therefore, it appears that firms experience write-downs not only due to exuberant inventory purchases, but also due to exogenous factors, which include non-stationary demand.

Third, to learn how firms manage inventory, we investigate how much of their sold inventory firms replace under growing demand. For comparison, firms are sorted into quintiles based on their sales growth rate: firms falling in the first quintile (i.e., the top 20%) are referred to as extreme growth firms while firms in the middle quintiles (i.e., the middle 60%) are defined as moderate growth firms. We find that firms experiencing no sales growth replace approximately 100% of inventory sold during a period. This corroborates predictions from the inventory theory: Under stationary demand, firms should use policies (e.g., base-stock or \((r, Q)\) policy, depending on the presence of fixed costs) that simply replace any sold stock. When there is moderate growth, firms purchase an additional 11.8% of sales growth, i.e., firms experiencing sales growth purchase more and carry more ending inventories than without growth. Interestingly, however, extreme sales growth firms replace less of their sold inventory than their moderate growth counterparts, i.e., extreme growth firms exhibit a more conservative ordering behavior than moderately growing.
firms. There has been no consensus so far in the theoretical literature whether fast growth should lead to more aggressive inventory policies (there are arguments for both directions). Our empirical analysis suggests a non-monotone relationship between sales growth and sales adjusted inventory purchases, i.e., it not necessarily true that faster growing firms stock up more to meet demand in the next selling season.

Fourth, we examine whether there is a positive correlation between inventory write-downs and firms’ purchasing policy. This is done by comparing the ordering behavior of firms that experience write-downs to those that do not. We find that the extreme growth firms, which experience write-downs replace more of their sold inventory than other extreme growth firms. However, we do not find evidence of excessive inventory buildup. In fact, average extreme growth firms, which experience write-downs, replace approximately the same amount of sold inventory as moderate growth firms replace. Future research may explore whether this finding implies that inventory policies used by moderately growing firms are inappropriate for extreme growth firms. This may lead to useful heuristics that would function well in a non-stationary demand environment with stochastic holding cost. To our knowledge, the theoretical literature has not considered this issue.

The rest of the paper is organized as follows. Section 2 surveys the related literature. Section 3 summarizes the data, defines the variables used and summarizes the economic impact of inventory write-downs observed in our data set. Section 4 develops our hypotheses. In Section 5, we discuss the empirical methodology we employ. We present our empirical results in Section 6. In the remainder of the paper, we conclude with managerial implications and a discussion of directions for future research.

2 Literature Review

This paper is closely related to the growing empirical literature that aims to link firms’ inventory/supply chain management practices to their financial performance. Hendricks and Singhal (2005b,a) examine the impact of supply chain glitches on firms’ operating performance. It has been shown that these glitches have a significant and long-lasting impact on various operating performance metrics. Singhal (2005) documents the negative effect of holding excessive inventories on stock price. Chen et al. (2005) examine the trend of inventories held by American companies and find that abnormally high stock levels are associated with poor long-term stock returns. Rumyantsev and Netessine (2007a) study whether a firm should use lean or responsive inventory policies. They report that mismatched changes in sales and inventory are associated with lower profitability. Kesavan and Mani (2010) test the predictive power of abnormal inventory growth at retail firms. They find that abnormal inventory growth is useful in forecasting retailers’ future earnings; however, both managers and analysts tend to overlook the information conveyed by such a metric. Our paper contributes to this literature by investigating the impact of inventory write-downs on firms’ operating performance.

This paper is also related to the studies that empirically compare inventory levels across
firms/industries/time. Rajagopalan and Malhotra (2001) study the trends of three types of inventories in the U.S. manufacturing industry during the period 1961-1994: materials, work-in-progress, and finished goods. Their analysis suggests that the results from various inventory-reduction efforts are encouraging but also somewhat mixed during this period. Several papers attempt to explain why firms differ in their inventory levels or key performance measures such as inventory turnover. Lai (2006) studies how the stock market affects firms’ inventory decisions. He identifies strong statistical evidence that firms decrease inventory when the market discounts high-inventory firms. Cachon and Olivares (2010) examine the drivers of finished goods inventory in the U.S. automobile industry. It has been found that the difference in finished goods inventory at the major manufacturers can be largely explained by two factors: the number of dealerships in the firm’s distribution network and the firm’s production flexibility. Gaur et al. (2005) and Gaur and Kesavan (2007) develop empirical models to study the correlation between inventory turnover and firm characteristics such as gross margin, size, and sales growth. They find that inventory turnover increases with sales growth, and the rate of the increase depends on firm size and the sign of the sales growth. Rumyantsev and Netessine (2007b) use U.S. company data to test whether the insights offered by classical inventory theory hold at firm levels. They show that firms facing uncertain demand, long lead times, and high gross margin tend to hold more inventory, and larger companies hold less inventories than smaller ones because of the economies of scale effect. Cachon and Olivares (2009) study the effect of competition on retail stock levels by using the data collected from General Motors’ dealerships. The results suggest that dealers carry more inventory when they face intense retail competition (i.e., competition drives the retailers to offer higher service levels). In this paper, we establish the link between high sales growth and the likelihood of inventory write-downs, which is new in the literature. Also we study the association between firms’ ordering behavior and inventory write-downs by examining inventory changes relative to sales growth over time.

There is an enormous theoretical literature on inventory management. The emphasis of this literature is on analysis of various model settings and characterization of optimal (or near-optimal) ordering policies. Readers are referred to Zipkin (2000) and Porteus (2002) for comprehensive reviews of these inventory models. The extant theoretical inventory literature does not explicitly consider inventory write-downs. It does, however, consider closely related issue of inventory obsolescence and conjectures that its financial impact may be significant (Song and Zipkin, 1993, 1996).

Finally, researchers in accounting have spent considerable time examining issues associated with inventory accounting. Primarily this literature has focused on the determinants and consequences of firms’ accounting choice between LIFO and FIFO costing methods (for a review of this literature, see Derstine and Huefner, 1974; Dopuch and Pincus, 1988; Niehaus, 1989; Fields et al., 2001). Despite considerable research examining inventory, very little research has examined the link between demand, purchasing policies and inventory write-downs. Despite this dearth of research in accounting a few studies are of note. Francis et al. (1996) provides evidence on the causes and shareholder wealth effects of asset write-downs. Although the focus of the research is on
discretionary asset write-offs, which exclude inventory write-downs, they provide strong descriptive
evidence that inventory write-downs are large in magnitude and are associated with significant de-
creases in shareholder wealth. Chen et al. (2010) investigate inventory write-downs using empirical
data in the semiconductor industry, but their focus is on managers’ incentives when deciding the
timing and magnitude of inventory write-downs. In a more recent study Allen et al. (2010) examine
a hand collected sample of inventory write-downs in the context of the accrual anomaly (Sloan,
1996; Thomas and Zhang, 2002). They find that firms with extreme reversals in inventory from
high to low (low to high) experience significantly negative (positive) earnings and size-adjusted
returns. Using a sample of inventory write-downs they document that reversals in inventory from
high to low are extremely overrepresented by firms experiencing inventory write-downs. They also
document that inventory write-downs negatively signed are negatively associated with changes in
prior periods inventory. They argue that inventory write-downs and the accrual anomaly result
from inventors’ lack of understanding that extreme accruals are associated with reversals in accruals
resulting from accrual estimation error. They point out that these errors can result from several
factors including vagaries of Generally Accepted Accounting Principles (Penman and Zhang, 2002),
demand shifts (Thomas and Zhang, 2002), earnings management (Xie, 2001) and over-investment
by high growth firms (Titman et al., 2009). Although there are many potential explanations for
write-downs, we complement and extend Allen et al. (2010) by more closely examining two possible
explanations: demand shifts and inventory investment decisions.

3 Data Description and Definition of Variables

We examine two primary data samples in our analysis. Our first large sample consists of data
from the COMPUSTAT fundamentals annual file spanning the years 1980-2008. Stock return data
for both samples are obtained from the CRSP monthly returns files. As in prior literature examining
changes in inventory, we eliminate all financial services companies (SIC 6000-6999) and utilities (SIC
4900-4999). We also limit the sample to domestic firms (pops = D and fic = USA) traded on the
NYSE, NASDAQ or AMEX. Our inventory write-down sample uses write-down data that is hand
collected from 10-K filings in conjunction with data from the COMPUSTAT fundamentals annual
file. The size of our write-down sample is restricted by the cost of hand collecting data and is thus
limited to the calendar years 2002-2004. More detailed description is given below.

The financial variables of interest in this study are changes in sales ($\Delta SALES$) and purchases
scaled by cost of goods sold ($PURCHASES/COGS$). Purchases is calculated as cost of goods
sold plus ending inventory minus beginning inventory. We measure change in sales as percentage
year-over-year changes. We require the availability of COMPUSTAT data for each of the variables
for periods $t$ and change in sales for periods $(t - 1)$ and $(t + 1)$. Because our interest is in examining
firms’ ordering policy with respect to inventory, we also eliminate firms with insignificant inventory
levels, less than 1% of average total assets. Our final large sample consists of 87,295 firm-year
observations from 1980-2008.
We hand collect our inventory write-down data from firms’ 10-K filings. We searched all 10-K filings on the DirectEdgar database during the calendar years 2001-2004. We began by conducting a keyword search for any occurrence of the word write within ten words of the word inventory. We then filtered these results using the CIK numbers of all firms that appeared in both the CRSP and COMPUSTAT databases for the years 2001-2004 and were traded on a major US exchange. After filtering we were left with 5,638 filings. We then searched and read through each 10-K for discussion or documentation of an inventory write-down during the fiscal year. Upon finding evidence of an inventory write-down, we collected the inventory write-down amount from the annual report. Our search yielded many inventory write-downs that were insignificant in size and serial in frequency. As a result we chose to limit our inventory write-down sample to write-downs equal to or larger than 1% of average total assets. In addition, because we empirically observe that many firms take subsequent inventory write-downs following a first large write-down, we limit our sample to a first significant inventory write-down. Because we collect data beginning in 2001 and are thus unaware of whether a firm experiences an inventory write-down in 2000, we limit our inventory write-down sample to 2002-2004. Our sample of first time significant inventory write-down firms with available data consists of 290 firm observations. For some of our analysis we merge this write-down sample with the subset of observation in our larger 1980-2008 sample in the calendar years 2002-2004 to obtain a sample of firms experiencing first-time inventory write-downs and firms not experiencing first-time inventory write-downs. As in previous research, we find that the distributions of our financial variables are characterized by a small number of extreme outliers. We therefore follow the standard procedure of winsorizing observations at the 1st and 99th percentiles of the distribution.

3.1 Economic Impact of Inventory Write-downs

Table 1 presents descriptive statistics for our 290 firms experiencing inventory write-downs. The average inventory write-down in our sample is $13.2 million. This represents 3.7% of a firm’s average total assets suggesting that the magnitude of these write-downs is quite significant. The mean firm in our sample experiences a -15.4% return on assets and a -21.5% equally-weighted market adjusted return in the year of an inventory write-down, suggesting that write-downs are associated with extreme negative consequences for shareholders.

To further confirm this result, we conduct a matched sample analysis. We begin with the entire sample of inventory write-down firm years in our sample and match each write-down firm year to one non-inventory write-down firm year. Firms are first matched on calendar year and then 2-digit SIC codes. If a proper two digit match is unavailable then we move to a one digit match. We then require that matched firms be within 20% of total assets and inventory as a percentage of total assets in period \((t - 1)\). From the remaining firms that meet these criteria, the firm with the closest change in sales to the write-down firm in period \((t - 1)\) is selected as a match. Similar matching sample analysis is common in the accounting and finance literatures (e.g., Erickson et al., 2006; Lang et al., 2006). Using the matched firm, we then compare the performance of the firms in period \(t\), the period our write-down sample experiences a significant first time write-down. Table
2 presents the results of our analysis. The matches appear to have very similar characteristics in period \((t - 1)\). The only exception is that inventory levels for write-down firms are statistically significantly higher than for the matched sample. Despite the statistical significance the magnitude of the difference is quite small at 0.6 percent of average total assets. The results of primary interest presented in Table 2 are the extreme differences between the sample groups in period \(t\) return on assets and stock returns. Inventory write-down firms return on assets is -0.071 lower than matched firms and inventory write-down firms experience stock returns -0.41 lower than matched firms. The results confirm the fact that write-downs are associated with extreme negative consequences for firms.

Table 3 provides descriptive evidence on the industries most likely to experience inventory write-downs. The first column lists the industry, the second column is the percentage breakdown of our inventory write-down sample and the third column is the percentage breakdown of the entire sample by industry during 2002-2004. The bolded industries are those where the proportion of inventory write-down firms is significantly over represented relative to the proportion of firms in the non write-down sample. Consistent with the notion that inventory write-downs are more likely for firms with extreme sales growth and short product life cycles, we find that write-downs are overrepresented in both heavy industry and computers, semiconductor and computer services.

We would like to highlight the time-series trend in key variables for our inventory write-down sample. Figure 1(a) presents the time-series mean and median change in inventory as a percentage of average total assets. The figure shows that in the periods preceding the inventory write-downs inventory is significantly increasing. This suggests that firms are purchasing significant amounts of inventory prior to the write-down period. Figure 1(b) documents the evolution of firms sales growth prior to and following an inventory write-down. Consistent with the notion that extreme sales growth may be one explanation for inventory write-downs, we find that sales are consistently increasing prior to the write-down period. The mean (median) sales in the periods \((t - 4)\) to \((t - 2)\) is consistently increasing and growing at a rate of over 25% (10%) each year. In period \((t - 1)\) sales growth dramatically slows and then approaches zero in the period of the inventory write-down.

The next two figures document the performance of inventory write-down firms and illustrates the consequences to firms that are forced to write down inventory. Figure 1(c) shows that return on assets (ROA) declines dramatically in the period of the write-down and then bounces back in period \((t + 1)\). Figure 1(d) shows that mean market-adjusted returns are positive in the periods \((t - 4)\) to \((t - 2)\) as sales growth increases, zero in the period prior to the write-down and extremely negative in the year of the write-down. The subsequent bounce back in ROA in the year immediately following the write-down is consistent with the “big bath” phenomenon where firms excessively write-down inventory (Riedl, 2004). This results in subsequent high margins and increased ROA when the inventory is sold. The bounce back in returns is consistent with the accruals anomaly documented first by Sloan (1996).
4 Hypothesis Development

In this section, we set up the hypotheses to relate sales dynamics\textsuperscript{5}, non-recurring inventory holding costs (write-downs), and inventory ordering policy. The first two hypotheses, presented in Section 4.1, link stochastic inventory write-downs to firms’ sales growth. The subsequent three hypotheses, presented in Section 4.2, connect stochastic inventory write-downs to firms’ ordering behavior. Combining the hypotheses in Section 4.1 and Section 4.2 sheds some light on how firms manage the trade-off between stochastic holding and shortage costs when these are induced by both \textit{exogenous} (sales growth) and \textit{endogenous} (inventory policy) factors. We motivate the hypotheses primarily based on results in the theoretical inventory literature.

4.1 Sales Growth and Inventory Write-downs

First, we test the following hypothesis:

\textbf{Hypothesis 1.} The probability of a future inventory write-down is increasing in sales growth.

To our knowledge, there is little or no empirical research that connects sales growth and non-recurring inventory costs. The validity of the above hypothesis may not be obvious because intuition suggests that high sales growth would purge a company of any potential excess inventory.

Under GAAP, companies should recognize an inventory write-down when the value of inventory declines below its original cost. Taking this principle into consideration, we would expect firms that carry higher levels of more rapidly depreciating inventory to be more prone to inventory write-downs than their opposites.

The theoretical result that links high sales growth to high output is found in Veinott (1965b, Theorem 5) who considers a very general dynamic inventory model with non-stationary demand. Veinott (1965b) finds that firms ought to respond to a stochastically increasing demand by upwards adjusting their base-stock inventory level.

The link between increased output and product life cycle is established in Klepper (1996) who asserts that the larger the output of a firm the greater its total spending on R&D (Klepper, 1996, Proposition 7). Since increased R&D spending is widely interpreted to imply greater product innovation and shorter product life cycle, we hypothesize that it also implies a faster pace at which the existing products held in inventory become obsolete.

That is, based on results in Veinott (1965b) and Klepper (1996), we conjecture that as sales growth increases, companies carry larger quantities of inventory that lose value more rapidly, increasing the probability of an inventory write-down. This reasoning also agrees with the anecdotal evidence from the high-tech industry (Teach, 2001).

Should an inventory write-down occur, one of the reasons could be that an inventory build up is followed by a slowdown in sales growth. Since fast growing firms are more prone to inventory write-downs, we would like to test whether extreme sales growth is more prone to a slowdown than

\textsuperscript{5}To test our hypotheses, we use firms’ sales data as a proxy for otherwise unobservable demand data.
moderate sales growth. Extreme growth is defined as the top quintile while moderate growth is
associated with the middle quintiles of sales growth rate.

Hypothesis 2. Extreme sales growth is less persistent than moderate sales growth.

There is a large empirical literature in accounting and finance that seeks to identify predictable
variation in earnings and profitability. One of the findings this literature reports is that high
rates of growth in returns on assets (ROA) and returns on equity (ROE) are not sustainable and
that, eventually, these rates decay toward some economy-wide level. Stigler (1963) explains this
phenomenon using a standard economic argument by noting that competition causes above average
growth in returns to erode: “...entrepreneurs seek to leave relatively unprofitable industries and
enter relatively profitable ones ...”

In particular, Fama and French (2000) perform a time series study using NYSE, AMEX, and
NASDAQ firms on COMPSTAT with total assets greater than $10 million and document that
changes in ROA mean revert; i.e., they are beset by short-term disturbances but tend toward a
long-run mean value. Moreover, the rate of mean reversion is faster when the growth in ROA is
further from its mean in either direction. Therefore, the higher the growth rate the less persistent
it appears to be.

To develop Hypothesis 2, we now identify a formal link between changes in ROA and changes
in sales. Using the standard DuPont analysis, a company’s ROA can be written as a product of
profit margin and asset turnover:

\[
ROA = \text{Profit margin} \times \text{Asset turnover} = \frac{\text{Net income}}{\text{Total sales}} \times \frac{\text{Total sales}}{\text{Avg. total assets}},
\]

from which

\[
ROA = \frac{\text{Net income}}{\text{Avg. total assets}} = \frac{\text{Total sales} \times \text{Profit margin}}{\text{Avg. total assets}}.
\]

The right side of the above equation reveals that constant profit margin and constant average total
assets mean that changes in ROA are driven by changes in total sales. That is:

\[
ROA_t - ROA_{t-1} = (\text{Total sales}_t - \text{Total sales}_{t-1}) \times \frac{\text{Profit margin}}{\text{Avg. total assets}}, \quad \text{or}
\]

\[
\text{Total sales}_t - \text{Total sales}_{t-1} = (ROA_t - ROA_{t-1}) \times \frac{\text{Avg. total assets}}{\text{Profit margin}}. \quad (1)
\]

Since the results in Fama and French (2000) imply that high growth in ROA is less persistent than
average growth in ROA, then from Equation (1) we hypothesize that the same applies to growth in
sales. Nissim and Penman (2001) offer some visual evidence of this but provide no formal test. We
formally test the hypothesis that extreme sales growth is less persistent than average sales growth.
4.2 Purchasing Behavior and Inventory Write-downs

We proceed to investigate firms’ purchasing behavior under growing demand and its relationship with non-recurring inventory costs (write-downs).

What would a manager do when facing rapidly growing demand? This question is important for understanding firms’ purchasing behavior because in such an environment often little can be said about an optimal policy and managers must rely on approximations, heuristics, or base-stock models with appropriate “fudge factors,” designed to balance conflicting pressures (for an excellent overview, see Zipkin, 2000, §9).

Inventory theory with non-stationary demand dates back to the early 1960’s. One of the notable results is that when demand stochastically increases over time, the simple base-stock policy is optimal under certain conditions (Karlin, 1960; Veinott, 1965a). In addition, the optimal base-stock levels do not decrease over time. This suggests that it is optimal for firms to purchase more inventory when facing increasing demand. Rapidly increasing demand may also impose a psychological effect on managers’ ordering behavior. It has been widely documented in the literature that over-reaction biases are commonplace in people’s decision making (see, for example, Armstrong and Brodie (1987) in marketing, Watson and Zheng (2008) in operations management, and Nosic and Weber (2009) in behavioral finance). Anecdotal evidence of firms over-reacting to demand changes abounds in the business press. Just to mention a couple of examples: Cisco Systems was overly optimistic about market demand before the gigantic write-down in 2001 (Burrows, 2003), and John Deere displeased its customers by carrying insufficient inventory because of being pessimistic about the effect of the recent recession (Singh, 2010). The over-reaction effect may also be aggravated by certain practices adopted in many real-world supply chains. For instance, in the networking equipment industry customers are allowed to duplicate orders at multiple suppliers; in the semiconductor industry, customers may place soft orders that might be cancelled at a later stage. Armony and Plambeck (2005) demonstrate that allowing customers to duplicate (and freely cancel) orders will make a firm overestimate the demand rate and hence over-invest in capacity. Based on this discussion, it is natural to conjecture that adjusted for sales, extreme growth firms would carry more inventories than moderately growing firms.

To understand how firms respond to sales growth, we test how much of their sold inventory firms replace. That is, we measure firms’ inventory purchase as a percentage of the sold inventory. (For ease of exposition, we say that a firm with high (low) inventory purchase as a percentage of the sold inventory uses aggressive (conservative) inventory policy.) The main advantage of this approach is that it is impervious to an inventory policy: For example, a particular firm facing i.i.d. demand may use a base-stock or \((Q,r)\) policy; although the policies are different, standard inventory theory predicts that under either policy the firm orders new stock so as to replace any sold stock. First, we test the following hypothesis, which examines the connection between sales growth rate and firms’ purchasing behavior.

**Hypothesis 3.** As a percentage of inventory sold, extreme growth firms replace more inventory than moderate growth firms.
Next, we examine the connection between inventory write-down and firms’ purchasing policy. Note that firms need to write down inventory when the value is lower than the cost. Thus ordering/carrying more inventory will clearly increase the likelihood of a write-down. Since we conjecture that there is a link between growth rate and inventory write-downs, a natural question arises: do extreme growth write-down firms exhibit different purchasing behavior from the other firms? We investigate this question by testing the following two hypotheses:

**Hypothesis 4.** As a percentage of inventory sold, extreme growth firms that experience inventory write-downs replace more inventory prior to a write-down than other extreme growth firms.

**Hypothesis 5.** As a percentage of inventory sold, extreme growth firms that experience inventory write-downs replace more inventory prior to a write-down than moderate growth firms.

Hypothesis 4 compares the ordering behavior of extreme growth firms with and without experiencing inventory write-downs. We expect that these firms exhibit different ordering behavior, which, if true, may help explain why some extreme growth firms were hit by significant write-downs while the rest successfully avoided them. This would offer useful managerial insight that helps firms reduce the probability of inventory write-downs. Hypothesis 5 goes one step further to compare the ordering behavior of extreme growth firms that experienced write-downs to that of moderately growing firms. We test this hypothesis because the average firms have moderate growth rates and their behavior may serve as a natural benchmark to evaluate that of the extreme growth write-down firms. As it becomes clear in Section 6.2 on page 17, this hypothesis, combined with Hypotheses 3 and 4, provide interesting managerial implications about the firms’ ordering behavior under extreme growth.

### 5 Methodology

We begin by testing our first hypothesis that the probability of a future inventory write-down is increasing in sales growth. To do this we model the probability that a firm will experience a first inventory write-down (FWD), equal to one for firm-years experiencing a first large inventory write-down and zero otherwise, as a function of previous periods’ sales growth and change in inventory:

\[
PR(WD = 1)_{t+1} = b_0 + b_1 \Delta SALES_t + b_2 \Delta SALES_{t-1} + b_3 \Delta INV_t + b_4 \Delta INV_{t-1} + \epsilon_{t+1}. \tag{2}
\]

We estimate Model (2) using a logistic regression with standard error estimates clustered on firms (Rogers, 1993). If the probability of an inventory write-down is increasing in past sales growth, we expect to find significant positive coefficients for \(b_1\) or \(b_2\). Consistent with previous research by Allen et al. (2010) we also expect that the probability of an inventory write-down is increasing in positive inventory changes. Therefore, we also expect to find positive and significant coefficients for \(b_3\) and \(b_4\).

Next, we present tests for our second hypothesis that extreme sales growth is less persistent than average sales growth. We model sales growth in period \((t + 1)\) as a function of period \(t\) sales growth:
growth and interaction terms between sales growth and indicator variables equal to one if a firm experiences extreme sales growth in period $t$ and zero otherwise. We define extreme sales growth by sorting firms into quintiles each year based on sales growth. If a firm falls in the top quintile in period $t$, $\text{ExtremePositive}_t$ equals one and zero otherwise. If a firm falls in the bottom quintile, $\text{ExtremeNegative}_t$ equals one and zero otherwise. We estimate the following model using a pooled sample and ordinary least squares (OLS).

\[ \Delta \text{SALES}_{t+1} = b_0 + b_1 \Delta \text{SALES}_t + b_2 \Delta \text{SALES}_t \times \text{ExtremePositive}_t \]
\[ + b_3 \Delta \text{SALES}_t \times \text{ExtremeNegative}_t + \epsilon_{t+1} \quad (3) \]

In all of our OLS regressions, we use two-way clustered standard errors, for firms and years, to adjust for both cross-sectional and serial correlation (Petersen, 2009). If Hypothesis 2 is descriptive, we expect to find significant negative coefficients on $b_2$ and $b_3$ in Model (3). In addition to testing whether firms falling in the most extreme quintiles of sales growth experience less persistent sales growth relative to other sample firms, we examine how extreme firms compare to firms experiencing more moderate sales growth. We do this by creating two more interaction terms. The first term, $\Delta \text{SALES}_t \times \text{High}_t$, is sales growth for firms falling in the second quintile of sales growth and $\Delta \text{SALES}_t \times \text{Low}$ is sales growth for firms falling in the fourth quintile of sales growth.

\[ \Delta \text{SALES}_{t+1} = b_0 + b_1 \Delta \text{SALES}_t + b_2 \Delta \text{SALES}_t \times \text{ExtremePositive}_t \]
\[ + b_3 \Delta \text{SALES}_t \times \text{ExtremeNegative}_t + b_4 \Delta \text{SALES}_t \times \text{High}_t \]
\[ + b_5 \Delta \text{SALES}_t \times \text{Low}_t + \epsilon_{t+1} \quad (4) \]

If sales growth for moderate growth firms is less persistent than sales growth for firms falling in the middle quintile of sales growth, we would expect to find significant negative coefficients for $b_4$ and $b_5$ in Model (4). We would also expect that if extreme sales growth both positive and negative ($\text{ExtremePositive}$ and $\text{ExtremeNegative}$) is less persistent than more moderate sales growth ($\text{High}$ and $\text{Low}$) to find that $b_2 < b_4$ and $b_3 < b_5$.

We now lay out tests of our Hypothesis 3 that, extreme growth firms replace more of their sold inventory than their moderately growing counterparts. We expect that firms make purchasing decisions at the beginning of and during a period based on expectations of sales in the upcoming period. Therefore, we use the approach of modeling levels of purchases scaled by cost of goods sold ($\text{Purchases}_t/\text{COGS}_t$) as a function of contemporaneous changes in sales. Scaling purchases by cost of goods sold has a straightforward and natural interpretation. Firms whose ratio is equal to one (assuming constant costs) are purchasing just enough inventory to replace inventory sold. Whereas, firms with a ratio higher than one are increasing inventory and firms with a ratio lower than one are depleting inventory levels. If managers believe that demand shifts are not completely transitory then we expect to find that as $\Delta \text{SALES}_t$ increases (decreases) $\text{Purchases}_t/\text{COGS}_t$ increases (decreases). We examine changes in the ratio of purchases to cost of goods sold in period.
t as a function of contemporaneous sales growth and sales growth interacted with indicator variables for extreme positive and negative growth.

\[ \frac{\text{Purchases}_t}{\text{COGS}_t} = b_0 + b_1 \Delta \text{SALES}_t + b_2 \Delta \text{SALES}_t \times \text{ExtremePositive}_k + b_3 \Delta \text{SALES}_t \times \text{ExtremeNegative}_k + \epsilon_t \] (5)

The \text{ExtremePositive} and \text{ExtremeNegative} indicator variables in Model (5) allow us to test how the purchasing and inventory holding policies of extreme growth firms differ from moderate growth firms. We condition extreme sales growth on two different time periods by setting \( k \) equal to \((t - 1)\) and \( t \). The appropriate conditioning period for extreme growth, \((t - 1)\) or \( t \), will depend on the available information managers can use for making purchasing and investing decisions. If managers face long lead times or high fixed costs, it is likely that managers will be forced to make inventory decisions earlier, which would suggest that purchases and inventory will be a function of the extremity of sales growth in \((t - 1)\). On the other hand, if firms face very short lead times and low fixed costs, then managers will likely be able to adjust their purchase decisions during the period based on the current extremity of sales growth. Because lead times and fixed costs associated with inventory are largely unobservable, we present results conditioning on both periods. We expect that if extreme growth firms’ inventory purchasing and holding decisions are similar to moderate growth firms we will find that \( b_2 = 0 \) and \( b_3 = 0 \). On the other hand, if we find that \( b_2 < 0 \) (\( b_2 > 0 \)) this would suggest that, compared to moderately growing firms, extreme positive growth firms tend to shrink (expand) their inventories.

Using only the time period subsample for which we have inventory write-down data (2002-2004), we next test both Hypotheses 4 and 5. Using the identical framework as our tests of Hypothesis 3, we are able to examine the purchasing and inventory holding behavior of firms that ex-post experience inventory write-downs by interacting our extreme growth interaction terms with an additional indicator variable for firms that ex-post experience inventory write-downs.

\[ \frac{\text{Purchases}_t}{\text{COGS}_t} = b_0 + b_1 \Delta \text{SALES}_t + b_2 \Delta \text{SALES}_t \times \text{ExtremePositive}_k + b_3 \Delta \text{SALES}_t \times \text{ExtremeNegative}_k + b_4 \Delta \text{SALES}_t \times \text{ExtremePositive}_k \times \text{FWD}_{t+1} + b_5 \Delta \text{SALES}_t \times \text{ExtremeNegative}_k \times \text{FWD}_{t+1} + \epsilon_t \] (6)

Similar to our previous tests, we condition on extreme sales growth in both periods \( t \) and \((t - 1)\). If extreme growth firms that experience subsequent write-downs purchase and hold more inventory than other extreme growth firms, then we expect to find that \( b_4 \) in Model (6) is positive and significant. We can test our final hypothesis that extreme growth firms that experience inventory write-downs purchase and hold more inventory prior to a write-down than moderate growth firms using the same model. By examining the sum of the coefficients \( b_4 \) and \( b_2 \), we can compare extreme growth firms that subsequently experience inventory write-downs to moderate growth firms. If extreme growth write-down firms purchase significantly more inventory than average sales growth...
firms, we expect to find that $b_4 + b_2 > 0$. On the other hand, if we find that $b_4 + b_2 = 0$ this would suggest that extreme growth write-down firms behave similarly to moderate growth firms.

6 Results

6.1 Sales Growth and Inventory Write-down Results

Before providing formal tests for our first hypothesis, we provide additional descriptive evidence on the relation between sales growth and inventory write-downs. Referring back to Figure 1(b), it appears that there is an increasing trend in sales growth especially in the periods $(t-3)$ and $(t-2)$ relative to the inventory write-down period, $t$. In Figure 2, we provide evidence on the distributional properties of the data points in Figure 1(b). Sales growth is ranked each year into quintiles using the entire sample of inventory write-down and non inventory write-down firms. Figure 2 presents the percentage of the inventory write-down firms falling into each quintile of sales growth in the period of the inventory write-down, $t$, and the three periods prior to inventory write-downs, $(t-1)$, $(t-2)$, and $(t-3)$. The mean size of the inventory write-down scaled by average total assets for firms falling in each quintile is presented at the top of each bar. If there is no relationship between sales growth and inventory write-downs, we would expect to find that approximately 20% of write-down firms fall in each quintile. Not surprisingly, in the period of the inventory write-down the lowest sales growth quintile is extremely overrepresented by write-down firms (41.3%). In the period immediately prior to a write-down it appears that the firms are distributed fairly evenly across quintiles. While in periods $(t-2)$ and $(t-3)$, the strongest representation of write-down firms appears in the fastest sales growth quintile (31.1% and 27.3% respectively). These also appear to be the most extreme write-downs as the mean write-down is -4.2% and -4.1% of average total assets in periods $(t-2)$ and $(t-3)$ respectively. This descriptive evidence is consistent with a positive relation between sales growth and inventory write-downs especially in periods $(t-2)$ and $(t-3)$. We now provide formal statistical tests for our first hypothesis.

Table 4 presents regression results for our first hypothesis. Equation (A) presents results for sales growth alone. Equation (B) presents results for just changes in inventory, and Equation (C) presents results for both sales growth and changes in inventory. The results are consistent with our hypothesis. It appears that sales growth in the period immediately prior to an inventory write-down has no significant explanatory power, consistent with the evidence provided in Figure 2(b). But, sales growth two periods prior to an inventory write-down does have significant explanatory power in the predicted direction. In Figure 1, sales growth is extremely positive for inventory write-down firms in periods $(t-4)$ to $(t-2)$. In period $(t-1)$ sales growth slows significantly. A comparison of the inventory write-down firms’ sales growth in period $(t-1)$ and average sales growth in the non-inventory write-down sample suggests there is no difference. This result suggests two possible scenarios: (1) Managers are slow to recognize inventory write-downs in accounting earnings and thus delay the write-downs until they are forced to write-down inventory by auditors, or (2) managers do not believe that sales growth will continue to decline and expect that as a result
inventory will maintain its value. Under either circumstance, it appears that positive sales growth, in particular in period \((t-2)\), is positively associated with inventory write-downs. In Equation (B), we document that inventory write-downs are also increasing in positive changes in inventory. This should not be surprising. If firms were able to decrease inventory prior to a write-down by selling the inventory for its book value, firms would not be forced to incur a write-down. The buildup in inventory along with decreasing sales growth is a signal that firms have excess inventory on their balance sheets that they are unable to sell.

Now that we have established Hypothesis 1, we next examine the persistence of sales growth. We begin by providing some descriptive evidence in the form of a figure, and then we statistically test whether extreme growth is less persistent than moderate growth. We first sort firms into sales growth quintiles based on sales growth in period \(t\) and then plot mean sales growth for period \(t\) and the subsequent four years. Figure 3, similar to Nissim and Penman (2001), clearly shows that mean sales growth reverts at a quick rate. The extreme sales growth both positive and negative appears to dissipate within one to two periods. The mean reversion in the negative sales growth term might be the result of a survivorship bias in the data as firms that experience negative sales growth must either turn the trend around or go out of business. For this reason, we spend little time examining the results of extreme negative sales growth firms but focus our attention on firms experiencing extreme positive sales growth. Although Figure 3 provides strong visual evidence that sales growth for extreme firms is unlikely to be persistent, we provide statistical evidence for this in Table 5.

Table 5, Equation (A) shows that on average sales growth exhibits some persistence although it is significantly less than one. Equation (B) includes interaction terms for extreme negative and positive growth. The negative and significant coefficients for both \(\Delta SALES_t \times ExtremePositive_t\) and \(\Delta SALES_t \times ExtremeNegative_t\) suggest that extreme sales growth is much less persistent than sales growth for non-extreme firms. In fact, extreme negative sales growth completely reverses \((0.361 - 0.412)\) and sales growth for extreme positive firms is 39% less persistent than for moderate growth firms. We further explore the persistence of sales growth in Equation (C) by introducing interaction terms for moderate sales growth, those falling in quintiles 2 and 4. The persistence of firms falling in the middle quintile of sales growth is represented by the coefficient on \(\Delta SALES_t\). The evidence in Equation (C) for extreme growth firms is consistent with Equation (B) and confirms that indeed extreme sales growth is less persistent than moderate sales growth. The insignificant coefficients on \(\Delta SALES_t \times HIGH_t\) and \(\Delta SALES_t \times LOW_t\) suggest that moderate sales growth is just as persistent as mean sales growth. To statistically confirm that \(\Delta SALES_t \times ExtremePositive_t\) and \(\Delta SALES_t \times ExtremeNegative_t\) differ from \(\Delta SALES_t \times HIGH_t\) and \(\Delta SALES_t \times LOW_t\) respectively, we conduct F-tests comparing the coefficients. Statistical equivalence is soundly rejected at the less than 1% level. All of our results confirm our second hypothesis that extreme sales growth is less persistent than moderate sales growth.
6.2 Purchasing Behavior and Inventory Write-down Results

Tables 6 and 7 provide results for tests of Hypothesis 3. Table 6 regresses changes in purchases scaled by cost of goods sold in period $t$ on contemporaneous changes in sales and interaction terms for firms experiencing extreme sales growth in the prior period, $(t-1)$. Table 7 is identical to Table 6 except that extreme growth is conditional on time $t$ rather than $(t-1)$. We condition on both time periods to show that our results are robust to the timing of inventory decisions. If one assumes that purchase decisions are made primarily at the beginning of the period then the more appropriate tests are likely those presented in Table 6. On the other hand, if one assumes that firms have the ability to significantly adjust purchase decisions during the period, perhaps the results presented in Table 7 are more descriptive. We present both tables to demonstrate the results are robust to either assumption. The primary coefficients of interest in our specifications are $\Delta Sales_t \times \text{ExtremePositive}_{t-1}$ in Table 6 and $\Delta Sales_t \times \text{ExtremePositive}_t$ in Table 7. As outlined in the methodology section, we expect that if positive extreme growth firms behave consistent with Hypothesis 3, we will find that these coefficients are positive and significant suggesting that extreme sales growth leads to more aggressive replacement of inventory relative to sales leading to higher ending inventory.

Contrary to our conjectures in Hypothesis 3, we find that consistently across all specifications the signs are negative and significant. This suggests that extreme growth firms replace less of their sold inventory than firms that do not experience extreme growth. To illustrate this, Equation (B) of Table 6 shows that firms experiencing no sales growth replace approximately 100% of their inventory sold (Const. = 1.007). This is reassuring because it corroborates a prediction from the standard inventory theory: That is, in a stationary demand environment, firms optimally use a policy (e.g., $(r, Q)$ or base-stock policy, depending on the presence of fixed costs) that merely replaces any sold stock.

Firms experiencing moderate sales growth either positive or negative replace approximately 100% of inventory sold during a period plus 11.8% of sales growth. For example, if a firm were to experience 10% positive (negative) sales growth in period $t$ the equation suggests that on average this firm would have a ratio of purchases to cost of goods sold of 1.0188 (0.9952). Thus, firms experiencing positive sales growth increase ending inventory levels; while, firms experiencing negative sales growth decrease ending inventory levels. This also corroborates predictions from the theoretical literature: When firms experience moderate sales growth, time-dependent base-stock or $(r, Q)$ policies, under which base-stock level and re-order point respectively move in parallel with lead-time demand, are known to perform well (Zipkin, 2000, §9.2.2) and Equation (B) in Table 6 appears consistent with the use of such policies.

However, the rejection of Hypothesis 3 indicates that firms do not necessarily adopt more aggressive inventory policies when facing faster sales growth. When firms experience large and sudden changes in demand, often, base-stock and $(r, Q)$ structures continue to be optimal; but due to the curse of dimensionality, computing the optimal policy parameters is not easy at all (for an extended discussion, see Zipkin, 2000, §9.7).
Essentially, firms have to resolve the conflict between two pressures when making inventory decisions: the pressure for high inventories includes reduction of backorders and stockouts, institutional reasons that lead rational inventory managers to amplify changes in demand (Gilbert, 2005; Kahn, 1987), and shortage gaming (Lee et al., 1997); pressure for low inventories includes high holding cost and risk of incurring obsolescence cost (Song and Zipkin, 1993, 1996).

Controlling for sales growth, the interaction terms in Equation (B) provide evidence that extreme growth firms appear to replace stock at a lower rate than moderate growth firms. The finding suggests that on average extreme growth firms choose to resolve the conflicting pressures in favor of lower inventories, possibly due to the heightened risk of inventory write-downs, which has been generally disregarded in the theoretical literature.

The last observation also raises the question whether firms with extreme growth that subsequently experience inventory write-downs behave similarly to other extreme growth firms. We examine this question and test our concluding two hypotheses by estimating similar regressions as those presented in Tables 6 and 7, but this time we include interaction effects for firms that ex-post experience inventory write-downs. This allows us to determine whether extreme growth firms that ex-post experience inventory write-downs behave similarly to non-inventory write-down extreme growth firms. Due to constraints associated with hand collecting inventory write-down data, our sample is limited to the 2002-2004 time period. Tables 8 and 9 present our results. Table 8 corresponds in nature to Table 6 because the extreme sales growth interaction variables are conditional on previous period’s sales growth. While, Table 9 corresponds to Table 7 because extreme sales growth is conditional on current period’s sales growth. To show that our previous results hold in this more limited sample period, we first estimate the models without inventory write-down interaction terms. In Equations (B) for both tables, we find similar results. Extreme positive growth firms appear to replenish inventory relative to current period sales growth at a lower rate than firms that experience moderate sales growth.

Now we turn our focus to the more extensive specification, Equation (C) of Table 8. We are primarily interested in the coefficient on $\Delta SALES_t \times ExtremePositive_{t-1} \times FW_{D_{t+1}}$. If extreme growth firms that ex-post experience inventory write-downs behave as other extreme growth firms, we expect the coefficients to be statistically insignificant. On the other hand, if we find that the coefficients are positive, this suggests that firms that experience extreme growth and inventory write-downs behave more aggressively than other extreme growth firms in that they replace inventory sold at higher rates. Consistent with Hypothesis 4, conditional on extreme positive sales growth in $(t - 1)$, we find that inventory write-down firms purchase more inventory and thus increase ending inventory levels more than non-inventory write-down firms in the period immediately prior to an inventory write-down.

For comparison with the results in previous tests, Table 9 examines the results conditional on extreme sales growth in period $t$. Conditioning on extreme growth in period $t$, is likely to provide weaker results than conditioning on period $(t - 1)$. Referring back to Figure 1(b), the significant decrease in sales growth in the period immediately prior to the write-down suggests
that very few inventory write-down firms experience extreme growth in the period immediately prior to an inventory write-down. This was also confirmed by our logistic results in Table 4. The results in Table 9 are signed similarly as those found in Table 8 but the coefficient on $\Delta \text{SALES}_t \times \text{ExtremePositive}_t \times \text{FWD}_{t+1}$ is smaller and insignificant. Despite this lack of significance the results in Table 8 along with solid ex-ante reasons for weak results in Table 9 provide support for the hypothesis that extreme growth firms that ex-post experience inventory write-downs replace inventory more aggressively than other extreme growth firms in the period immediately prior to a write-down.

To test our last hypothesis, that extreme growth firms that experience inventory write-downs purchase and hold more inventory prior to a write-down than moderate growth firms, we conduct an F-test to determine whether the joint effect of $\Delta \text{SALES}_t \times \text{ExtremePositive}_{t-1} \times \text{FWD}_{t+1}$ and $\Delta \text{SALES}_t \times \text{ExtremePositive}_{t-1}$ differs from zero. The test statistic is extremely insignificant with an F-statistic of 0.05 and a P-value of 0.819. This result suggests that extreme growth inventory write-down firms’ purchasing behavior is no different than the behavior of average firms. Therefore, we are unable to provide any evidence in support of Hypothesis 5.

The tests of Hypotheses 4 and 5 suggest that extreme growth firms with write-downs tend to carry more inventory than other extreme growth firms. However, we do not find evidence of exuberant inventory purchases among these firms: In fact, inventory policy of an average extreme growth firm with write-downs is statistically indistinguishable from that of a firm experiencing only moderate or no sales growth. Future research may explore whether this finding implies that inventory policies used by moderately growing firms are inappropriate for extreme growth firms. To our knowledge, the theoretical literature has not considered this issue. We can, however, measure the economic impact of write-downs on extreme growth firms. The results are summarized in Section 3.

7 Summary and Future Research

Inventory write-downs that cause substantial losses to firms have been frequently reported in business media during the past few decades. As a result of mismatch between supply and demand, inventory write-downs may have a severe impact on firms’ operating performance. For the 290 firms in our inventory write-down sample, there is a -15.4% return on assets on average in the year of the inventory write-down (the corresponding market-adjusted return is -21.5%). The economic impact of write-downs is so significant that firms need to take such potential adverse events into account when making inventory decisions. Although both practitioners and academics surely understand its importance, there has been little research devoted to the study of (how to manage) inventory write-downs in today’s competitive, fast changing market environment. In this paper, we empirically analyze inventory write-downs and try to derive useful insights to firms in managing inventory and sales growth. Below we summarize the managerial implications from our study and point out some promising directions for future research.
It has been shown that the probability of an inventory write-down increases in sales growth (Hypothesis 1). Intuitively, one may conjecture that extreme growth firms should be concerned about securing supplies to avoid shortages. But our result suggests that these firms should also be alert to the potential risk of inventory write-downs. This is mainly because high sales growth is not sustainable. In particular, we have found empirical evidence that high sales growth is less persistent than moderate sales growth (Hypothesis 2). Thus, managers should carefully factor the lower persistence of high growth in their inventory decisions. Overlooking or under-estimating such an effect will raise the level of excess inventory when the demand rate deteriorates and therefore increase the risk of an inventory write-down. Additionally, investing in information systems to improve demand visibility and forecasting capability seem to be rather important when dealing with unpredictable growth. The quicker the firm can identify a turnaround in the growth pattern, the more likely it can react promptly to avoid potential excess inventory buildup.

The link between sales growth and inventory write-downs inspires us to investigate firms’ inventory purchasing policies under high sales growth. Our analysis shows that no-growth firms generally follow a one-for-one replenishment policy (i.e., they simply replace what they sell). This seems to be consistent with theoretical predictions that simple base-stock type policies perform well under stationary demand environments. With moderate growth, firms on average purchase an additional 11.8% of sales growth. However, Hypothesis 3, which states that high growth firms use a more aggressive purchasing policy, has been rejected by the empirical test. In fact, it has been found that the high growth firms replace less inventory sold than moderate growth firms. Overall we can see that the aggressiveness of firms’ inventory policy is not necessarily monotone in sales growth rate: Although moderate growth firms tend to replace more stock sold than no growth firms, high growth firms behave more conservatively than the moderate growth firms. As discussed in Section 4.2, various arguments from the literature would lead to a conjecture that firms purchase inventory more aggressively when facing fast growth rate. In contrast, we find that, on average, this is not the case.

Why do high growth firms exhibit such conservative ordering behavior? We postulate two non-mutually exclusive stories to answer this question. The first potential story is that managers are aware of the lower persistence of high growth, and therefore take precautionary actions accordingly. The second potential story is that high growth firms appear to behave cautiously with respect to purchases and inventory not because they are cautious but because they cannot meet demand. The high growth in demand results in inventory shortages due to limited production capacity or long supply lead times. These inventory shortages show up in the data as a conservative purchasing policy and lower inventory levels. A prominent example that fits into this story is Cisco Systems: As reported by the media, Cisco experienced huge amount of backlogged demand and tried every means (e.g., ramp up capacity and place orders for parts and components) to catch up with the fast sales growth before the 2001 inventory write-down (Burrows, 2003).

By using cross-firm comparison, our analysis also demonstrates that there is an association between inventory write-downs and firms’ purchasing and inventory holding policies. The acceptance
of Hypothesis 4 is not surprising. It confirms the intuition that under high growth, the write-down firms adopt more aggressive ordering and inventory policies than the no-write-down firms. This finding, though intuitive, offers useful guidelines to firms seeking appropriate inventory strategies under high growth. The overly aggressive policies of the write-down firms could be due to following reasons: First, the write-down firms under-estimate the low persistence of high sales growth; second, the pressure of inventory shortages make the write-down firms over-invest in capacity and raw materials, causing them to overshoot demand; third, the write-down firms are unable to adjust their short-term inventory because they use inflexible ordering systems or have locked in capacity through long-term supply contracts. Therefore, in order to reduce the inventory write-down risk, firms under high growth should be cognizant of the low persistence of sales growth, exert caution when investing in capacity, and ensure that the ordering systems are as nimble and accurate as possible.

The rejection of Hypothesis 5 means that the high growth write-down firms exhibit similar behavior to the moderate growth firms, which is not intuitive. The implication of this result is two-fold. On one hand, it suggests that high growth write-down firms are unable to properly adjust their inventory strategy when entering the high sales growth region. This could be either because they do not know how to do it, or simply because they fail to realize that such an adjustment is necessary and important. Second, the ordering policy that works fine under moderate growth may actually heighten the inventory write-down risk under high growth. Therefore, it is critical for firms to carefully adjust their purchasing policies when switching between different phases of growth (future research may address how to identify each phase). This may require a very flexible and agile supply system. For instance, firms may try to procure from suppliers that can provide responsive deliveries. Instead of locking in long-term supply/capacity to take advantage of low cost, firms may try to use shorter-term contracts with flexible supply terms.

In summary, the above results indicate that growing demand and obsolescence risks present remarkable challenges to operations managers. To meet this challenge, (high growth) firms need to pay attention to both ends of their supply chains. For the customer end, firms should invest in information technology to improve demand transparency throughout the supply chain. This will help firms better forecast future demand changes. For the supplier end, firms should exert effort to shorten supply lead times and negotiate flexible supply contracts. In fact, many high-tech firms have learned the lesson from Cisco’s great inventory correction and are trying to make their supply chains shorter, more transparent and as flexible as possible (Teach, 2001).

Inventory management appears particularly challenging for firms experiencing high sales growth. A more cautious purchasing policy can reduce the potential risk of inventory write-downs. But an overly cautious policy may lead to lost sales and unsatisfactory service level. Future research may address the question of an appropriate ordering policy for high growth firms. This appears to be an under-explored research area in the literature. Song and Zipkin (1996) were among the first to look into inventory control with the prospect of obsolescence (their paper was published in a special issue on new directions in operations management). However, they assume a constant unit
value of inventory and do not consider the impact of inventory write-downs. Graves (1999) studies an inventory model where demand is non-stationary and governed by an autoregressive process. Wagner (2002, p.224) points out the need for useful heuristics: “…these should take notice of rules of thumb that may seem too simple … but may function well in a non-stationary environment that contains [multiple] uncertainties.”
References


### Table 1: Summary statistics for write-down firm years

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inventory Write-down(_t) (millions)</td>
<td>13.208</td>
<td>45.684</td>
<td>0.07</td>
<td>559</td>
<td>290</td>
</tr>
<tr>
<td>WD(_t)</td>
<td>-0.037</td>
<td>0.036</td>
<td>-0.195</td>
<td>-0.01</td>
<td>290</td>
</tr>
<tr>
<td>Total Assets(_t) (millions)</td>
<td>496.027</td>
<td>1347.836</td>
<td>1.985</td>
<td>14964</td>
<td>290</td>
</tr>
<tr>
<td>Sales(_t) (millions)</td>
<td>412.962</td>
<td>991.787</td>
<td>0.996</td>
<td>10542</td>
<td>290</td>
</tr>
<tr>
<td>ROA(_t)</td>
<td>-0.154</td>
<td>0.253</td>
<td>-0.877</td>
<td>0.268</td>
<td>290</td>
</tr>
<tr>
<td>PURCHASES(_t)/COGS(_t)</td>
<td>0.948</td>
<td>0.162</td>
<td>0.651</td>
<td>1.532</td>
<td>290</td>
</tr>
<tr>
<td>∆SALES(_t)</td>
<td>0.027</td>
<td>0.436</td>
<td>-0.545</td>
<td>2.764</td>
<td>290</td>
</tr>
<tr>
<td>INV(_t)</td>
<td>0.203</td>
<td>0.138</td>
<td>0.01</td>
<td>0.648</td>
<td>290</td>
</tr>
<tr>
<td>∆INV(_t)</td>
<td>-0.024</td>
<td>0.081</td>
<td>-0.197</td>
<td>0.281</td>
<td>290</td>
</tr>
<tr>
<td>RET(_t)</td>
<td>-0.023</td>
<td>0.820</td>
<td>-0.982</td>
<td>4.587</td>
<td>285</td>
</tr>
<tr>
<td>EWRET(_t)</td>
<td>-0.215</td>
<td>0.661</td>
<td>-1.636</td>
<td>3.699</td>
<td>285</td>
</tr>
</tbody>
</table>

Sample consists of 290 firms experiencing a first time inventory write-down of more than 1% of average total assets between calendar years 2002 and 2004. WD is measured as the hand collected inventory write-down amount scaled by average total assets. ROA is measured as income before extraordinary items scaled by average total assets. PURCHASES\(_t\)/COGS\(_t\) is measured as purchases (ending inventory plus cost of goods sold minus beginning inventory) divided by cost of goods sold. ∆SALES is measured as the year-over-year percentage change in sales. INV is measured as inventory scaled by total assets. ∆INV is measured as change in inventory scaled by average total assets. RET is the annual buy-hold returns measured beginning four months after the beginning of the fiscal year until three months after the end of the fiscal year. EWRET is the buy-hold annual return less the compounded monthly annual equally-weighted return. All variables with the exception of returns are winsorized at the 1st and 99th percentiles.
The WD sample consists of 274 inventory write-down firms and the matched sample consists of 274 matched firms. Firms are first matched on calendar year then 2-digit SIC codes and then 1-digit SIC codes if a proper 2 digit match is unavailable. Matched firms must be within 20% of total assets and inventory as a percentage of total assets in period \((t-1)\). From the remaining firms that meet the previous criteria, the firm with the closest change in sales to the write-down firm in period \((t-1)\) is selected as a match. ROA is measured as income before extraordinary items scaled by average total assets. \(\Delta SALES\) is measured as the year-over-year percentage change in sales. RET is the annual buy-hold returns measured beginning four months after the beginning of the fiscal year until three months after the end of the fiscal year. All variables with the exception of returns are winsorized at the 1st and 99th percentiles. * Significant at the .10 level ** Significant at the .05 level *** Significant at the .01 level

<table>
<thead>
<tr>
<th></th>
<th>WD Sample</th>
<th></th>
<th>Matched Sample</th>
<th></th>
<th>Difference in Means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
<td>Mean</td>
<td>Median</td>
<td></td>
</tr>
<tr>
<td>TotalAssets(_{t-1})</td>
<td>427.8</td>
<td>123.0</td>
<td>425.6</td>
<td>114.7</td>
<td>2.2</td>
</tr>
<tr>
<td>INV(_{t-1})</td>
<td>0.204</td>
<td>0.185</td>
<td>0.199</td>
<td>0.184</td>
<td>0.006***</td>
</tr>
<tr>
<td>(\Delta SALES(_{t-1}))</td>
<td>0.173</td>
<td>0.031</td>
<td>0.115</td>
<td>0.029</td>
<td>-0.143*</td>
</tr>
<tr>
<td>ROA(_t)</td>
<td>-0.173</td>
<td>-0.095</td>
<td>-0.030</td>
<td>0.030</td>
<td>-0.071***</td>
</tr>
<tr>
<td>RET(_t)</td>
<td>-0.033</td>
<td>-0.235</td>
<td>0.377</td>
<td>0.069</td>
<td>-0.41***</td>
</tr>
</tbody>
</table>
**Table 3: Industry classification for inventory write-down firms and sample firms**

<table>
<thead>
<tr>
<th>Industry</th>
<th>Write-down Sample</th>
<th>Entire Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture, Forestry and Fishing</td>
<td>0.0</td>
<td>1.5</td>
</tr>
<tr>
<td>Mining and Construction</td>
<td>0.7</td>
<td>3.3</td>
</tr>
<tr>
<td>Food and Tobacco</td>
<td>2.1</td>
<td>3.6</td>
</tr>
<tr>
<td>Textile and Apparel</td>
<td>0.7</td>
<td>2.4</td>
</tr>
<tr>
<td>Lumber, Furniture and Paper</td>
<td>1.4</td>
<td>5.1</td>
</tr>
<tr>
<td>Chemicals</td>
<td>1.0</td>
<td>3.6</td>
</tr>
<tr>
<td>Refining and Extractive</td>
<td>0.3</td>
<td>3.5</td>
</tr>
<tr>
<td><strong>Heavy Industry</strong></td>
<td><strong>47.2</strong></td>
<td><strong>34.6</strong></td>
</tr>
<tr>
<td>Computers, Semiconductor and Computer Services</td>
<td><strong>25.2</strong></td>
<td><strong>12.8</strong></td>
</tr>
<tr>
<td>Transport, Pipelines and Telecom</td>
<td>0.3</td>
<td>3.4</td>
</tr>
<tr>
<td>Wholesale, Retail and Restaurants</td>
<td>10.3</td>
<td>15.7</td>
</tr>
<tr>
<td>Services and Conglomerates</td>
<td>5.5</td>
<td>6.8</td>
</tr>
<tr>
<td>Drugs and Medical Equipment</td>
<td>5.2</td>
<td>3.9</td>
</tr>
</tbody>
</table>

Industries in bold are significantly overrepresented in the write-down sample relative to the total sample at the .01 level. Industries are classified based on SIC codes as follows: Agriculture, Forestry and Fishing (0-999), Mining and Construction(1000-1299,1400-1999), Food and Tobacco (2000-2141), Textile and Apparel (2200-2399), Lumber, Furniture and Paper (2400-2796), Chemicals (2800-2824,2840-2899), Refining and Extractive (2900-2999, 1300-1399), Heavy Industry (3000-3569, 3580-3669, 3680-3999), Computers, Semiconductor and Computer Services (7370-7379, 3570-3679), Transport, Pipelines and Telecom (4000-4899), Wholesale, Retail and Restaurants (5000-5999), Services and Conglomerates (7000-7369), Drugs and Medical Equipment (2830-2836, 3829-3851).
Figure 1: Inventory write-down sample from 2002-2004 (N=290)
The entire sample of firms is sorted each year on sales growth and ranked into quintiles. Quintiles are sorted from the highest sales growth on the left to lowest sales growth on the right. The figure above the columns represents the mean inventory write-down scaled by average total assets for firms in that sales growth quintile.

Figure 2: Sales growth quintiles for write-down firms in periods $t$, $(t - 1)$, $(t - 2)$ and $(t - 3)$ relative to write-down
Table 4: Logistic regressions of inventory write-down \((t+1)\) regressed on changes in inventory and sales (2002-2004)

<table>
<thead>
<tr>
<th></th>
<th>(PR(WD = 1)_{t+1})</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>((A))</td>
<td>((B))</td>
<td>((C))</td>
</tr>
<tr>
<td>(\Delta SALES_t)</td>
<td>-.070 (\text{**}) ((.158))</td>
<td>-.205   (\text{**}) ((.205))</td>
<td></td>
</tr>
<tr>
<td>(\Delta SALES_{t-1})</td>
<td>.283*** ((.080))</td>
<td>.190* ((.097))</td>
<td></td>
</tr>
<tr>
<td>(\Delta INV_t)</td>
<td>2.182* ((1.154))</td>
<td>2.708** ((1.240))</td>
<td></td>
</tr>
<tr>
<td>(\Delta INV_{t-1})</td>
<td>4.243*** ((1.035))</td>
<td>3.557*** ((1.146))</td>
<td></td>
</tr>
<tr>
<td>Const.</td>
<td>-3.392*** ((.062))</td>
<td>-3.414*** ((.064))</td>
<td>-3.431*** ((.066))</td>
</tr>
<tr>
<td>Obs.</td>
<td>8192</td>
<td>8223</td>
<td>8184</td>
</tr>
</tbody>
</table>

Sample consists of all non-financial and utility firm-years listed on either NYSE/NASDAQ/AMEX, CRSP and COMPUSTAT between 2002 and 2004 with inventory greater than 1% of average total assets and available data to calculate change in sales data for periods \((t-1), t,\) and \((t+1),\) and purchases to cost of goods sold in period \(t.\) The dependent variable, FWD, equals 1 for the 290 firms experiencing a first-time inventory write-down of more than 1% of average total assets and 0 otherwise. \(\Delta SALES\) is measured as the year-over-year percentage change in sales. \(\Delta INV\) is measured as change in inventory scaled by average total assets. * Significant at the .10 level ** Significant at the .05 level *** Significant at the .01 level.
The entire sample of firms is sorted each year on sales growth and ranked into quintiles. Sample consists of all non-financial and utility firm-years listed on either NYSE/NASDAQ/AMEX, CRSP and COMPUSTAT between 1980 and 2008 with inventory greater than 1% of average total assets and available data to calculate change in sales data for periods \((t - 1), t,\) and \((t + 1)\), and purchases to cost of goods sold in period \(t\).
Table 5: Sales growth persistence

<table>
<thead>
<tr>
<th></th>
<th>(A)</th>
<th>(B)</th>
<th>(C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta SALES_t )</td>
<td>.197***</td>
<td>.361***</td>
<td>.334***</td>
</tr>
<tr>
<td></td>
<td>(.010)</td>
<td>(.031)</td>
<td>(.046)</td>
</tr>
<tr>
<td>( \Delta SALES_t \times ExtremePositive_t )</td>
<td>- .139***</td>
<td>- .111***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.023)</td>
<td>(.041)</td>
<td></td>
</tr>
<tr>
<td>( \Delta SALES_t \times ExtremeNegative_t )</td>
<td>- .412***</td>
<td>- .383***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.045)</td>
<td>(.057)</td>
<td></td>
</tr>
<tr>
<td>( \Delta SALES_t \times HIGH_t )</td>
<td></td>
<td>.036</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.033)</td>
<td></td>
</tr>
<tr>
<td>( \Delta SALES_t \times LOW_t )</td>
<td></td>
<td>- .078</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.076)</td>
<td></td>
</tr>
<tr>
<td>Const.</td>
<td>.080***</td>
<td>.057***</td>
<td>.057***</td>
</tr>
<tr>
<td></td>
<td>(.008)</td>
<td>(.008)</td>
<td>(.008)</td>
</tr>
<tr>
<td>Obs.</td>
<td>87335</td>
<td>87335</td>
<td>87335</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>.062</td>
<td>.067</td>
<td>.067</td>
</tr>
</tbody>
</table>

Eq (C) \( \Delta SALES_t \times ExtremePositive_t = \Delta SALES_t \times HIGH_t \) F-statistic=48.26, P-value=0.00

Eq (C) \( \Delta SALES_t \times ExtremeNegative_t = \Delta SALES_t \times LOW_t \) F-statistic=10.68, P-value=0.00

Sample consists of all non-financial and utility firm-years listed on either NYSE/NASDAQ/AMEX, CRSP and COMPUSTAT between 1980 and 2008 with inventory greater than 1% of average total assets and available data to calculate change in sales data for periods \( (t-1), t, \) and \( (t+1) \), and purchases to cost of goods sold in period \( t \). \( \Delta SALES_t \) is measured as the year-over-year percentage change in sales. \( ExtremePositive_t \) and \( ExtremeNegative_t \) are indicator variables equal to 1 if a firm ranks in the top or bottom quintiles respectively of sales growth in year \( t \) and 0 otherwise. \( High_t \) and \( Low_t \) are indicator variables equal to 1 if a firm ranks in the 4th or 2nd quintile respectively of sales growth in year \( t \) and 0 otherwise. Standard errors are estimated using two-way clustering on firms and years. * Significant at the .10 level ** Significant at the .05 level *** Significant at the .01 level
Table 6: Purchases scaled by cost of goods sold in $t$ regressed on change in $t$ sales

<table>
<thead>
<tr>
<th></th>
<th>(A)</th>
<th>(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta SALES_t$</td>
<td>.093***</td>
<td>.118***</td>
</tr>
<tr>
<td></td>
<td>(.003)</td>
<td>(.004)</td>
</tr>
<tr>
<td>$\Delta SALES_t \times ExtremePositive_{t-1}$</td>
<td></td>
<td>-.032***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.004)</td>
</tr>
<tr>
<td>$\Delta SALES_t \times ExtremeNegative_{t-1}$</td>
<td></td>
<td>-.036***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.004)</td>
</tr>
<tr>
<td>Const.</td>
<td>1.008***</td>
<td>1.007***</td>
</tr>
<tr>
<td></td>
<td>(.002)</td>
<td>(.002)</td>
</tr>
<tr>
<td>Obs.</td>
<td>87335</td>
<td>87335</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.132</td>
<td>.136</td>
</tr>
</tbody>
</table>

Sample consists of all non-financial and utility firm-years listed on either NYSE/NASDAQ/AMEX, CRSP and COMPUSTAT between 1980 and 2008 with inventory greater than 1% of average total assets and available data to calculate change in sales data for periods $(t-1)$, $t$, and $(t+1)$, and purchases to cost of goods sold in period $t$. The dependent variable, $PURCHASES_t/COGS_t$, is measured as purchases (ending inventory plus cost of goods sold minus beginning inventory) divided by cost of goods sold. $\Delta SALES_t$ is measured as the year-over-year percentage change in sales. $ExtremePositive_{t}$ and $ExtremeNegative_{t}$ are indicator variables equal to 1 if a firm ranks in the top or bottom quintiles respectively of sales growth in year $t$ and 0 otherwise. Standard errors are estimated using two-way clustering on firms and years. * Significant at the .10 level ** Significant at the .05 level *** Significant at the .01 level
Table 7: Purchases scaled by cost of goods sold in $t$ regressed on change in $t$ sales

\[
\frac{PURCHASES_t}{COGS_t}
\]

<table>
<thead>
<tr>
<th></th>
<th>(A)</th>
<th>(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta SALES_t$</td>
<td>.093***</td>
<td>.132***</td>
</tr>
<tr>
<td></td>
<td>(.003)</td>
<td>(.008)</td>
</tr>
<tr>
<td>$\Delta SALES_t \times ExtremePositive_t$</td>
<td></td>
<td>-.053***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.008)</td>
</tr>
<tr>
<td>$\Delta SALES_t \times ExtremeNegative_t$</td>
<td>.089***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.009)</td>
<td></td>
</tr>
<tr>
<td>Const.</td>
<td>1.008***</td>
<td>1.012***</td>
</tr>
<tr>
<td></td>
<td>(.002)</td>
<td>(.002)</td>
</tr>
<tr>
<td>Obs.</td>
<td>87335</td>
<td>87335</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.132</td>
<td>.15</td>
</tr>
</tbody>
</table>

Sample consists of all non-financial and utility firm-years listed on either NYSE/NASDAQ/AMEX, CRSP and COMPUSTAT between 1980 and 2008 with inventory greater than 1% of average total assets and available data to calculate change in sales data for periods $t-1$, $t$, and $t+1$, and purchases to cost of goods sold in period $t$. The dependent variable, $\frac{PURCHASES_t}{COGS_t}$, is measured as purchases (ending inventory plus cost of goods sold minus beginning inventory) divided by cost of goods sold. $\Delta SALES_t$ is measured as the year-over-year percentage change in sales. $ExtremePositive_t$ and $ExtremeNegative_t$ are indicator variables equal to 1 if a firm ranks in the top or bottom quintiles respectively of sales growth in year $t$ and 0 otherwise. Standard errors are estimated using two-way clustering on firms and years. * Significant at the .10 level ** Significant at the .05 level *** Significant at the .01 level.
Table 8: Purchases scaled by cost of goods sold in $t$ regressed on change in $t$ sales, extreme growth indicators for $(t - 1)$ and write-down indicators for $(t + 1)$

<table>
<thead>
<tr>
<th></th>
<th>(A)</th>
<th>(B)</th>
<th>(C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta SALES_t$</td>
<td>.074*** (.005)</td>
<td>.110*** (.006)</td>
<td>.110*** (.006)</td>
</tr>
<tr>
<td>$\Delta SALES_t \times \text{ExtremePositive}_{t-1}$</td>
<td>-.037*** (.010)</td>
<td>-.040*** (.010)</td>
<td></td>
</tr>
<tr>
<td>$\Delta SALES_t \times \text{ExtremeNegative}_{t-1}$</td>
<td>-.050*** (.017)</td>
<td>-.049*** (.018)</td>
<td></td>
</tr>
<tr>
<td>$\Delta SALES_t \times \text{ExtremePositive}<em>{t-1} \times WD</em>{t+1}$</td>
<td></td>
<td></td>
<td>.043*** (.016)</td>
</tr>
<tr>
<td>$\Delta SALES_t \times \text{ExtremeNegative}<em>{t-1} \times WD</em>{t+1}$</td>
<td></td>
<td></td>
<td>-.031* (.018)</td>
</tr>
<tr>
<td>Const.</td>
<td>.997*** (.004)</td>
<td>.996*** (.003)</td>
<td>.996*** (.003)</td>
</tr>
<tr>
<td>Obs.</td>
<td>8386</td>
<td>8174</td>
<td>8174</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.129</td>
<td>.121</td>
<td>.122</td>
</tr>
</tbody>
</table>

Sample consists of all non-financial and utility firm-years listed on either NYSE/NASDAQ/AMEX, CRSP and COMPUSTAT between 2002 and 2004 with inventory greater than 1% of average total assets and available data to calculate change in sales data for periods $(t - 1)$, $t$, and $(t + 1)$, and purchases to cost of goods sold in period $t$. The dependent variable, $PURCHASES_t/COGS_t$, is measured as purchases (ending inventory plus cost of goods sold minus beginning inventory) divided by cost of goods sold. $\Delta SALES_t$ is measured as the year-over-year percentage change in sales. $\text{ExtremePositive}_{t-1}$ and $\text{ExtremeNegative}_{t-1}$ are indicator variables equal to 1 if a firm ranks in the top or bottom quintiles respectively of sales growth in year $(t - 1)$ and 0 otherwise. $WD_{t+1}$ is an indicator variable equal to 1 if a firm experiences a first time inventory write-down of at least 1% of total assets in period $(t + 1)$ and 0 otherwise. Standard errors are estimated using two-way clustering on firms and years. * Significant at the .10 level ** Significant at the .05 level *** Significant at the .01 level.
Table 9: Purchases scaled by cost of goods sold in \( t \) regressed on change in \( t \) sales, extreme growth indicators for \( t \) and write-down indicators for \((t+1)\)

<table>
<thead>
<tr>
<th>( \Delta SALES_t )</th>
<th>(A)</th>
<th>(B)</th>
<th>(C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>.074***</td>
<td>.148***</td>
<td>.148***</td>
<td></td>
</tr>
<tr>
<td>(.005)</td>
<td>(.025)</td>
<td>(.025)</td>
<td></td>
</tr>
<tr>
<td>( \Delta SALES_t \times ExtremePositive_t )</td>
<td>- .091***</td>
<td>- .091***</td>
<td></td>
</tr>
<tr>
<td>(.025)</td>
<td>(.025)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \Delta SALES_t \times ExtremeNegative_t )</td>
<td>.094***</td>
<td>.103***</td>
<td></td>
</tr>
<tr>
<td>(.018)</td>
<td>(.016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \Delta SALES_t \times ExtremePositive_t \times WD_{t+1} )</td>
<td>.007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(.010)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \Delta SALES_t \times ExtremeNegative_t \times WD_{t+1} )</td>
<td>- .159***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(.028)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Const.</td>
<td>.997***</td>
<td>1.005***</td>
<td>1.005***</td>
</tr>
<tr>
<td>(.004)</td>
<td>(.002)</td>
<td>(.002)</td>
<td></td>
</tr>
<tr>
<td>Obs.</td>
<td>8386</td>
<td>8386</td>
<td>8386</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>.129</td>
<td>.173</td>
<td>.175</td>
</tr>
</tbody>
</table>

\( Eq \ (C) \ \Delta SALES_t \times ExtremePositive_t + \Delta SALES_t \times ExtremePositive_{t-1} \times WD_{t+1} = 0 \), \( F \)-statistic=9.26, \( P \)-value=0.002

Sample consists of all non-financial and utility firm-years listed on either NYSE/NASDAQ/AMEX, CRSP and COMPUSTAT between 2002 and 2004 with inventory greater than 1% of average total assets and available data to calculate change in sales data for periods \((t-1), t, \) and \((t+1), \) and purchases to cost of goods sold in period \( t \). The dependent variable, \( PURCHASES_t / COGS_t \), is measured as purchases (ending inventory plus cost of goods sold minus beginning inventory) divided by cost of goods sold. \( \Delta SALES_t \) is measured as the year-over-year percentage change in sales. \( \Delta SALES_t \) is measured as the year-over-year percentage change in sales. \( ExtremePositive_t \) and \( ExtremeNegative_t \) are indicator variables equal to 1 if a firm ranks in the top or bottom quintiles respectively of sales growth in year \( t \) and 0 otherwise. \( WD_{t+1} \) is an indicator variable equal to 1 if a firm experiences a first time inventory write-down of at least 1% of total assets in period \((t+1)\) and 0 otherwise. Standard errors are estimated using two-way clustering on firms and years. Standard errors are estimated using two-way clustering on firms and years. * Significant at the .10 level ** Significant at the .05 level *** Significant at the .01 level
Title: Single-stage approximations for optimal policies in serial inventory systems with non-stationary demand

Author: Kevin Shang

Affiliation: Fuqua School of Business, Duke University, Durham, NC 27708

Contact information: khshang@duke.edu, 919-660-7965
Companies often face non-stationary demand due to product life cycles and seasonality, and this non-stationary demand complicates supply-chain managers’ inventory decisions. The present paper suggests a simple heuristic for determining stocking levels in a serial inventory system. Unlike the exact optimization algorithm, the heuristic can generate a near-optimal solution by solving independent, single-stage systems with the original problem data. This result enables us to reduce the computation time and complexity by allowing each location to simultaneously solve its own inventory problem. We further examine myopic solutions for these single-stage systems. Specifically, in a numerical study, we find that the change of the myopic solution is consistent with that of the optimal solution. We then derive a closed-form expression for the myopic solution and use it to approximate the optimal local base-stock level and to gain insights into how to manage safety stocks. The closed-form approximation shows how the local base-stock level is affected by future demand; it also explains the observation that the safety stock at an upstream stage is often stable and may not increase when demand variability increases. Finally, we discuss how the heuristic leads to a coordination scheme that enables a decentralized supply chain to achieve the near-optimal solution.

1. Introduction

Customer demand is often non-stationary in practice. Causes of non-stationary demand include product life cycles, seasonality, trends, and economic conditions. Non-stationary demand makes it difficult for managers to determine optimal stocking levels in a supply chain. One difficulty is computational – finding the optimal system-wide solution often requires solving several interrelated, recursive cost functions between stages across time. The other difficulty concerns implementation: if a supply chain is composed of independent firms, these firms may not be willing to implement the optimal solution without appropriate incentives. In a non-stationary demand environment, the optimal solution is often time-varying. Thus, designing an incentive scheme that can induce each stage to choose the optimal stocking level in each time period poses a difficult challenge.

This paper provides a heuristic that can simplify the computation and help facilitate the implementation of the optimal solution. To illustrate the approach, we consider an N-stage, serial inventory model with a finite horizon. Materials flow from stage N to stage N − 1, N − 1 to N − 2, etc. until stage 1, where in each period a random, non-stationary demand occurs. Clark and Scarf (1960) first study this model with N = 2. They show that (time-varying) echelon base-stock policies are optimal. Although the structure of the policy is simple, obtaining the optimal solution
is quite complex, especially for the upstream stage. This is because the optimal value function for the upstream stage depends on the downstream stage’s optimal base-stock level. Thus, one has to solve two sets of optimal value functions simultaneously. Clearly, the complexity will grow quickly when the chain becomes longer because the optimal value function of an upstream stage depends on all of its downstream solutions.

The heuristic we propose can generate an approximation for stage $j$ ($1 \leq j \leq N$) without knowing stage $i$’s base-stock level, $i < j$. More specifically, we show that the optimal base-stock level for stage $j$ is bounded by the optimal solutions of two single-stage systems with the original problem data. This result is established in three steps. First, we show that the optimal value function for stage $j$ is bounded above and below by that of a revised $j$-stage system. We refer to these revised systems as the upper-bound system and the lower-bound system, respectively. The upper-bound system is constructed by requiring stage $i (< j)$ to always order up to stage $i+1$’s echelon inventory level in each period. On the other hand, the lower-bound system is constructed by regulating stage $i$’s holding and order cost parameters. Second, we show that the optimal base-stock level for stage $j$ in the upper-bound (lower-bound) system is a lower (upper) bound to that in the original system. Finally, we show that solving these revised $j$-stage systems is equivalent to solving a single-stage system with parameters obtained from the original problem data. A numerical study suggests that the gap between these two solution bounds is generally quite small. This result motivates us to propose a heuristic solution for each stage by solving a single-stage system with a weighted average of the cost parameters obtained from the upper- and the lower-bound systems. In a numerical study, we find that the heuristic generates 58% of the optimal solutions, and that 98% of the heuristic solutions are within the $\pm 1$ unit range of the optimal solutions.

The results described in this paper make three contributions. First, the suggested heuristic can generate a near-optimal solution for a stage without knowing the base-stock levels of its downstream stages. This separation feature not only simplifies the computation, but also shortens the computation time by allowing each stage to solve its own problem in parallel. In other words, the heuristic can generate a solution at least $N$ times faster than the exact algorithm, provided that parallel processing is possible. Second, the effectiveness of the heuristic solution motivates us to investigate the corresponding myopic solution (an upper bound to the heuristic solution; see, for example, Zipkin 2000). We find that the change of the myopic solution obtained from the heuristic single-stage system is consistent with that of the optimal solution when the system parameters vary. Thus, we can provide a closed-from expression for the myopic solution to approximate the optimal
local base-stock level. The closed-form expression clearly shows how the system parameters affect the local base-stock level and the safety stock. It also explains why the safety stock of an upstream stage is often stable and may not increase when the variance of the demand increases over time. Finally, for a supply chain consisting of independent, self-interested firms (i.e., decentralized control system), we show that our heuristic can lead to a remarkably simple, time-consistent contract that induces each stage to choose the near-optimal solution.

Several researchers have provided methods to simplify the computation for the Clark and Scarf model. Federgruen and Zipkin (1984) consider an infinite-horizon version of the model with i.i.d. demand and show that the optimal policy can be obtained by recursively solving two cost functions that have the form of a single-period problem. Chen and Zheng (1994) reinterpret Federgruen’s and Zipkin’s results, simplify the optimality proof, and present an optimization algorithm to facilitate the computation. Although the computation effort is much reduced with these results, finding an upstream solution still is not easy because it depends on the downstream solutions. Thus, there is a stream of research that aims to further simply the computation and to reveal insights by solving single-stage problems. Noteworthy examples include Dong and Lee (2003), Shang and Song (2003), Gallego and Özer (2005), and Chao and Zhou (2007). To our knowledge, the existing heuristic solutions for the Clark-Scarf model are all for the infinite-horizon problem. Given that the Clark-Scarf model is a building block for many supply chain models and that non-stationary demand is prevalent in practice, the solution approach presented here has wide applicability.

Several papers have derived solutions for practical issues in supply chains under non-stationary demand. Erkip et al. (1990), for example, consider a one-depot-multi-warehouse system in which the warehouses’ demands are correlated. They derive an expression for the optimal safety stock as a function of the level of correlation through time. Ettl et al. (2000) consider a supply chain network that implements base-stock policies subject to service level requirements. They approximate the lead time demand for each location and suggest a rolling-horizon approach to find the base-stock levels for the non-stationary demand case. Abhyankar and Graves (2001) consider a two-stage serial system with a Markov-modulated Poisson demand process. They implement an inventory hedging policy to protect against cyclic demand variability. Graves and Willems (2008) consider a problem of allocating safety stocks in a supply chain network where the demand is bounded and there is a guaranteed service time between stages and customers. They propose an algorithm to determine safety stocks under a constant service time policy. Similar to the above papers, the present work aims to provide a simple control policy.
Finally, our paper is also related to the coordination literature. Most coordination papers consider an infinite-horizon model with stationary demand. Because of the regenerative process, these infinite-horizon models are equivalent to single-period problems. These coordination papers often analyze a decentralized Nash equilibrium solution and provide contracts to induce the system to achieve the centralized (first best) solution. Noteworthy examples include Lee and Whang (1999), Chen (1999), Cachon and Zipkin (1999), and Shang et al. (2009). However, when the system fails to form a regenerative process, studying the decentralized behaviors becomes more difficult. Donohue (2000) studies a two-period model with demand forecasts. She suggests using time-varying contract terms to coordinate the system. Parker and Kapuscinski (2010) consider a two-stage serial inventory system with capacity limits, where each stage aims to minimize its own costs. They show that there exists a Markov equilibrium policy for a dynamic game in the decentralized control system. In general, it is very difficult to derive a coordination contract in a finite-horizon model. Were such a contract to exist, it would be too difficult to implement because the contract terms are often time-varying.

2. Model and Main Results

We consider an $N$-stage serial inventory system, where stage 1 orders from stage 2, stage 2 from stage 3, etc. until stage $N$, which orders from an ample outside supplier. There is a lead time $\tau_j$ between stage $j$ and stage $j+1$, and $\tau_j$ is a positive integer. Denote $\tau[i,j] = \sum_{k=i}^{j} \tau_k$ and $\tau[i,j] = 0$ if $i > j$. Let $h_j$ be the echelon holding cost rate at stage $j$ and let $b$ be the backorder cost rate at stage 1. Let $p_j$ as the unit order cost for stage $j$. We use $t$ to index the time period and count the time backwards. That is, if $t$ is the current period, $t-1$ will be the next period, etc. Let $T$ be the planning horizon. Denote $D(t)$ the demand in period $t$. The demands are independent between periods, but the demand distributions may differ from period to period. Let $D[t,s] = \sum_{i=s}^{t} D(i)$, representing the total demand in period $t$, $t-1$, $t-2$, ..., $s$, where $t \geq s$. The sequence of events in a period is as follows: (1) each stage receives a shipment sent one period ago from its upstream stage at the beginning of the period; (2) each stage places an order at the beginning of the period; (3) each stage sends a shipment to its downstream stage; (4) demand occurs at stage 1 during the period; (5) inventory cost is evaluated at the end of the period.

Clark and Scarf (1960) show that time-varying, echelon base-stock policies are optimal. Let
be the optimal echelon base-stock level for stage \( j \). The echelon base-stock policy is executed as follows: stage \( j \) reviews \( x_j \) at the beginning of each period, where \( x_j \) is the echelon inventory level for stage \( j \) (= inventory at stage \( j \) + inventory in-transit to and at stage \( i(< j) \) - backorders). The stage orders up to \( s_j(t) \) if \( x_j < s_j(t) \), and does not order otherwise. It is well known that \( s_j(t) \) depends on the downstream base-stock levels and can be found by solving \( j \) sets of dynamic programs sequentially. More specifically, finding \( s_1(t) \) is equivalent to solving a single-stage system. With the known \( s_1(t) \), one can form a dynamic program to compute the functional equation for stage 2, assuming that stage 2 has ample supply. The optimal base-stock level \( s_2(t) \) is the optimal solution obtained from the functional equation. Continuing this procedure, with the known \( s_i(t) \), \( i < j \), one can compute the functional equation for stage \( j \) and the corresponding optimal solution \( s_j(t) \). See Appendix A for the detailed algorithm.\(^1\) In other words, when finding the optimal solution for stage \( j \), we can only focus on echelon \( j \) that includes stage 1 to stage \( j \) by viewing the echelon has an ample supply from its upstream stage.

Below we demonstrate that \( s_j(t) \) is bounded by the solution obtained from two single-stage systems with the original problem data. The lower-bound (upper-bound) solution is generated from an upper-bound (lower-bound) system presented in §2.1 (§2.2). We only present the main results and refer the reader to Appendix B for proofs.

### 2.1 Upper-Bound System

Consider echelon \( j \) with a more restrictive policy: stage \( i \) always orders up to \( x_{i+1} \) in each period except \( t \leq \tau[1, i] \) for \( i < j \). (When \( t \) is in this interval, stage \( i \) would not order because of the end of the horizon.) Let \( s'_j(t) \) be the resulting optimal echelon base-stock level for stage \( j \). Clearly, such a policy is suboptimal and the resulting total cost is an upper bound to that of the original system. For this reason, we call this restrictive system the upper-bound system. It is not clear, however, whether there exists an order relationship between \( s'_j(t) \) and the optimal solution \( s_j(t) \).

To see this, consider a two-stage system. Under this restrictive ordering policy, stage 1 will order no less than it does in the original system. To minimize the cost in the upper-bound system, stage 2 should order less. Thus, it is not clear whether the net effect of the restrictive policy will be to make the resulting echelon base-stock level \( s'_2(t) \) smaller than \( s_2(t) \). This uncertainty is particularly great when the demand changes drastically between periods. Nevertheless, we show that the order

\(^1\)Clark and Scarf (1960) develop the dynamic program for the stage cost by introducing an induced penalty cost function (a penalty cost charged to an upstream stage for insufficient inventory supply). We revise their algorithm by considering the echelon cost in the dynamic program. This step is necessary for our analysis.
relationship exists.

**Theorem 1** If \( x_{j-1} < s_{j-1}(t), \ s_j(t) \geq s_j^f(t) \) for \( t > \tau[1,j] \).

Let us take a closer look at the upper-bound system. To construct the upper-bound system for echelon \( j, j = 2, \ldots, N \), we regulate stage \( i(<j) \) to always order up to \( x_{i+1} \) in each period. By doing so, any unit ordered by stage \( j \) will eventually arrive at stage 1 in \( \tau[1,j] \) periods. Thus, we can see each stage \( i, i = 2, \ldots, j \) is a transit point and that the echelon \( j \) is effectively a single-stage system with a lead time of \( \overline{T}_j = \tau[1,j] \) periods. Next, let us consider the unit order cost. For each unit ordered by stage \( j \) in period \( t, p_j \) is incurred. This unit will arrive at stage \( j \) in period \( t - \tau_j \), during which stage \( j - 1 \) will order it with an order cost of \( p_{j-1} \). In addition, stage \( j \) will incur a holding cost \( h_j \) for this unit in period \( t - \tau_j \). Thus, the effective order cost is \( (p_{j-1} + h_j) \). Continuing this logic, this unit will arrive at stage 2 in period \( t - \tau[2,j] \), during which stage 1 will incur an order cost of \( p_1 \) and stage 2 will incur a holding cost \( h_2 \) for this unit. In summary, the unit order cost for the upper-bound system for echelon \( j, j = 2, \ldots, N \), is

\[
p_j^u = p_j + \sum_{i=2}^{j} \alpha\tau[i,j](p_{i-1} + h_i).
\]

Once this unit arrives at stage 1, it will kept as inventory. Thus, it will incur either the local holding cost rate \( h_j^u = h_1[j,j] \) or the local effective backorder cost rate \( b_j = b + h[j + 1, N] \) (see Shang and Song (2003) for an explanation of the formulation of \( b_j \)).

To summarize, let \( S_j^u[p_j^u, h_j^u, b_j, \overline{T}_j] \) denote the upper-bound system for echelon \( j, j = 2, \ldots, N \), where \( p_j^u \) is the unit order cost, \( h_j^u \) the holding cost rate, \( b_j \) the backorder cost rate, and \( \overline{T}_j \) the lead time. Set \( h_{j+1} = 0 \).

**Proposition 2** The lower-bound solution \( s_j^l(t) \) can be obtained by solving \( S_j^l[p_j^l, h_j^l, b_j, \overline{T}_j] \), for \( j = 1, \ldots, N \).

### 2.2 Lower-Bound System

The approach of constructing the lower-bound system for echelon \( j \) is different from that of constructing the upper-bound system. More specifically, we set \( h_i = 0 \) and \( p_i = 0 \) for \( i < j \). Clearly, the resulting cost is a lower-bound to echelon \( j \)'s cost. Under this construction, stage \( i \) would choose to order up to \( x_{i+1} \) as there is no intention to carry inventory at stage \( i + 1 \). Thus, it turns out that the optimal order policies are the same for the upper- and the lower-bound systems. However, as in
the upper-bound system, it is not clear whether an order relationship exists between the resulting solution, \( s_u^j(t) \), and the optimal solution \( s_j(t) \). Consider a two-stage system. Under such new cost parameters, stage 1 will always order up to \( x_2 \). If we follow the same logic as before, stage 2 should order less. Thus, there is little intuition that \( s_u^2(t) \) is larger than \( s_2^j(t) \). The following theorem shows that, indeed, the order relationship holds.

**Theorem 3** \( s_u^j(t) \geq s_j(t) \), for \( t > \tau[1,j] \).

Since the downstream stages use the same optimal policy, a similar logic can be applied to show that the system is equivalent to a single-stage system. We can simply set \( h_i = 0 \) and \( p_i = 0 \) for \( i < j \) to obtain the cost parameters for the lower-bound system. More specifically, let \( S_l^j[p_l^j, h_l^j, b_j, \Xi_j] \) denote the lower-bound system for echelon \( j \), where the order cost \( p_l^j = p_j + \alpha \tau[j,j] \), the holding cost rate \( h_l^j = h_j \), and the effective backorder cost rate \( b_j \), and the lead time \( \Xi_j \) periods.

**Proposition 4** The upper-bound solution \( s_u^j(t) \) can be obtained by solving \( S_l^j[p_l^j, h_l^j, b_j, \Xi_j] \), for \( j = 1, ..., N \).

### 3. Myopic Solution

For managers, it is crucial to learn how the system parameters affect the stocking decision in a supply chain. This section provides a closed-form expression for the optimal local base-stock level and the safety stock at each stage.

We propose using a myopic solution. More specifically, in §4, we suggest a heuristic that solves a single-stage system with a weighted average of the cost parameters obtained from the upper- and lower-bound systems. More specifically, let \( p_a^j = wp_u^j + (1 - w)p_l^j \) and \( h_a^j = wh_u^j + (1 - w)h_l^j \), where \( 0 \leq w \leq 1 \). We call the resulting single-stage system heuristic system \( j \), denoted by \( S_h^j(p_a^j, h_a^j, b_j, \Xi_j) \) and we define the resulting optimal solution as \( s_h^j(t) \). In §4, we numerically show that \( s_h^j(t) \) is an effective approximation to the optimal solution \( s_j(t) \). Furthermore, it is well known that the myopic solution is an upper bound to the optimal solution of a single-stage system (Zipkin 2000, p. 378-379). These two results together motivate us to derive a closed-form approximation, referred to as \( s_m^j(t) \), for \( s_h^j(t) \). In our numerical study, we find that in all cases \( s_m^j(t) \) moves in the same direction as the optimal solution \( s_j(t) \). Thus, we can use the myopic solution to derive an approximation for the optimal local base-stock level to examine the system behaviors. Below we lay out the detailed steps.

Let the myopic solution for \( S_h^j(p_a^j, h_a^j, b_j, \Xi_j) \) be \( s_m^j(t) \).
Proposition 5 For $t > \tau_j$, 

$$s_j^m(t) = \arg \min_{s_j} \left\{ P \left( D[t, t - \tau_j] \leq s_j \right) > \beta_j \right\},$$

where 

$$\beta_j = \frac{\alpha^\tau_j b_j - p^\mu_j (1 - \alpha)}{\alpha^\tau_j (b_j + h^\alpha_j)}.$$ 

Note that when $t < \tau_j$, stage $j$ will not order. When $t = \tau_j + 1$, $s_j^m(t)$ is equal to the solution obtained from the above equation except for the removal of the term $(1 - \alpha)$ in the numerator due to the termination value being equal to zero.

To obtain a closed-form expression, we apply normal approximation on $D[t, t - \tau_j]$. Let the mean of $D[t, t - \tau_j]$ be $\lambda[t, t - \tau_j]$ and the standard deviation be $\sigma[t, t - \tau_j] = \sqrt{\text{Var}[D[t, t - \tau_j]]}$. We can form a closed-form expression for $s_j^m(t)$: 

$$s_j^m(t) = \lambda[t, t - \tau_j] + \sigma[t, t - \tau_j] \Phi^{-1}(\beta_j).$$

Thus, the local base-stock level is $s_j^{m'}(t) = s_j^m(t)$, and for $j = 2, ..., N$, 

$$s_j^{m'}(t) = s_j^m(t) - s_{j-1}^m(t) = \lambda[t - \tau_{j-1}, t - \tau_j] + \sigma[t, t - \tau_j] \Phi^{-1}(\beta_j) - \sigma[t, t - \tau_{j-1}] \Phi^{-1}(\beta_{j-1}). \quad (1)$$

The first term in Equation (1) is the average pipeline inventory in period $t$, which depends on the average $\tau_j$ periods of future demand in period $t - \tau_{j-1} - 1, t - \tau_{j-1} - 2, ..., t - \tau_j$. The second term is the safety stock for stage $j$ in period $t$, denoted as $ss_j^m(t)$, which depends on the cost ratios $\beta_j$ and $\beta_{j-1}$, and the variability of the demand in period $[t, t - \tau_j]$.

Equation (1) allows us to analytically investigate how the system parameters affect the optimal base-stock level and the safety stock at each stage. For example, if we are interested in the change to the amount of safety stock of the upstream stage in a two-stage system, we can define the change to stage 2’s safety stock in period $t$ as 

$$\Delta ss_2^m(t) = ss_2^m(t - 1) - ss_2^m(t)$$

$$= \left( \sigma[t - 1, t - 1 - \tau_2] - \sigma[t, t - \tau_2] \right) \Phi^{-1}(\beta_2) - \left( \sigma[t - 1, t - 1 - \tau_1] - \sigma[t, t - \tau_1] \right) \Phi^{-1}(\beta_1). \quad (2)$$

From the above equation, we can see that $\Delta ss_2^m(t)$ will be fairly small unless there is a significant difference between $\text{Var}[D(t - 1 - \tau_2)]$ and $\text{Var}[D(t - 1 - \tau_1)]$. This implies that the safety stock at
the upstream stage should be fairly stable. (A similar conclusion is observed in Graves and Willems (2008) in their numerical study.) In addition, \( \Delta s_{ss}^m(t) \) may not be positive even if \( \text{Var}[D(t)] < \text{Var}[D(t-1)] \), \( \forall t \). More specifically, when either \( p_2 \) is large, or \( h_2 \) is large, or \( b \) is small, \( \Phi^{-1}(\beta_2) \) tends to be smaller than \( \Phi^{-1}(\beta_1) \), causing the difference in (2) to become negative even when the demand variance increases over time. This suggests that the safety stock at an upstream stage may not increase with the demand variability.

**Example.** We consider a two-stage system with \( \tau_1 = \tau_2 = 1 \), \( p_2 = 6 \), \( h_2 = 1 \), \( p_1 = 4 \), \( h_1 = 1 \), and \( b = 15 \). The demand follows a Poisson distribution with mean rate shown in Table 1. We report the optimal echelon, heuristic, and myopic base-stock levels in each period. We also report the optimal local base-stock level as well as the corresponding safety stock for stage 2. As can be seen, the safety stock may decrease although the demand rate increases (e.g., from \( t = 10 \) to \( t = 9 \)).

<table>
<thead>
<tr>
<th>Period ( (t) )</th>
<th>Demand rate</th>
<th>10</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_1(t) )</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>24</td>
<td>26</td>
<td>22</td>
<td>16</td>
<td>10</td>
<td>5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( s_2(t) )</td>
<td>16</td>
<td>22</td>
<td>29</td>
<td>33</td>
<td>31</td>
<td>24</td>
<td>16</td>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( s_1^m(t) )</td>
<td>16</td>
<td>23</td>
<td>29</td>
<td>33</td>
<td>31</td>
<td>24</td>
<td>16</td>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( s_2^m(t) )</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>24</td>
<td>26</td>
<td>23</td>
<td>18</td>
<td>12</td>
<td>7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( s_2'(t) = s_2(t) - s_1(t) )</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( ss_2(t) = s_2(t) - E[D[t-2,t-2]] )</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: A two-stage example with the optimal, heuristic, and myopic solutions, as well as the local base-stock levels and the resulting safety stocks.

4. **Numerical Study**

We first conduct a numerical study to report the gap between \( s_j^e(t) \) and \( s_j^u(t) \) and examine where within this interval the optimal solution \( s_j(t) \) is located. We seek to determine an effective weight \( w \) that will be used for generating the heuristic solution.

We consider two-stage systems with a horizon of \( T = 10 \) periods. Assume that the period demand follows a Poisson distribution with rate \( \lambda(t) \) in period \( t \). We test the following demand patterns: constant (C), linear increasing (I), linear decreasing (D), concave (V), and convex (X) forms. For the constant demand, we set \( \lambda(t) = 5.5 \), for \( 1 \leq t \leq 10 \); for the increasing demand, \( \lambda(t) = 11 - t \), for \( 1 \leq t \leq 10 \); for the decreasing demand, \( \lambda(t) = t \), for \( 1 \leq t \leq 10 \); for the convex
demand, \( \lambda(t) = 11 - 2t \) for \( 1 \leq t \leq 5 \) and \( \lambda(t) = 2t - 10 \) for \( 6 \leq t \leq 10 \); finally, for the concave demand, \( \lambda(t) = 2t - 1 \) for \( 1 \leq t \leq 5 \), and \( \lambda(t) = 22 - 2t \) for \( 6 \leq t \leq 10 \).

We first fix the cost parameters at stage 1 and change the cost parameters at stage 2. Specifically, we set \( h_1 = 1 \), \( p_1 = 4 \), and \( h_2 \in \{0.5, 1, 1.5\} \), \( p_2 \in \{2, 6\} \). The other parameters are \( \tau_1 = \tau_2 = 1 \) and \( b \in \{15, 50\} \). The total number of instances is 60. We then swap the stage index in the above order and holding cost parameters to generate another 60 instances. The total number of instances is 120 and the total number of optimal base-stock levels for stage 2 is 960.

To examine the overall effectiveness of the bounds, we define the following two measures. Let

\[
\xi = \frac{s^u_2(t) - s^l_2(t)}{s_2(t)}, \quad \theta = \frac{s_2(t) - s^l_2(t)}{s^u_2(t) - s^l_2(t)}.
\]

The first measure \( \xi \) signifies the size of the gap between the solution bounds with respect to the optimal base-stock level; the second measure \( \theta \) signifies the position of the optimal base-stock level between the solution bounds. When \( s^u_2(t) = s^l_2(t) \), we set \( \theta = 0.5 \).

We find that a total of 533 lower bound solutions and a total of 122 upper bound solutions are equal to \( s_2(t) \). Overall, \( s_2(t) \) tends to be closer to \( s^l_2(t) \). But when \( b \) increases, \( s_2(t) \) tends to move toward \( s^u_2(t) \). Also, the average gap is smaller when \( b \) is large. The average \( \xi \) and \( \theta \) for the 960 solutions are 8.50% and 0.27, respectively. Because \( \theta \) is roughly equal to 0.3, we then set \( w = 0.7 \) to generate the heuristic solution.

To examine the performance of the resulting heuristic solution \( s^h_2(t) \), and the myopic solution \( s^m_2(t) \), we solve the same 120 two-stage instances above. Let

\[
\epsilon_i = \left| \frac{s^i_2(t) - s_2(t)}{s_2(t)} \right| \times 100\%, \quad i \in \{a, m\},
\]

which denotes the performance of the heuristic and myopic solution, respectively. Table 2 summarizes the average \( \epsilon \). The parenthesis shows the number of optimal solutions generated from the heuristic and the myopic solutions. Overall, the heuristic generates 769 optimal solutions (80.1%). Interestingly, the myopic solution is also quite effective - it generates 649 optimal solutions (67.6%). The myopic solution is particularly effective when demand rate is increasing or constant. This is intuitive. When demand rate increases, it is more likely that the inventory will be consumed at the end of a period. Thus, the manager can determine an effective order quantity without considering the long-term effect of the left-over inventory. On the other hand, we observe two situations in which the myopic solution is least effective: (1) when the demand is decreasing over time and (2) a few periods before the end of the horizon.
Table 2: Summary of the effectiveness of the solution bounds

<table>
<thead>
<tr>
<th>b</th>
<th>Demand Form</th>
<th>C</th>
<th>I</th>
<th>D</th>
<th>V</th>
<th>X</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>Avg $\epsilon_a$ (#)</td>
<td>1.27% (76)</td>
<td>1.76% (69)</td>
<td>2.56% (80)</td>
<td>1.85% (76)</td>
<td>1.98% (67)</td>
<td>1.88% (368)</td>
</tr>
<tr>
<td></td>
<td>Avg $\epsilon_m$ (#)</td>
<td>1.97% (70)</td>
<td>1.91% (66)</td>
<td>6.75% (52)</td>
<td>5.18% (48)</td>
<td>2.34% (60)</td>
<td>3.63% (296)</td>
</tr>
<tr>
<td>50</td>
<td>Avg $\epsilon_a$ (#)</td>
<td>0.66% (81)</td>
<td>0.52% (86)</td>
<td>0.84% (79)</td>
<td>0.67% (80)</td>
<td>1.26% (75)</td>
<td>0.79% (401)</td>
</tr>
<tr>
<td></td>
<td>Avg $\epsilon_m$ (#)</td>
<td>0.89% (76)</td>
<td>0.52% (86)</td>
<td>3.11% (57)</td>
<td>2.46%(62)</td>
<td>1.34% (72)</td>
<td>1.66% (353)</td>
</tr>
</tbody>
</table>

Table 3: Summary of the effectiveness of the heuristic and the myopic solutions for the four-stage system.

<table>
<thead>
<tr>
<th>Solution</th>
<th>stage 2</th>
<th>stage 3</th>
<th>stage 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s^H_{ij}(t)$</td>
<td>1.50%</td>
<td>1.57%</td>
<td>2.03%</td>
<td>1.70%</td>
</tr>
<tr>
<td>$s^M_{ij}(t)$</td>
<td>1.57%</td>
<td>1.68%</td>
<td>2.28%</td>
<td>1.84%</td>
</tr>
</tbody>
</table>
5. Implementation Issues

So far, we have focused the discussion on the centralized control system, i.e., the supply chain belongs to a single firm. We have demonstrated how our heuristic can help the owner reduce the computational complexity and duration of computation by allocating the computation to each stage.

Now, let us consider a decentralized control system, i.e., the supply chain is composed of independent firms, each pursuing its own best interest. There are two questions of interest. First, is there an equilibrium solution between these individual firms in the finite horizon? If so, by how much the equilibrium solution does deviate from the centralized (first-best) solution? To answering this question, one has to analyze a dynamic game; in general, it is difficult to show the existence of a Markov equilibrium solution (Parker and Kapuscinski 2010). Second, how does one design a contract (or mechanism) that can coordinate the firms to achieve the centralized solution? Because the centralized solution is often time-varying, a coordination contract, if it existed, would have non-stationary parameters. Thus, it is more difficult to implement the contract.

Here, we focus only on the second question. As we shall demonstrate below, our heuristic leads to a simple, time-consistent contract that can induce the supply chain partners to choose the heuristic solution \( s_a^j(t) \) in each period.

To demonstrate the coordination scheme, let us consider a two-stage system, where stage \( j \) has full information and is responsible for determining its echelon base-stock level. (This is the echelon information scenario described in, e.g., Cachon and Zipkin (1999), Shang et al. (2009) and Parker and Kapuscinski (2010).) We assume that stage \( j \) chooses an arbitrary base-stock level \( s_j^0(t) \) in period \( t, t = 1, ..., T \). Denote the resulting supply chain cost incurred in the \( T \)-period horizon as \( f_T^0(x_1, x_2) \), where \( x_j \) is the echelon inventory level for stage \( j \) at the beginning of period \( T \). Clearly, \( f_T^0(x_1, x_2) \) is the sum of the cost incurred at stage 1, \( c_T^0(x_1) \), and at stage 2, \( g_T^0(x_2) \). (Here, stage 1 and stage 2 are cost centers. In practice, there can be different ways to calculate the cost for the cost center. For example, a natural way is to charge \( p_j \) per unit of inventory ordered, \( h_j \) per unit of the echelon on-hand inventory for stage \( j \), \( j = 1, 2 \), and \( b \) per unit of the backorders for stage 1 in each period. In any case, the bottom line is \( f_T^0(x_1, x_2) = c_T^0(x_1) + g_T^0(x_2). \)

We assume that there is an integrator who knows the centralized solution and is responsible for payment transfers between the firms. The integrator can be one of these firms, a team of them, or a third-party organization. The integrator designs a contract for each stage \( j \) with three cost
parameters \((\theta_j p_j^a, \theta_j h_j^a, \theta_j b_j)\), aiming to induce the stage manager to choose \(s_j^a(t)\) in each period, \(t = 1, 2, \ldots, T\), where \(\theta_j\) is an adjustment factor for stage \(j\), \(0 \leq \theta_j \leq 1\). Let the resulting near-optimal cost be \(f_T^a(x_1, x_2)\) and assume \(f_T^a(x_1, x_2) < f_T^0(x_1, x_2)\).

The contract is implemented as follows: in each period, the integrator first compensates the actual cost incurred for stage \(j\). Then, stage \(j\) pays the integrator based on the accounting echelon inventory level \(\bar{x}_j\) according to the cost terms. (The accounting echelon inventory level is defined as the echelon inventory level by assuming that there is an ample supply from upstream.) More specifically, with the perceived lead time \(T_j\) periods, stage \(j\) pays the integrator \(\theta_j p_j^a\) per unit of \((\bar{x}_j)^+\) and \(\theta_j b_j\) per unit of \((\bar{x}_j)^-\) in each period, where \((x)^+ = \max\{x, 0\}\) and \((x)^- = \max\{0, -x\}\). From stage \(j\)’s perspective, because his actual cost will be covered by the integrator, he only needs to minimize the payment to the integrator in the \(T\) period. Thus, the problem for stage \(j\) is exactly the same as solving the heuristic system \(j\), and stage \(j\) will therefore choose \(s_j^a(t)\) as the optimal solution. (\(\theta_j\) is a constant, which will not affect the optimal solution.)

To see why such a contract is implementable, we have to demonstrate that each player, including the integrator, is better off in this game. Let us denote stage 1’s payment (respectively, stage 2’s payment) to the integrator as \(\theta_1 c_T^1(\bar{x}_1)\) (respectively, \(\theta_2 g_T^1(\bar{x}_2)\)). Clearly, the stages are willing to accept the contract if the outflow payment is smaller than the inflow payment, i.e.,

\[
\theta_1 c_T^1(\bar{x}_1) < c_T^0(x_1), \quad \theta_2 g_T^1(\bar{x}_2) < g_T^0(x_2).
\]  

Similarly, the integrator is willing to perform this function if she can make profits after both stages implement the solution \(s_j^a(t)\), i.e.,

\[
\theta_1 c_T^1(\bar{x}_1) + \theta_2 g_T^1(\bar{x}_2) > f_T^0(x_1, x_2).
\]  

Because \(c_T^0(x_1) + g_T^0(x_2) = f_T^0(x_1, x_2) > f_T^0(x_1, x_2)\), there exist \(\theta_1\) and \(\theta_2\) such that Equations (3) and (4) hold. Thus, the contract is implementable.

In fact, the resulting solution \((s_1^a(t), s_2^a(t))\) is a Markov equilibrium in this dynamic game. (This means that each player will choose \(s_j^a(t)\) in each period without deviation.) This can be verified by the fact that each firm optimizes its inventory decision independently. Thus, the solution \((s_1^a(t), s_2^a(t))\) is a Nash equilibrium in each period.

Although the suggested coordination scheme can only achieve a near-optimal solution, it has the advantage of easy implementation because the contract terms are stationary. A key enabler of the coordination scheme is the integrator, who decouples the players’ decisions in each period.
In practice, the integrator can be one of the firms, a subset of the firms, or a third-party financial institution. We refer the reader to Shang et al. (2009) for examples of the integrator.

6. Concluding Remarks

This paper presents an approach to generating bounds for the optimal echelon base-stock levels for a serial inventory system with a finite horizon. The demands between periods are independent but are allowed to be nonidentical. The solution bounds are obtained by solving two revised problems. The revised problem that generates the lower bound solution is formulated by regulating stages' order decisions; the revised problem that generates the upper bound solution is constructed by regulating stages’ holding cost and order cost parameters. We prove that solving these revised problems is equivalent to solving single-stage models with the original problem data. We suggest a heuristic by solving a series of single-stage systems with a weighted average of cost parameters. A numerical study suggests that the heuristic is near-optimal. We therefore derive the myopic solution for the local base-stock level in the heuristic system to reveal insights concerning the local inventories and safety stocks. We also demonstrate how our heuristic can lead to a coordination scheme that achieves the near-optimal solution.

The solution bounds and the heuristic can be extended to a one-warehouse-multi-retailer distribution system in which the retailers are identical. Under the so-called balance assumption (i.e., the inventory levels of the retailers can be freely and instantaneously re-distributed as needed), the distribution system is equivalent to a two-stage serial system, where the downstream stage can be viewed as a composite stage that includes all retailers’ demands. As shown in Federgruen and Zipkin (1984), one can apply the Clark and Scarf serial algorithm to obtain the echelon base-stock level for the warehouse and the composite stage. Then, one could apply the myopic allocation rule in each period to the retailers to determine the retailers' base-stock levels. Federgruen and Zipkin reports that this approach can generate a very effective solution. Clearly, we can apply the same technique to generate single-stage approximations for the resulting two-stage system, and then apply the myopic allocation for the system. It will be interesting to examine whether our results can be applied to the non-identical retailer system. We leave this for the future research.

Finally, our solution bounds can be extended to a system with Markov-modulated demand. More specifically, assume that the demand process is driven by a homogeneous, discrete-time Markov chain $\mathbf{W}$ with $K$ states. It is known that a state-dependent echelon base-stock policy
is optimal (e.g., Chen and Song 2000, Muharremoglu and Tsitsiklis 2008). Let $s_j(k, t)$ be the optimal echelon base-stock level for stage $j$ when the demand state is $k$ in period $t$, $k = 1, ..., K$. Following a similar analysis, we can derive single-stage bounds $s^\ell_j(k, t)$ and $s^u_j(k, t)$ for each demand state $k$ in each period $t$ such that $s^\ell_j(k, t) \leq s_j(k, t) \leq s^u_j(k, t)$. The proof is available from the author.

Acknowledgements: The author would like to thank the editor Stephen C. Graves and anonymous reviewers for useful suggestions, and Paul Rebeiz for the computation.

References


Online Companion for
Single-Stage Bounds for Optimal Policies in Serial Inventory
Systems with Non-stationary Demand

Appendix A: Optimal Value Functions

To facilitate the analysis of the proofs in Appendix B, we provide dynamic program formulations
for the Clark-Scarf model, the upper-bound system, and the lower-bound system. Note that our
formulation for the Clark-Scarf model is different from that in Clark and Scarf (1960), who decouple
the total cost by introducing an induced-penalty cost function. With the penalty cost function,
Clark and Scarf decouple the total cost into each stage. We, on the other hand, decouple the total
cost into each echelon. This alternative decomposition scheme is necessary for our analysis because
the functional equation for each stage in the the upper- and the lower-bound systems turns out to
be the echelon cost. Also, to simplify the notation and the analysis, we only focus on the two-stage
system with $\tau_1 = \tau_2 = 1$. A similar but more tedious analysis can be carried out for the general
system.

The Clark and Scarf Model

Let $x_j$ be the echelon inventory level at stage $j$ after a shipment is received, and let $y_j$ be the echelon
inventory in-transit position at stage $j$ after an order is placed. Define $L_j(x_j, t)$ the inventory cost
incurred for stage $j$ in period $t$ when the echelon inventory level is $x_j$, namely,

$$
L_1(x_1, t) = \mathbb{E}[h_1(x_1 - D(t)) + (b + h_1 + h_2)(x_1 - D(t))^+] - 
$$

$$
L_2(x_2, t) = \mathbb{E}[h_2(x_2 - D(t))],
$$

where $(x)^+ = \max\{0, -x\}$. The total inventory holding and backorder cost incurred in period $t$ is
then $L_1(x_1, t) + L_2(x_2, t)$. Let $f_t(x_1, x_2)$ be the optimal total discounted cost for the system with
initial echelon inventory levels $(x_1, x_2)$ when there are $t$ periods to go. An alternative dynamic
program based on the echelon system for the Clark-Scarf model is as follows: Let $f_0(x_1, x_2) = 0$.
For $t \geq 1$,

$$
f_t(x_1, x_2) = \min_{x_1 \leq y_1 \leq x_2 \leq y_2} \left\{ p_1(y_1 - x_1) + p_2(y_2 - x_2) + L_1(x_1, t) + L_2(x_2, t) + \alpha \mathbb{E}[f_{t-1}(y_1 - D(t), y_2 - D(t))] \right\}
$$
\[ = C_t(x_1) + G_t(x_2), \]

where

\[
C_t(x_1) = L_1(x_1, t) - p_1 x_1 + [U_t(\max\{x_1, s_1(t)\}) - U_t(s_1(t))],
\]

(5)

\[
G_t(x_2) = U_t(\min\{x_2, s_1(t)\}) + L_2(x_2, t) - p_2 x_2 + V_t(\max\{x_2, s_2(t)\}),
\]

(6)

\[
U_t(y_1) = p_1 y_1 + \alpha E[C_{t-1}(y_1 - D(t))],
\]

\[
V_t(y_2) = p_2 y_2 + \alpha E[G_{t-1}(y_2 - D(t))],
\]

\[
s_1(t) = \arg\min_{y_1} U_t(y_1),
\]

\[
s_2(t) = \arg\min_{y_2} V_t(y_2).
\]

Note that Equations (5) and (6) should be revised for \( t = 1 \) and \( t = 2 \). When \( t = 1 \), both stages will not order, so \( C_1(x_1) = L_1(x_1, 1) \) and \( G_1(x_2) = L_2(x_2, 1) \); when \( t = 2 \), stage 2 will not order, so \( G_2(x_2) \) is the same as (6) except that the last term is changed to \( V_2(x_2) \).

Under the above cost decomposition scheme, the total system cost \( f_t(x_1, x_2) \) consists of two echelon cost functions. \( G_t(x_2) \) is the cost for echelon 2, which includes the costs directly and indirectly determined by \( x_2 \), assuming that stage 1 will always order up to its optimal base-stock level in each period. \( C_t(x_1) \) includes the remaining costs directly determined by \( x_1 \).

**The Upper Bound System**

We first define the following single-period inventory holding and backorder cost function:

\[
L(x_1, t) = E[(h_1 + h_2)(x_1 - D(t)) + (b + h_1 + h_2)(x_1 - D(t))^2].
\]

Let \( \overline{f}_t(x_1, x_2) \) denote the optimal total cost when the state is \( (x_1, x_2) \) at the beginning of period \( t \), and \( \overline{f}_0(x_1, x_2) = 0 \). For any \( t \geq 1 \),

\[
\overline{f}_t(x_1, x_2) = \min_{x_2 \leq y_2} \left\{ p_1(x_2 - x_1) + p_2(y_2 - x_2) + L(x_1, t) + L_2(x_2, t)
\right.
\]

\[
+ \alpha E[\overline{f}_{t-1}(x_2 - D(t), y_2 - D(t))]
\]

\[
= \overline{C}_t(x_1) + \overline{G}_t(x_2),
\]

where

\[
\overline{C}_t(x_1) = L_1(x_1, t) - p_1 x_1,
\]

\[
\overline{G}_t(x_2) = U_t(\min\{x_2, s_1(t)\}) + L_2(x_2, t) - p_2 x_2 + V_t(\max\{x_2, s_2(t)\}).
\]
\[
\begin{align*}
\mathcal{G}_t(x_2) &= \mathcal{U}_t(x_2) + L_2(x_2, t) - p_2 x_2 + \mathcal{V}_t(\max\{s_2^t(t), x_2\}), \\
\mathcal{U}_t(x_2) &= p_1 x_2 + \alpha \mathbb{E}[\mathcal{C}_{t-1}(x_2 - D(t))], \\
\mathcal{V}_t(y_2) &= p_2 y_2 + \alpha \mathbb{E}[\mathcal{G}_{t-1}(y_2 - D(t))], \\
s_2^t(t) &= \arg\min_{y_2} \{\mathcal{V}_t(y_2)\}.
\end{align*}
\]

(We have to revise the above \(\mathcal{G}_t(x_1)\) and \(\mathcal{G}_t(x_2)\) when \(t = 1\) and \(t = 2\) in a similar fashion as in the above Clark and Scarf formulation.)

**The Lower Bound System**

Let \(f^*_t(x_1, x_2)\) denote the optimal discounted total cost with initial echelon inventory levels \((x_1, x_2)\) when there are \(t\) periods before termination, and \(f^*_0(x_1, x_2) = 0\). We define the one-period inventory cost as follows:

\[
\begin{align*}
L_1(x_1, t) &= \mathbb{E}[(b + h_2)(x_1 - D(t))] , \\
L_2(x_2, t) &= \mathbb{E}[h_2(x_2 - D(t))].
\end{align*}
\]

For any \(t \geq 1\),

\[
\begin{align*}
f^*_t(x_1, x_2) &= \min_{x_1 \leq y_1 \leq x_2 \leq y_2} \left\{ p_2(y_2 - x_2) + L_1(x_1, t) + L_2(x_2, t) \\
&\quad \quad + \alpha \mathbb{E}[f^*_{t-1}(y_1 - D(t), y_2 - D(t))] \right\} \\
&= \mathcal{C}_t(x_1) + \mathcal{G}_t(x_2),
\end{align*}
\]

where

\[
\begin{align*}
\mathcal{C}_t(x_1) &= L_1(x_1, t), \\
\mathcal{G}_t(x_2) &= \mathcal{U}_t(x_2) + L_2(x_2, t) - p_2 x_2 + \mathcal{V}_t(\max\{x_2, s_2(t)\}), \\
\mathcal{U}_t(x_1) &= \alpha \mathbb{E}[\mathcal{C}_{t-1}(x_2 - D(t))], \\
\mathcal{V}_t(y_2) &= p_2 y_2 + \alpha \mathbb{E}[\mathcal{G}_{t-1}(y_2 - D(t))], \\
s_2(t) &= \arg\min_{y_2} \{\mathcal{V}_t(y_2)\}.
\end{align*}
\]

(We have to revise the above \(\mathcal{C}_t(x_1)\) and \(\mathcal{G}_t(x_2)\) when \(t = 1\) and \(t = 2\) in a similar fashion as in the above Clark and Scarf formulation.)
Appendix B: Proofs

Theorem 1

We use the notation “Δ” to represent the difference of a function, i.e.,

$$\Delta f(x) = f(x + 1) - f(x).$$

Part (1) can be shown by a simple induction.

To show part (2), we need to show $\Delta V_t(y_2) \geq \Delta V_t(y_2)$ for all $y_2$ when $x_1 < s_1(t)$. We prove this result by induction. Notice that under the assumption of $x_1 < s_1(t)$, $C_t(x_1) = L(x_1, t) - p_1 x_1$, which is equal to $\overline{C}_t(x_1)$. Thus, in each period $t$, $U_t(\cdot) = \overline{U}_t(\cdot)$.

We start the induction from $t = 3$. When $t = 3$,

$$\Delta V_3(y_2) = p_2 + \alpha E[G_2(y_2 + 1 - D(3)) - G_2(y_2 - D(3))]$$

$$= p_2 + \alpha E[U_2(\min\{y_2 + 1 - D(3), s_1(2)\}) - U_2(\min\{y_2 - D(3), s_1(2)\}) + (h_2 - p_2)$$

$$+ \Delta V_2(y_2 - D(3))]$$

$$\leq p_2 + \alpha E[U_2(y_2 + 1 - D(3)) - U_2(y_2 - D(3)) + (h_2 - p_2) + \Delta V_2(y_2 - D(3))]$$

$$= \Delta V_3(y_2).$$

Thus, $s_2(3) \geq s_2^U(3)$. (In fact, the condition is not required for $t = 3$.)

Suppose $t - 1$ holds true, i.e., $\Delta V_{t-1}(y_2) \leq \Delta \overline{V}_{t-1}(y_2)$ and $s_2(t - 1) \geq s_2^U(t - 1)$. For period $t$,

we have

$$\Delta V_t(y_2) = p_2 + \alpha E[G_{t-1}(y_2 + 1 - D(t)) - G_{t-1}(y_2 - D(t))]$$

$$= p_2 + \alpha E[U_{t-1}(\min\{y_2 + 1 - D(t), s_1(t - 1)\}) + V_{t-1}(\max\{y_2 + 1 - D(t), s_2(t - 1)\})$$

$$- U_{t-1}(\min\{y_2 - D(t), s_1(t - 1)\}) - V_{t-1}(\max\{y_2 - D(t), s_2(t - 1)\}) + (h_2 - p_2)$$

$$\leq p_2 + \alpha E[U_{t-1}(y_2 + 1 - D(t)) - U_{t-1}(y_2 - D(t)) + (h_2 - p_2) + \Delta V_{t-1}(\max\{y_2 - D(t), s_2(t - 1)\})$$

$$= p_2 + \alpha E[U_{t-1}(y_2 + 1 - D(t)) - U_{t-1}(y_2 - D(t)) + (h_2 - p_2) + \Delta V_{t-1}(\max\{y_2 - D(t), s_2^U(t - 1)\})$$

$$\leq p_2 + \alpha E[\overline{U}_{t-1}(y_2 + 1 - D(t)) - \overline{U}_{t-1}(y_2 - D(t)) + (h_2 - p_2) + \Delta \overline{V}_{t-1}(\max\{y_2 - D(t), s_2^U(t - 1)\})$$

$$= p_2 + \alpha E[\overline{U}_{t-1}(y_2 + 1 - D(t)) - \overline{U}_{t-1}(y_2 - D(t))]$$

$$= \Delta \overline{V}_t(y_2).$$

The last inequality is due to $\Delta V_{t-1}(y_2) \leq \Delta \overline{V}_{t-1}(y_2)$ and the third equality is due to $s_2^U(t - 1) \leq s_2(t - 1)$. 

20
Proposition 2

The single-period inventory holding and backorder cost for the two-stage system can be expressed as

\[ L_1(x_1, t) + L_2(x_2, t) = L(x_1, t) + h_2(x_2 - x_1). \]

One can view \( L(x_1, t) \) as the local inventory holding and backorder cost for stage 1 and \( h_2(x_2 - x_1) \) as the inventory holding cost incurred by the inventory held at stage 2 and in transit to stage 1.

It is more clear to prove this result by redefining the state variables. Let \( m = x_2 - x_1 \) and \( z = y_2 - x_2 \). With slight abuse of notation, the optimality equation for the two-stage upper-bound system is

\[
\overline{f}_t(x_1, m) = \min_{z \geq 0} \left\{ p_2z + (p_1 + h_2)m + L(x_1, t) + \alpha E[\overline{f}_{t-1}(x_1 + m - D(t), z)] \right\},
\]

where \( \overline{f}_0(x_1, m) = 0, \overline{f}_1(x_1, m) = L(x_1, 1) + h_2m, \) and

\[
\overline{f}_2(x_1, m) = \min_{z \geq 0} \left\{ p_2z + (p_1 + h_2)m + L(x_1, 2) + \alpha E[\overline{f}_1(x_1 + m - D(2), z)] \right\}
= (p_1 + h_2)m + L(x_1, 2) + \alpha E[L(x_1 + m - D(2), 1)] + \min_{z \geq 0} \{(p_2 + \alpha h_2)z\}
= (p_1 + h_2)m + L(x_1, 2) + \alpha E[L(x_1 + m - D(2), 1)].
\]

For \( t \geq 3 \), we have

\[
\overline{f}_t(x_1, m) = (p_1 + h_2)m + L(x_1, t) + \alpha E[L(x_1 + m - D(t), t - 1)] + \overline{G}_t(x_1 + m),
\]

where

\[
\overline{G}_t(x_1 + m) = \min_{z \geq 0} \left\{ [p_2 + \alpha(p_1 + h_2)]z + \alpha^2 E[L(x_1 + m + z - D(t) - D(t - 1), t - 2)] + \alpha E[\overline{G}_{t-1}(x_1 + m + z - D(t))] \right\}.
\]

This dynamic program has the same structure as that of the single-stage problem (Karlin and Scarf 1958) with lead time of two periods and with the specified cost parameters in the proposition.

Theorem 3

Part (1) is straightforward and omitted.

We show Part (2) by induction. To show \( s_2^n(t) \geq s_2(t) \), we need to show \( \Delta V_t(y_2) \geq \Delta V_1(y_2) \) for all \( y_2 \). Denote \((x)^+ = \max\{x, 0\}\).
Let us first consider $t = 3$.

\[
\Delta V^3(y_2) = p_2 + \alpha E[G_2(y_2 + 1 - D(3)) - G_2(y_2 - D(3))]
\]

\[
= p_2 + \alpha E\left[U_2(\min\{y_2 + 1 - D(3), s_1(2)\}) - U_2(\min\{y_2 - D(3), s_1(2)\})\right]
\]

\[
+ (h_2 - p_2) + \Delta V_2(y_2 - D(3))
\]

\[
= p_2 + \alpha E\left[p_1(\min\{y_2 + 1 - D(3), s_1(2)\}) - p_1(\min\{y_2 - D(3), s_1(2)\})\right]
\]

\[
+ \alpha E[C_1(\min\{y_2 + 1 - D(3), s_1(2)\} - D(2))] - \alpha E[C_1(\min\{y_2 - D(3), s_1(2)\} - D(2))]
\]

\[
+ (h_2 - p_2) + \Delta V_2(y_2 - D(3))
\]

\[
\geq p_2 + \alpha E\left[\alpha E\left[h_1(\min\{y_2 + 1 - D(3), s_1(2)\} - D(2) - D(1))^+\right]
\]

\[
+ (b + h_2)(\min\{y_2 + 1 - D(3), s_1(2)\} - D(2) - D(1))^-
\]

\[
- \alpha E\left[h_1(\min\{y_2 - D(3), s_1(2)\} - D(2) - D(1))^+\right]
\]

\[
+ (b + h_2)(\min\{y_2 - D(3), s_1(2)\} - D(2) - D(1))^-
\]

\[
+ (h_2 - p_2) + \Delta V_2(y_2 - D(3))
\]

\[
\geq p_2 + \alpha E\left[\alpha E\left[\alpha E\left[L_1(y_2 + 1 - D(3) - D(2) - D(1))^+\right]
\]

\[
- (b + h_2)(\min\{y_2 - D(3), s_1(2)\} - D(2) - D(1))^-
\]

\[
+ (h_2 - p_2) + \Delta V_2(y_2 - D(3))
\]

\[
\geq p_2 + \alpha E\left[\alpha E[L_1(y_2 + 1 - D(3) - D(2), 1)] - \alpha E[L_1(y_2 - D(3) - D(2), 1)]
\]

\[
+ (h_2 - p_2) + \Delta V_2(y_2 - D(3))\right]\quad (\text{because } V_2(\cdot) = V_2(\cdot))
\]
Thus, we have
\[
2 \geq t \geq t + (h_2 - p_2) + \Delta V_2(y_2 - D(3))
\]

\[
= \Delta V_3(y_2).
\]

Thus, we have \( s_2(y_2) \geq s_2(3) \).

Suppose \( t - 1 \) holds true, that is, \( \Delta V_{t-1}(y_2) \geq \Delta V_{t-1}(y_2) \) for all \( y_2 \), and \( s_2(t - 1) \geq s_2(t - 1) \). Then, for period \( t \),

\[
\Delta V_t(y_2) = p_2 + \alpha E \left[ G_{t-1}(y_2 + 1 - D(t)) - G_{t-1}(y_2 - D(t)) \right]
\]

\[
= p_2 + \alpha E \left[ U_{t-1}(\min\{y_2 + 1 - D(t), s_1(t - 1)\}) - U_{t-1}(\min\{y_2 - D(t), s_1(t - 1)\}) \right.
\]

\[
+ (h_2 - p_2) + \Delta V_{t-1}(\max\{y_2 - D(t), s_2(t - 1)\})
\]

\[
= p_2 + \alpha E \left[ p_1(\min\{y_2 + 1 - D(t), s_1(t - 1)\}) - p_1(\min\{y_2 - D(t), s_1(t - 1)\}) \right.
\]

\[
+ \alpha E[C_{t-2}(\min\{y_2 + 1 - D(t), s_1(t - 1)\} - D(t - 1)])
\]

\[
- \alpha E[C_{t-2}(\min\{y_2 - D(t), s_1(t - 1)\} - D(t - 1))]
\]

\[
+ (h_2 - p_2) + \Delta V_{t-1}(\max\{y_2 - D(t), s_2(t - 1)\})
\]

\[
\geq p_2 + \alpha E \left[ \alpha E \left[ L_1(\min\{y_2 + 1 - D(t), s_1(t - 1)\} - D(t - 1), t - 2) \right.ight.
\]

\[
- L_1(\min\{y_2 - D(t), s_1(t - 1)\} - D(t - 1), t - 2)
\]

\[
+ (h_2 - p_2) + \Delta V_{t-1}(\max\{y_2 - D(t), s_2(t - 1)\})
\]

\[
\geq p_2 + \alpha E \left[ \alpha E \left[ L_1(\min\{y_2 + 1 - D(t), s_1(t - 1)\} - D(t - 1), t - 2) \right.ight.
\]

\[
- L_1(\min\{y_2 - D(t), s_1(t - 1)\} - D(t - 1), t - 2)
\]

\[
+ (h_2 - p_2) + \Delta V_{t-1}(\max\{y_2 - D(t), s_2(t - 1)\})
\]

\[
= \Delta V_t(y_2).
\]

Thus, we have \( s_2(y_2) \geq s_2(t) \).
Propositions 4

The proof is similar to that of proposition 2, and thus omitted.

Proposition 5

The dynamic program for a single-stage system with lead time $L$ periods, order cost $p$, holding cost rate $h$, and backorder cost rate $b$ is as follows:

$$f_t(x) = \min_{y \geq x} \left\{ p(y - x) + \alpha^L \mathbb{E}[L(y - D[t, t - L + 1], t - L)] + \alpha \mathbb{E}[f_{t-1}(y - D(t))] \right\}$$

$$= \min_{y \geq x} \{ u_t(y) \} - px.$$  

where $f_0(x) = 0$. Here, $x$ and $y$ are the inventory order position before and after ordering at the beginning of period $t$, respectively. $L(x, t)$ is the one-period inventory holding and backorder cost in period $t$ when the initial inventory order position is $x$. The optimal base-stock level $s(t)$ can be found by minimizing $u_t(y)$ in each period: $s(t) = \arg \min_y \{ u_t(y) \}$, and the resulting optimal value function is $f_t(x) = u_t(\max\{x, s(t)\}) - px$.

Let $s^m(t)$ denote the myopic solution in period $t$ for the above inventory problem. For $t = L + 1$,

$$s^m(L + 1) = \arg \min_y \{ py + \alpha^L \mathbb{E}[L(y - D[L + 1, 1])] \}.$$  

Thus, $s^m(L + 1)$ is the smallest $y$ such that $P(D[L + 1, 1]) \leq y \geq (\alpha^L b - p)/\alpha^L (b + h)$.

For $t > L + 1$,

$$f_t(x) = \min_{y \geq x} \left\{ p(y - x) + \alpha^L \mathbb{E}[L(y - D[t, t - L + 1], t - L)] + \alpha \mathbb{E}[u_{t-1}(\max\{y - D(t), s(t - 1)\}) - p(y - D(t))] \right\}.$$  

Thus, $s^m(t) = \arg \min_y \{ p(1 - \alpha)y + \alpha^L \mathbb{E}[L(y - D[t, t - L + 1], t - L)] \}$, or equivalently, $s^m(t)$ is the smallest $y$ such that $P(D[L + 1, 1]) \leq y \geq (\alpha^L b - p(1 - \alpha))/\alpha^L (b + h)$. 

24
Supply Streams

Jing-Sheng Song  Paul Zipkin

The Fuqua School of Business, Duke University, Durham, NC 27708
jssong@duke.edu  zipkin@duke.edu
Presenter: Paul Zipkin
February 20, 2011
Supply Streams

February 9, 2011

Abstract

A supply stream is a continuous version of a supply chain. It is like a series inventory system, but stock can be held at any point along a continuum, not just at discrete stages. We assume stationary parameters. For a discrete-stage model, a stationary echelon base-stock policy is known to be optimal. We show that the solutions to discrete-stage systems converge monotonically to a limit, as the distances between the stages become small, and this limit solves the continuous-stage system. We examine two such models in greater detail, one with Poisson demand and the other with demand approximated by a Brownian motion. For Poisson demand we develop tractable computational methods. For Brownian motion we show that, in the continuous-stage limit, the model is equivalent to one describing first-passage times. This linkage leads to some interesting and useful results. Specifically, we obtain a partial differential equation that characterizes the solution, and we find closed-form solutions for certain special cases, the first in the inventory literature.

Keywords: Inventory/production: multiechelon, uncertainty; probability: diffusion; partial differential equations.
1 Introduction

A supply stream is a continuous version of a supply chain. It is like a series inventory system, but stock can be held at any point along a continuum, not just at discrete stages. Time too is continuous.

Imagine that a single product is carried on barges on some waterway. Demand and supply occur in whole barge-loads. Demand occurs, randomly, only at one end of the waterway, and the source is at the other end. At any point in time, one can launch a barge at the source. The barge usually travels at a certain maximum speed towards the destination. At any point along the way, however, one can slow the barge down or even stop it. Later, the barge can start moving again. Once the barge reaches the destination, its goods are stored in inventory until they are demanded.

Why stop a barge in midstream or even slow it down? The answer is, to save on holding costs. We assume a cost of holding inventory on the barge that increases as it moves towards the destination. This relation is sometimes called a cost-time profile; see Fooks [21], Schraner and Hausman [39] and Chaudhari [5]. On the other hand, of course, as the barge moves, the closer it gets to being available to meet demand and thus avoid shortage costs. Similar stories can be told about other modes of transportation.

As a model of traffic, this one is stylized in many ways, but a few deserve mention at the outset. First, those stops and starts and changes in speed are costless. Imagine that the crew can just cut the engines, drop anchor, and turn to unpaid leisure activities. Second, there are no capacity limits and therefore no congestion. Each barge has a maximum speed, but any number of barges can pass a given point at any time. Third, the speed of each barge is unaffected by exogenous influences, such as weather, or internal impediments, such as breakdowns. Fourth, there are no explicit scale economies, like fixed costs. Here, a barge-load is the system’s basic quantity unit.

A supply stream can also describe a fluid that flows from a source to a destination. Think of a pipeline or a channel moving gasoline, natural gas, or water. The fluid moves only in one direction, with a maximum velocity, but otherwise we can costlessly control the flow rate at each point. (Again, as a description of fluid dynamics, this model omits some important real factors. Our pipes have infinite capacities and therefore zero pressures.)

Or, imagine a production process with many stages. We approximate it by a continuum. A unit of product can move through the system at a maximum rate, but we can stop it anywhere and hold it in inventory or slow it down. Many production processes really are continuous in this sense. Oil refining and chemical processing are examples. (Our assumption that product can be stored at any point is a stretch in those contexts, however. Also, they too have finite capacities.)

The model is a natural continuous limit of discrete-stage models. Such discrete-stage models are
well understood. Their analysis was pioneered by Clark and Scarf [9] and refined by Federgruen and Zipkin [19] and Chen and Zheng [6]. Concepts and methods from this literature, such as echelon inventory, play prominent roles in our discussion.

We formulate the continuous-stage model in Section 2. Demand is a compound-Poisson process, and all parameters are stationary. There are costs for holding inventory, which vary continuously over space, and a penalty cost for backorders at the point of customer demand. The objective is the long-run average total cost. We construct a discrete-stage model by allowing the system to retain inventory only at fixed, discrete points. We know that a policy of simple form, called a stationary echelon base-stock policy, is optimal for this model, and we have a simple algorithm to compute one. We then recover analogous results for the continuous model by taking limits, as the distances between the stocking points become small. We demonstrate that the optimal costs and optimal policies of the discrete-stage models converge monotonically, and these limits solve the original continuous-stage model. The result also provides simple bounds on the optimal policy and cost.

The next two sections study two particular demand models more closely. For Poisson demand (Section 3), we show how to evaluate any given policy and to find the best one with tractable algorithms.

In Section 4 we approximate demand by a Brownian motion. (More precisely, we approximate the optimal cost in equilibrium by treating demand as a Brownian motion.) We show that, in the limit, the model is equivalent to one that describes the first-passage time of standard Brownian motion to a moving boundary. (A similar equivalence holds for certain discrete-stage models, as pointed out in Section 2.) This linkage is interesting and fruitful. That model has been studied intensively, and the results immediately apply to the the supply-stream model. In particular, we obtain a partial differential equation that characterizes the solution, as well as closed-form solutions for certain special cases. These are the only closed-form solutions of inventory-network models we know of. We point out and explain some interesting qualitative features of these solutions. Their behavior is different from what we might expect by extrapolation from single-stage models. We also discuss various numerical solution methods.

Continuous-time, continuous-state models have been extensively used to study queuing network models, see, for example, Harrison [25] and Chen and Yao [7]. Their applications to inventory models have been relatively scant. Examples include Bather [2], Reiman et al. [37] and Plambeck and Ward [34], [35]. We hope this paper can help generate more interest in this domain.
2 Basics

2.1 Formulation

We use the continuous variable $t \geq 0$ to indicate time. In space, the system covers a finite, semi-open interval, denoted $[0, U)$. Customer demand occurs at point $u = 0$, and point $u = U$ represents an outside supplier with ample stock. When we order a unit from the supplier, it starts at $U$ and travels towards 0. We scale space so that the time it would take a unit traveling at maximum speed to reach point 0 from $u \in [0, U)$ is precisely $u$. Once a unit enters the system at $U$, we can choose to hold it at any point $u$ along the way or to make it move more slowly than the maximum speed. Echelon $u$ means the subsystem comprising point $u$ and all points downstream, that is, the interval $[0, u]$, $0 \leq u < U$. When a demand occurs, if there is enough stock at point 0, the demand is satisfied. Otherwise, the demand is satisfied as much as possible, and the unsatisfied portion is backlogged, to be filled later.

Demand is represented by a compound-Poisson process $D = \{D(t) : t \geq 0\}$ with positive jumps. Here, $D(0) = 0$, and $D(t)$ is the cumulative demand in $(0, t]$, $t > 0$. Also, let $D(s, t) = D(t) - D(s)$, $s \leq t$. Thus, the sample paths of $D$ are nondecreasing and càdlàg (right-continuous with left limits, or RCLL). Also, $D$ has stationary, independent increments. The underlying Poisson process has rate $\lambda$, and the jumps are distributed as a positive random variable $X$. The average demand rate is $\mu$, with $0 < \mu \equiv E[D(1)] = \lambda E[X] < \infty$. Let $\Psi_0(t, x) = \Pr\{D(t) > x\}$ denote the complementary cumulative distribution function (cdf) of $D(t)$, and $\Psi_1(t, x) = E[\lceil x - D(t) \rceil]$ its loss function. For the range of $D$, we consider two cases, integer demand ($X$ is a positive integer with $\Pr\{X = 1\} > 0$) or continuous demand ($X$ has a positive density on the interval $(0, X_+)$ for some $X_+ > 0$). The results apply to both cases unless indicated otherwise.

The control is described by the supply process $S = \{S(t, u) : t \geq 0, u \in [0, U)\}$, where $S(t, u)$ is the cumulative supply to echelon $u \in [0, U)$ in time $(0, t]$, $t > 0$. The set of admissible controls, denoted $\mathcal{S}$, are those satisfying the following conditions:

- $S$ is adapted to $D$ (in particular, nonanticipating).
- The sample paths of $S$ are nondecreasing and càdlàg in both $t$ and $u$.
- $S$ satisfies

$$S(s, u) \leq S(t, u + s - t) , \; t < s < t + U - u.$$  \hspace{1cm} (1)

This constraint expresses the condition that goods can move towards 0 at rate no more than 1. When $D$ is integer valued, so is $S$, because only whole units can be used to meet demand; when $D$ is continuous, $S$ is real valued. The initial supply $S(0, u)$ is a given constant. It is nonnegative, nondecreasing and càdlàg in $u$. 

4
These properties are consistent with our usual notions of orders and shipments. Suppose \( S(0, u) = 0 \), and we order one unit at time 0. We subsequently move the unit so that its position describes a continuous, nonincreasing curve \( \bar{u}(t) \), where \( \bar{u}(0) = U, \bar{u}(t) < U, t > 0 \), and, according to (1), \( \bar{u}(s) - \bar{u}(t) \geq -(s - t) \) for \( s > t \). Then, for \( t > 0 \), \( S(t, u) = 1 \) for \( u \geq \bar{u}(t) \) and 0 otherwise. This function is indeed càdlàg as well as nondecreasing in both \( t \) and \( u \).

In general, an order at time \( t \) does not affect \( S(t, u) \) itself but rather \( S(s, u) \) for \( s > t \). The same is true of a decision to ship a unit from its current position to a lower one. Letting \( t- \) or \( u- \) denote a limit from below and \( t+ \) or \( u+ \) a limit from above, we can say that an order at time \( t \) increments \( S(t+, U-) \), and a shipment from point \( u \) increments \( S(t+, u-) \).

There is a positive linear holding cost for inventory at each point in \([0, U)\), whether the unit is moving or not, and a positive linear penalty cost for backorders at point 0.

\[
\begin{align*}
b &= \text{unit backorder penalty cost rate} \\
\hat{h}(u) &= \text{unit holding cost rate at space point } u.
\end{align*}
\]

Also, let

\[
\begin{align*}
h(u, v) &= -[\hat{h}(v) - \hat{h}(u)] , \ u \leq v \\
h(u) &= h(0, u) \\
H(u) &= \int_0^u h(v) dv \\
H(u, v) &= H(v) - H(u) , \ u \leq v.
\end{align*}
\]

Assume that \( \hat{h}(u) \) is continuous on \([0, U_+)\) and continuously differentiable \((C^1)\) on \((0, U_+)\), for some \( U_+ > U \), with \( \hat{h}(U) = 0 \). The function \( h(u) \) measures how much cheaper it is to hold inventory at \( u \) instead of 0. Thus, \( h(0) = 0 \), and \( h(U) = \hat{h}(0) \). Denote the derivative of \( h \) by \( h'(u) \), and assume \( h'(u) > 0, u \in (0, U_+) \). Also, assume the limit \( h'(0+) \) exists and is finite. (Some of these assumptions can be relaxed, as mentioned in Subsection 2.4 below.) Rescale the costs so that \( \hat{h}(0) + b = 1 \).

The objective is to minimize the long-run average total cost of the system. As usual in steady-state analysis of inventory systems, we can and do ignore linear purchase and shipment costs. The average total of such costs is a constant, independent of the control, for any control with finite average cost.

Denote

\[
\begin{align*}
I(t, u) &= \text{net inventory in echelon } u \text{ at time } t \\
&= S(t, u) - D(t).
\end{align*}
\]
This is the (on-hand, non-net) inventory in echelon $u$, minus backorders at point 0. Note that the backorders is $B(t) \equiv [I(t, 0)]^-$. Thus, the inventory in echelon $u$ is $I(t, u) + B(t)$. The total cost rate at time $t$ can be written as

$$
C(t|\mathbf{S}) = \int_{U^-}^U \hat{h}(u)dI(t, u) + \hat{h}(0)[I(t, 0)]^+ + bB(t)
$$

$$
= \left\{ \hat{h}(u)[I(t, u) + B(t)] \right\}_{u=0^+}^{U^-} - \int_0^U \hat{h}'(u)[I(t, u) + B(t)]du
$$

$$
+ \hat{h}(0)[I(t, 0) + B(t)] + bB(t)
$$

$$
= \int_0^U h'(u)I(t, u)du + bB(t)
$$

$$
= \int_0^U h'(u)I(t, u)du + (\hat{h}(0) + \hat{b}) B(t)
$$

$$
= \int_0^U h'(u)I(t, u)du + [I(t, 0)]^- .
$$

The average cost is

$$
C(\mathbf{S}) = \lim \sup_{T \to \infty} \left\{ \frac{1}{T} \int_0^T C(t|\mathbf{S})dt \right\} .
$$

Suppose we follow a stationary policy, such that for each $u$ the random variables $I(t, u)$ have a stationary distribution over $t$. Let $I(u)$ denote a random variable with this same distribution. The mean cost rate in steady state is then

$$
E \left[ \int_{U^-}^U h'(u)I(u)du + [I(0)]^- \right].
$$

Under standard ergodic assumptions, $C(\mathbf{S})$ equals this quantity with probability 1.

### 2.2 Discrete-Stage Model

#### 2.2.1 Definition

Next, we consider policies where stock can be held only at discrete stages. This leads to a discrete-stage model. Time is still continuous, and costs accumulate continuously, as above. Our treatment follows Zipkin [45], Section 8.3, but with different notation.

There are $N$ stages, with $N = 2^k$ for a nonnegative integer $k$. The stages are spaced equally at distance $\varepsilon = U/N$ apart, specifically at $u = 0, \varepsilon, 2\varepsilon, \ldots, U - \varepsilon$. Stage 0 obtains its resupply from stage $\varepsilon$, stage $\varepsilon$ from stage $2\varepsilon$, etc. Stage $U - \varepsilon$ obtains stock from the outside source at $U$. This is a restriction of the model above. Goods between stages must keep moving at rate 1. We can express this condition as another constraint,

$$
S(s, n\varepsilon) = S(t, n\varepsilon + s - t) , \ t < s < t + \varepsilon , \ 0 \leq n < N ,
$$

in addition to (1). Let us refer to this as system $k$. Let $\mathcal{S}_k \subset \mathcal{S}$ denote the set of admissible controls for system $k$, namely, those which satisfy (3) as well as the three conditions above.
2.2.2 Echelon Base-Stock Policies

For $u = \varepsilon, 2\varepsilon, ..., U$, $I(t, u-)$ is called the echelon inventory transit position at $u - \varepsilon$. This is $I(t, u - \varepsilon)$ plus inventory in transit from stage $u$, or in other words, $I(t, u)$ minus the inventory at $u$ itself. It is something we partly control. As explained above, when we order a unit at time $t$, we increment $S(t+, U-)$ and thereby $I(t+, U-)$. When we pull a unit from point $u$ and send it towards $u - \varepsilon$, we increment $I(t+, u-)$.

It turns out that a stationary policy based on monitoring and controlling the $I(t, u-)$, called an echelon base-stock policy, is optimal for this system. Such a policy is specified by a càglàd (left-continuous with right limits, or LCRL) step function $y_k = y_k(u)$, $u \in (0, U]$, with steps at $u = n\varepsilon$. When $\mathbf{D}$ and hence $I(t, u)$ are integer valued, so is $y_k$. When $\mathbf{D}$ is continuous, $y_k$ is real valued.

The idea is, try to keep $I(t, u-)$ as close as possible to $y_k(u)$. The value $y_k(U)$ governs external orders. Do not order while $I(t, U-) > y_k(U)$. When $I(t, U-) \leq y_k(U)$, immediately order the difference. Thus, perhaps after a finite (with probability 1) initial period when $I(t+, U-) > y_k(U)$, the policy keeps $I(t+, U-) = y_k(U)$, and external orders precisely equal demands. The rest of $y_k$ controls flow within the system. For $u = n\varepsilon$, while $I(t, u- > y_k(u)$, make no shipments from $u$. When $I(t, u-) \leq y_k(u)$, immediately dispatch the difference from $u$, provided there is enough stock there. If not, dispatch what is there. So, again perhaps after an initial period, $I(t+, u-) \leq y_k(u)$. It is not hard to see that this policy describes an admissible control.

One can restrict attention to nonnegative $y_k$. Given any $y_k$ with $y_k(u) < 0$ for some $u$, consider the revised policy $[y_k(u)]^+$. This policy has lower backorder costs than $y_k$ and no greater holding costs.

Here is some notation: When $x$ is an integer variable, for a function $a(x)$, denote

$$\Delta_x a(x) = a(x + 1) - a(x),$$

as usual. Also, when $\varepsilon$ is understood, for a function $a(u)$ of the real variable $u$, denote

$$\Delta_u a(u) = a(u + \varepsilon) - a(u).$$

An echelon base-stock policy does lead to a stationary distribution for the $I(t, u)$. (In fact, the system arrives at this steady state after a finite initial period, with probability 1.) This distribution can be described recursively as follows: Let $\mathbf{D} = \{\hat{D}(u) : u \geq 0\}$ denote a process identical in law to $\mathbf{D}$, but with parameter $u$ instead of $t$. We use $\hat{\mathbf{D}}$ to describe equilibrium behavior, whereas $\mathbf{D}$ describes real, finite time.

\begin{align*}
I(U-) &= y_k(U) \\
I(u) &= I((u + \varepsilon)-) - \Delta_u \hat{D}(u), \quad u = n\varepsilon, \quad 0 \leq n \leq N \\
I(u-) &= \min \{y_k(u), I(u)\}, \quad u = n\varepsilon, \quad 0 < n < N.
\end{align*}

7
Here, $\Delta_u \bar{D}(u)$ is the random variable $\bar{D}(u, u + \varepsilon)$. It follows that

$$I(u) = \min_v \{ y_k(v) - \bar{D}(u, v) : v = m \varepsilon , \ n < m \leq N \}, \ u = n \varepsilon , \ 0 \leq n < N.$$  \hspace{1cm} (5)

Also, by (3), for $u \in (n \varepsilon, (n + 1)\varepsilon)$, $I(u) = I((n + 1)\varepsilon -) - \bar{D}(u, (n + 1)\varepsilon)$.

For any such policy $y_k$, let

$$c_k(u, x|y_k) = E \left[ \int_0^u h'(v)I(v)dv + [I(0)]^- \ | \ I(u-) = x \ , \ y_k \right], \ u = n \varepsilon , \ n > 0$$

$$\bar{c}_k(u, x|y_k) = E \left[ \int_0^u h'(v)I(v)dv + [I(0)]^- \ | \ I(u) = x \ , \ y_k \right], \ u = n \varepsilon , \ n \geq 0.$$  

(The conditioning event for $c_k$ means, start (4) at $u = n\varepsilon$ instead of $U$ with $I(u-) = x$, and continue the recursion from that point on using $y_k$. Interpret $\bar{c}_k$ similarly.) The variable $x$ is integer-valued for integer demand and real for continuous demand. We have $\bar{c}_k(0, x|y_k) = [x^-]$, and for $u = n\varepsilon$, $n > 0$, $\bar{c}_k(u, x|y_k) = c_k(u, y_k(u) \wedge x|y_k)$.

For $u = n\varepsilon$, $n \geq 0$,

$$c_k(u + \varepsilon, x|y_k) = E \left[ \int_0^{u+\varepsilon} h'(v)I(v)dv + [I(0)]^- \ | \ I((u + \varepsilon)-) = x \ , \ y_k \right]$$

$$= E \left[ \int_0^{u+\varepsilon} h'(v) (x - \bar{D}(v, u + \varepsilon)) dv + \bar{c}_k(u, x - \Delta_u \bar{D}(u)|y_k) \right]$$

$$= \int_0^{u+\varepsilon} h'(v) [x - \mu(u + \varepsilon - v)] dv + E \left[ \bar{c}_k(u, x - \Delta_u \bar{D}(u)|y_k) \right].$$

The integral in this expression equals

$$\{ h(v) [x - \mu(u + \varepsilon)] - \mu[H(v) - vh(v)] \}_{v = u}^{u+\varepsilon}$$

$$= \Delta_u h(u) [x - \mu(u + \varepsilon)] - \mu [\Delta_u H(u) - \Delta_u uh(u)]$$

$$= \Delta_u h(u) x - \mu [\Delta_u H(u) - \varepsilon h(u)].$$

So, the following algorithm evaluates $c_k$ and $\bar{c}_k$ recursively:

$$\bar{c}_k(0, x|y_k) = [x^-]$$

$$c_k(u + \varepsilon, x|y_k) = \Delta_u h(u) x - \mu [\Delta_u H(u) - \varepsilon h(u)] + E[\bar{c}_k(u, x - \Delta_u \bar{D}(u)|y_k)], \ u \geq 0$$

$$\bar{c}_k(u, x|y_k) = c_k(u, y_k(u) \wedge x|y_k), \ u \geq \varepsilon.$$  \hspace{1cm} (6)

The average cost (2) is $\bar{c}_k(U, y_k(U)|y_k) = c_k(U, y_k(U)|y_k)$.

This has the same form as a standard discrete-stage model, except for the term $\mu [\Delta_u H(u) - \varepsilon h(u)]$. The standard model has a term $\mu \varepsilon \Delta_u h(u)$ instead. The reason is, in the standard model, stock in
transit between stages is charged the holding cost of the upstream stage. Here, the holding cost varies continuously. This change just adds a constant to each of the functions above, independent of the policy $y_k$ and $x$. (It does depend on $u$, of course.) Also, in the standard model, the expected cost rate we call $c_k(u + \varepsilon, x|y_k)$ is normally assigned to point $u$. This is merely a difference in labeling. Our labeling turns out to be convenient.

For $y_k \geq 0$ and $x \leq 0$, an induction shows that

$$\bar{c}_k(u, x|y_k) = c_k(u, x|y_k) = h(u)x - \mu H(u) + E[\bar{D}(u) - x] \quad \text{for } x \leq 0$$

This is independent of $y_k$ and $k$.

A similar algorithm determines an optimal policy $y^*_k$ and its cost $c^*_k$.

$$\bar{c}^*_k(0, x) = [x]^-$$

$$c^*_k(u + \varepsilon, x) = \Delta_u h(u)x - \mu [\Delta_u H(u) - \varepsilon h(u)] + E[c^*_k(u, x - \Delta_u \bar{D}(u))], \quad u \geq 0$$

$$y^*_k(u) = \arg\min_x \{c^*_k(u, x)\}$$

$$\bar{c}^*_k(u, x) = c^*_k(u, y^*_k(u) \wedge x), \quad u \geq \varepsilon.$$ 

The main results are, for each $k$,

- for all $(u, x)$, $0 \leq \bar{c}^*_k(u, x) \leq c^*_k(u, x)$;
- for each $u$, $c^*_k(u, x)$ and $\bar{c}^*_k(u, x)$ are convex and $\bar{c}^*_k(u, x)$ is nonincreasing in $x$;
- $0 \leq y^*_k < \infty$, and $\bar{c}^*_k(u, x) = c^*_k(u, x) = J(u, x)$, $x \leq 0$;
- $c^*_k(u, x)$ is in fact strictly convex in $x$ for $x \geq 0$, so the minimal $y^*_k(u)$ in (8) is unique;
- for integer demand, $\Delta_x c^*_k \leq \Delta_x \bar{c}^*_k$;
- for continuous demand, $c^*_k$ and $\bar{c}^*_k$ are $C^1$ in $x$ for $x > 0$, with $\partial c^*_k/\partial x \leq \partial \bar{c}^*_k/\partial x$; and
- $y^*_k$ is indeed the best echelon base-stock policy, that is, $c^*_k(\cdot, \cdot) = c_k(\cdot, \cdot|y^*_k) \leq c_k(\cdot, \cdot|y_k)$ for any other $y_k$.

As mentioned above, this last fact implies that $y^*_k$ is optimal over all policies for system $k$. Let $c^*_k = c^*_k(U, y^*_k(U))$ denote the optimal average cost.

We remark that, in the usual presentation of the discrete-stage (and discrete-time) model with continuous demand, it is assumed that the distribution of the analogue of $\Delta_u \bar{D}(u)$ is absolutely continuous everywhere. In that case, $c^*_k$ is $C^1$ in $x$ for all $x$. Here, there is a kink at $x = 0$, due to the mass of $\Psi_0$ there. This is why we say only that $c^*_k$ is $C^1$ for $x > 0$. (Of course, it is also $C^1$ for $x < 0$.) By convexity, however, the limit $\partial c^*_k(u, 0+)/\partial x$ exists, with $\partial c^*_k(u, 0+)/\partial x \geq -[1 - h(u)]$. 


2.2.3 First-Passage Times and Stockout Probabilities

Here is an interesting characterization of $y^*_k$, due to van Houtum and Zijm [43]. For the case of continuous demand and real $x$, assume $y^*_k(u) > 0$ for all $u = n\varepsilon$, $n > 0$. Then,

$$\frac{\partial}{\partial x} c^*_k(u,x) = h(u) - \Pr\{ I(0) \leq 0 \mid I(u-) = x , y^*_k \} , \ x > 0. \quad (9)$$

In particular, since $y^*_k(u)$ solves $\frac{\partial c^*_k}{\partial x} = 0$, $\Pr\{ I(0) \leq 0 \mid I(u-) = y^*_k(u) , y^*_k \} = h(u). \quad (10)$

Moreover, as in (5),

$$[I(0) \mid I(u-) = y^*_k(u) , y^*_k] = \min_v \{ y^*_k(v) - \bar{D}(v) : v = m\varepsilon , \ 0 < m \leq n \}.$$

So, the event $\{ I(0) > 0 \mid I(u-) = y^*_k(u) , y^*_k \}$ is equivalent to $\{ \bar{D}(v) < y^*_k(v) : v = m\varepsilon , \ 0 < m \leq n \}$. As pointed out by de Kok and Fransoo [14], this can be interpreted as a statement about a first-passage time. Define the random variable

$$\tau_k = \inf_n \{ n\varepsilon : n > 0 , \ \bar{D}(n\varepsilon) \geq y^*_k(n\varepsilon) \}.$$

This is the first-passage time of the process $\bar{D}$, observed at the discrete points $n\varepsilon$, to the function $y^*_k$. So, for $u = n\varepsilon$,

$$\{ I(0) > 0 \mid I(u-) = y^*_k(u) , y^*_k \} \iff \{ \tau_k > u \}.$$

This and (10) imply

$$\Pr\{ \tau_k \leq u \} = h(u).$$

In other words, the $y^*_k(u)$ are those values, such that the first-passage time $\tau_k$ has cdf $h$. (A similar but more complex relation holds for integer demand; see Doğru, van Houtum and de Kok [15].) This connection will be useful later.

Notice that (10) specifies the stockout probability under the optimal policy, namely $h(u)$, for a system of length $u$. This relation is independent of the demand distribution. Once the holding-cost structure is known, this performance measure is immediate. This finding generalizes a well-known property of the single-stage model; see, e.g., Zipkin [45], Section 6.4. Because $h(u)$ is increasing, the longer the system is, the higher is the stockout probability. This is perhaps not obvious; larger $u$ means lower holding cost, so one might expect to find a lower stockout probability. But this logic is flawed. The result refers to the stockout probability at 0, not $u$, and between $u$ and 0 lie all the points $v < u$ with their larger holding costs.

Sobel [41] derives a related performance measure – the fill rate – for discrete-time systems. This quantity does involve the demand distribution. Consistent with the finding here, however, longer supply chains have lower fill rates.
2.3 Monotonicity and Limits

In the continuous-stage system, a stationary echelon-base stock policy can be described as a limit of discrete-stage policies above. It is specified by a càgla function \( y = y(u) \). Again, its range is determined by that of \( D \). The value \( y(U) \) governs external orders. When \( I(t, U^-) \) falls below \( y(U) \), order the difference. Again, in steady state, \( I(U^-) = y(U) \), and external orders equal demands. The rest of the policy again tries to keep \( I(t, u^-) \) as close as possible to \( y(u) \). When \( I(t, u^-) \) falls below \( y(u) \), immediately dispatch the difference from \( u \), to the extent possible given \( I(t, u) \). Again, perhaps after a finite initial period, \( I(t+, u^-) \leq y(u) \). Clearly, the corresponding \( S \in S \). It is not hard to show that, in equilibrium, as in (5),

\[
I(u) = \inf_v \{ y(v) - \hat{D}(u, v) : u < v \leq U \}. \tag{11}
\]

Note that each \( y_k \) above counts as a policy for the continuous system. In fact, any policy \( y(u) \) that is constant over an interval \( u \in (v, w) \) retains no inventory in the interval in steady state. (Of course, all inventory must pass through the interval, but none of it stops or slows down there.) So, a step-function \( y \) retains inventory only at its jump points.

The functions \( c_k(u, x|y_k) \) are defined in Subsection 2.2.2 above for discrete \( u = n \varepsilon \). We can interpolate them for all \( u \) as follows: For \( n \varepsilon < u \leq (n + 1) \varepsilon \), set

\[
c_k(u, x|y_k) = h(n \varepsilon, u)x - \mu [H(n \varepsilon, u) - (u - n \varepsilon)h(n \varepsilon)] + E[\bar{c}_k(n \varepsilon, x - \hat{D}(n \varepsilon, u)|y_k)]. \tag{12}
\]

Also, define \( c_k(0, x|y_k) = [x]^-. \) Similarly, we can interpolate the functions \( c_k^+ \). Since each \( \bar{c}_k^+(n \varepsilon, x) \) is convex in \( x \), so is \( c_k^+(u, x) \) for all \( u \). Also, set \( y_k^-(u) = \arg \min_x \{ c_k^+(u, x) \}, u > 0 \). (Note that this is not the same as the step function \( y_k^* \), although they agree at the points \( u = n \varepsilon \).) Also, interpolate \( \bar{c}_k^+(u, x) = c_k^+(u, y_k^-(u) \land x), n \varepsilon < u < (n + 1) \varepsilon \).

The recursion (6) evaluates any step-function policy \( y_k \). Now, consider a policy \( y \) for the continuous-stage problem. Construct a sequence of step-function policies \( y_k \) that approximate \( y \) in a sensible way, e.g., \( y_k(u) = y([u/\varepsilon] \varepsilon) \). It is not hard to see that the functions \( c_k(u, x|y_k) \) converge to the limit \( c(u, x|y) \), where

\[
c(u, x|y) = E \left[ \int_0^u h'(v)I(v)dv + [I(0)]^- \mid I(u^-) = x , y \right], \tag{13}
\]

and, as in (11),

\[
[I(v) \mid I(u^-) = x , y] = \min \left\{ x - \hat{D}(v, u) , \inf_w \{ y(w) - \hat{D}(v, w) : v < w < u \} \right\}. \tag{14}
\]

(The sequence \( c_k(u, x|y_k) \) just expresses the definition of this integral.) The average cost (2) of the policy is thus \( c(U, y(U)|y) \). Also, by (7), for \( y \geq 0 \) and \( x \leq 0 \),

\[
c(u, x|y) = J(u, x), \tag{15}
\]

11
independent of $y$.

Demonstrating the convergence of the $c^*_k$ is a bit harder:

**Proposition 1**  
(a) For each $(u,x)$, $c^*_k(u,x)$ and $\bar{c}^*_k(u,x)$ are nonincreasing in $k$. These functions converge pointwise to a common limit $c^*$ as $k \to \infty$. This function $c^*(u,x)$ is convex and nonincreasing in $x$ for each $u$.

(b) For integer (continuous) demand, $\Delta x c^*_k$ and $\Delta x \bar{c}^*_k$ ($\partial c^*_k / \partial x$ and $\partial \bar{c}^*_k / \partial x$) are nonincreasing in $k$. These functions too converge to limits as $k \to \infty$. For each $u > 0$, $y^*_k(u)$ and $\bar{y}^*_k(u)$ are nondecreasing in $k$, and they converge to a limit $y^*(u)$ (for now possibly $\infty$). This $y^*(u)$ minimizes $c^*(u,x)$ over $x$.

(c) For integer (continuous) demand, $\Delta x \bar{c}^*_k$ ($\partial \bar{c}^*_k / \partial x$) converges to its limit uniformly in $(u,x)$. (Thus, $c^*$ is $C^1$ in $x$, $x > 0$.) For both cases, $c^*$ is continuous in $u$, and $y^*$ is càglàd.

(d) The policy $y^*$ is optimal among echelon base-stock policies. Specifically, $c^*(\cdot, \cdot) = c(\cdot, \cdot | y^*) \leq c(\cdot, \cdot | y)$ for all $y$.

(e) For $u > 0$ and $x < y^*(u)$, $c^*$ is $C^1$ in $u$, and

$$\frac{\partial}{\partial u} c^*(u, x) = h'(u)x + \lambda[E[c^*(u, x - X)] - c^*(u, x)].$$

(16)

For integer demand, presuming $y^*(u) < \infty$, this holds also for $x = y^*(u)$.

(f) For integer demand,

$$y^*(u) = \inf \left\{ x : \Delta x c^*(u-, x') < 0 \ , \ x' < x \right\},$$

(17)

$$y^*(u) = \inf \left\{ x : x \geq 0 \ , \ h'(u) + \lambda E[\Delta x c^*(u, x - X)] \geq 0 \right\},$$

(18)

and $y^*(u)$ is finite for all $u > 0$. Analogous results hold for continuous demand.

All proofs of this paper are in the Appendix.

Let $c^{**} = c^*(U, y^*(U))$ denote the limiting optimal average cost.

**Proposition 2**  
The policy $y^*$ is optimal for the continuous-stage system.

Note that, by the proof of Proposition 1(c), the first-passage-time interpretation for continuous demand also carries over to the limit. Specifically,

$$\Pr \{ \tau \leq u \} = \Pr \{ I(0) \leq 0 \mid I(u-) = y^*(u) \ , \ y^* \} = h(u).$$

Again, under the optimal policy, $h(u)$ is the cdf of $\tau$, as well as the stockout probability of a system of length $u$. 

12
Proposition 1 also implies that \( c_0^*(u, x) \) is an upper bound on \( c^*(u, x) \), and \( y_0^-(u) \) is a lower bound on \( y^*(u) \). These bounds are easy to compute. By (12),

\[
c_0^*(u, x) = h(u)x - \mu H(u) + \Psi_1(u, x).
\]

And,

\[
y_0^-(u) = \arg \min_x \{c_0^*(u, x)\}
\]

In the case of continuous demand, \( y_0^-(u) \) solves

\[
0 = \frac{\partial}{\partial x} c_0^*(u, x) = h(u) - \Psi_0(u, x),
\]

provided this equation has a positive solution; otherwise, \( y_0^-(u) = 0 \). For integer demand, the corresponding first-order condition is \( y_0^-(u) = \min \{x : h(u) - \Psi_0(u, x) \geq 0\} \). This calculation for each \( u \) is just the solution of a single-stage system, which has the same form as a newsvendor problem.

This result is analogous to the bounds for discrete-stage systems obtained by Gallego and Zipkin [22], Shang and Song [40], Dong and Lee [16], and Gallego and Özer [23]. (They also derive lower bounds on \( c_k^* \) and upper bounds on \( y_k^* \). We have not been able to extend those results to the continuous-stage model.)

2.4 Discussion and Extensions

Notice that \( y_0^-(u) \) need not be increasing in \( u \). It is true that \( \bar{D}(u) \) is stochastically increasing in \( u \), and so \( \Psi_0(u, x) \) is increasing. So, the solution to the equation \( \Psi_0(u, x) = \bar{h} \) for a fixed constant \( \bar{h} \) is increasing in \( u \). But \( h(u) \) is not fixed. It is increasing, and this tends to reduce the solution \( y_0^-(u) \). The net effect of these opposing influences depends on the particulars of \( h \) and \( \Psi_0 \). The same contention appears in (8) for larger \( k \). In \( c_k^*(u + \epsilon, x) \), the expectation over \( \Delta_u \bar{D}(u) \) tends to push \( y_k^*(u + \epsilon) \) above \( y_k^*(u) \), but the positive linear term \( \Delta_u h(u)x \) tends to pull it down. Thus, we should not be surprised to find a \( y_k^* \) that is decreasing in some places. The same logic applies to the limiting \( y^* \).

However, for both the discrete- and continuous-stage systems, given a policy \( y \) which is not nondecreasing, one can construct another policy \( \hat{y} \) which is nondecreasing and which is equivalent to \( y \). Specifically, \( \hat{y}(u) = \min \{y(v) : u \leq v \leq U\} \). (For example, consider system \( k \) with \( y_k(u) > y_k(u + \epsilon) \) for some \( u \). After a finite time, the system reaches and remains in states with \( I(t, (u + \epsilon)^-) \leq y_k(u + \epsilon) \). So, \( I(t, u-) \leq I(t, (u + \epsilon)^-) \leq y_k(u + \epsilon) \). The policy tries to pull more stock from \( u + \epsilon \), to raise \( I(t, u-) \) closer to \( y_k(u) \), but it can’t; there is no more stock there to pull. One can prove the equivalence formally using (5).) So, having computed an optimal policy \( y_k^* \)
(or \(y^*\)), one can construct another policy \(\hat{y}_k^*\) (or \(\hat{y}^*\)), also optimal, which is nondecreasing. (The first-passage-time interpretations, however, apply only to \(y_k^*\) and \(y^*\), not \(\hat{y}_k^*\) and \(\hat{y}^*\).)

This notion allows us to relax the assumption that \(h'(u) > 0\). For system \(k\), in (8), if \(\Delta_u h(u - \varepsilon) \leq 0\), then \(c_k^*(u, x)\) is decreasing in \(x\), so \(y_k^*(u) = \infty\), and \(\bar{c}_k^*(u, x) = c_k^*(u, x)\). The recursion can continue anyway. The resulting policy \(y_k^*\) includes some infinite values, which are awkward to interpret or implement. However, provided \(y_k^*(U) < \infty\), the policy \(\hat{y}_k^*\) has only finite values. The same considerations apply to \(y^*\). As in Proposition 1(f), to ensure \(y^*(U) < \infty\), we must still assume \(h'(U) > 0\). Otherwise, the model is ill posed, like a single-stage model with non-positive holding cost.

Also, we can allow discontinuous (but still càdlàg) \(h\) and \(\hat{h}(U) > 0\). These extensions complicate the notation and arguments, but not the basic results.

Incidentally, the restriction to equally spaced points above is purely for convenience. One can define the recursions (6) and (8) for any sequence of points \(u\). We shall see an example in the next section. The monotonicity results in Proposition 1 go through for any refinement of a given sequence (that is, a new sequence obtained by adding more points to an old one). The convergence results also remain valid, provided the sequence becomes dense.

To further understand the model and its solution, the following sections examine a specific demand process and an approximation.

3 Poisson Demand

This section assumes \(D\) is a Poisson process with rate \(\mu = \lambda\). The variable \(x\) is integer valued. Several of the properties established in Proposition 1 simplify in this case. Based on these results, we develop a simple, explicit method to compute an optimal policy.

3.1 Properties

First, consider a fixed echelon base-stock policy, specified by a nonnegative, nondecreasing, integer-valued, càdlàg step function \(y\) with finite \(y(U)\). So, \(y\) has a finite number of jump points. Label them \(u_n\), with \(0 = u_0 < u_1 < u_2 < \cdots < u_N = U\). The policy reduces the system to a discrete-stage system, but with unequal step sizes. We can thus evaluate the policy by an algorithm like (6):

\[
c(0, x|y) = [x].
\]  

\[
c(u_{n+1}, x|y) = h(u_n, u_{n+1})x - \mu [H(u_n, u_{n+1}) - (u_{n+1} - u_n)h(u_n)]
\]  

\[
+ E[\bar{c}(u_n, x - \hat{D}(u_n, u_{n+1})|y)], \quad x \leq y(u_{n+1}), \quad n \geq 0
\]

\[
\bar{c}(u_n, x|y) = c(u_n, y(u_n) \land x|y), \quad n > 0.
\]
The average cost is \( c(U, y(U)|y) \).

We can explicitly write the expectation above. Denote
\[
\psi(\delta, i) = \Pr\{\hat{D}(\delta) = i\} = \frac{(\mu \delta)^i e^{-\mu \delta}}{i!},
\]
the Poisson probability mass function, and set \( \delta = u_{n+1} - u_n \). Using (15), for \( 0 < x \leq y(u_{n+1}) \),
\[
E[\hat{c}(u_n, x - \hat{D}(u_n, u_{n+1})|y)] = \sum_{i=0}^{x-1} \psi(\delta, i)\hat{c}(u_n, x - i|y) + \sum_{i=x}^{\infty} \psi(\delta, i)J(u_n, x - i)
\]
\[
= \sum_{i=0}^{x-1} \psi(\delta, i)\hat{c}(u_n, x - i|y) + \Psi_1(\delta, x)[1 - h(u_n)] + \Psi_0(\delta, x - 1)\mu[u_n - H(u_n)].
\]

To find the optimal policy and cost, we must proceed more carefully, because we do not yet know the jump points of \( y^* \). We can obtain some useful information from Proposition 1. Here, (16) becomes
\[
\frac{\partial}{\partial u}c^*(u, x) = h'(u)x - \lambda \Delta_x c^*(u, x - 1), \quad x \leq y^*(u);
\]
(23) since we know \( y^*(u) \) is finite, we can write (17) as
\[
y^*(u) = \inf \{ x : \Delta_x c^*(u-, x - 1) < 0 \} ;
\]
(24) and (18) becomes
\[
y^*(u) = \inf \{ x : x \geq 0 , \ h'(u) + \lambda \Delta_x c^*(u, x - 1) \geq 0 \}
\]
(25) These conditions are enough to construct the entire solution.

Let \( u^*_n \) denote the jump points of \( y^* \). Suppose we have found \( u^*_n \) and \( c^*(u^*_n, x) \), and we seek the next jump point. Suppose \( y^*(u) \) remains at the constant value \( \bar{y}^* = y^*(u^*_n+) \) on some interval \( (u^*_n, v) \), but \( v \) is a potential jump point. As in (22), for \( u \in (u^*_n, v) \),
\[
c^*(u, x) = h(u^*_n, u)x - \mu [H(u^*_n, u) - (u - u^*_n)h(u^*_n)] + E[c^*(u^*_n, x - \hat{D}(u^*_n, u))], \quad x \leq \bar{y}^*.
\]
(26) Two things could happen that would force a jump at \( v \). The first is \( \Delta_x c^*(v, \bar{y}^* - 1) = 0 \), for then (24) would be violated if we kept \( y^*(v+) = \bar{y}^* \). The second is \( \Delta_x c^*(v, \bar{y}^* - 1) = -h'(v)/\lambda \), for then keeping \( y^*(v+) = \bar{y}^* \) would violate (25). The remedy in each case is clear: In the first case, set \( y^*(v+) = \bar{y}^* - 1 \), and in the second, set \( y^*(v+) = \bar{y}^* + 1 \).

3.2 Solution

We can now see how to find the next jump point, \( u^*_{n+1} \). For \( v > u^*_n \), keep \( y^*(v) = \bar{y}^* = y^*(u^*_n) \), provided (24) and (25) remain valid, that is,
\[
-h'(v)/\lambda < \Delta_x c^*(v, \bar{y}^* - 1) < 0.
\]
Set $u_{n+1}^*$ to the first point where $\Delta_x c^*$ hits one of these boundaries. At that point, switch $y^*(v^+)\uparrow$ or down, depending on whether the lower or upper boundary is hit. (Note that $\Delta_x c^*(v, y^* - 1)$ cannot touch a boundary at a critical point. For instance, if $\Delta_x c^*(v, y^* - 1) = 0$, (23) implies $(\partial / \partial u) \Delta_x c^*(v, y^* - 1) < 0$.)

To track $\Delta_x c^*(u, y^* - 1)$ over $u > u_n^*$, note that (26) implies

$$\Delta_x c^*(u, x) = h(u_n^*, u) + E[\Delta_x c^*(u_n^*, x - D(u_n^*, u))], \ x < y^*.$$ 

More explicitly, setting $\delta = u - u_n^*$,

$$\Delta_x c^*(u_n^* + \delta, x) = h(u_n^*, u_n^* + \delta) + \sum_{i=0}^{x} \psi(\delta, i) \Delta_x c^*(u_n^*, x - i)$$

$$- [1 - h(u_n^*)] \Psi_0(\delta, x), \ 0 \leq x < y^*,$$

Regarding $\delta$ as a variable, everything in (27) is constant, except $h(u_n^*, u_n^* + \delta)$ and the Poisson probabilities. It is thus not hard to solve numerically the two equations in $u$

$$\Delta_x c^*(u, y^* - 1) = 0,$$

$$\Delta_x c^*(u, y^* - 1) = -h'(u)/\lambda$$

to find $u_{n+1}^*$. (Since $y^*$ is normally nondecreasing, one can solve the second equation first, and then check that the first equation has no smaller solution.)

Once we have located $u_{n+1}^*$, we can compute $\Delta_x c^*$ there by (27). Using those data, we can begin searching for the next jump point.

To set the initial $y^*(0^+)$, the logic is similar to the switching analysis above. We just summarize the results: First, for $h'(0^+) > 0$, $y^*(0^+) = 0$ or 1. To decide between these values, compare $\lambda$ and $h'(0^+)$. If $\lambda < h'(0^+)$, then $y^*(0^+) = 1$, and if $\lambda > h'(0^+)$, then $y^*(0^+) = 0$. (In case of a tie, a higher-order analysis is needed. If $h$ is twice differentiable, compare $\lambda^2$ and $h''(0)$ similarly, and so on.) For $h'(0^+) = 0$, $y^*(0^+)$ can be larger than 1. Set $y^*(0^+) = y^*$ to the smallest value, such that (24) and (25) remain valid for small $u$.

The overall procedure, then, is to start at $u = 0$, and determine the jump points of $y^*$ sequentially. In contrast to the discrete-stage model, the procedure itself selects the locations of stages where stock is held.

One question remains: Is the number of jump points finite? It is not hard to construct examples where $h'(u)$ fluctuates rapidly, which leads to equally rapid fluctuations in $y^*(u)$. We suspect that it is possible to ensure finite jumps by imposing some regularity condition on $h'(u)$, but we have not done so. In all the examples we have tried using “nice” cost functions, including those mentioned below, the number of jumps is indeed finite.

16
Incidentally, by Proposition 1(c), $c^*(u, x)$ is continuous in $u$ on all of $[0, U)$, and so the same is true of $\Delta_x c^*(u, x)$. In particular, this remains true at jump points for those $x$ which shift from one side of the boundary to the other. Consequently, by (23), even $\partial c^*/\partial u$ is continuous for all $u$ and $x$. So, we can view (23) alternatively as a differentio-difference equation over $(0, U)$. This equation, together with the boundary conditions (24)-(25), characterizes $c^*$.

4 Brownian-Motion Approximation

4.1 Discussion

Assume the equilibrium demand process $\bar{D}$ has positive, finite variance rate $\sigma^2 = V[\bar{D}(1)]$. We shall approximate $\bar{D}$ by a Brownian motion $\tilde{D}$ with the same first two moment parameters, $\mu$ and $\sigma^2$. The state variable $x$ is now continuous.

For a single-stage system, the normal approximation of leadtime demand (which corresponds to $\bar{D}(U)$ here) is a well established technique. It is often accurate. It is not only makes computation easier, it also reveals the primary effects of the system’s parameters on the optimal policy and its cost. It is thus of great practical and pedagogical value. The aim of this section is to investigate whether analogous results can be found for supply streams. As we shall see, the answer is yes, partly. We show that the effects of $\mu$ and $\sigma^2$ on the solution are quite transparent, as for single-stage systems. (See (28) below.) Also, for certain cost functions $h$, we obtain the entire solution in closed form, and for others we obtain nearly explicit solutions. But in general, the computation remains challenging. We explain the challenges and present methods for dealing with them.

Recall that we assumed above that the demand increments are nonnegative. The theory underlying the discrete-stage model relies on this assumption. A Brownian motion, of course, violates it. So, we cannot allow $D$ to be a Brownian motion. However, it is possible to perform the calculations indicated in (6) and (8) using normal demand increments, treating the results as approximations. (Federgruen and Zipkin [19] do just that for the case of two stages.) This is exactly analogous to the use of normal demand in a single-stage model.

Specifically, consider the discrete model of Subsection 2.2 with one stage, that is, $k = 0$ and $N = 1$. A base-stock policy has only one parameter, $y = y(U)$. After a finite time, $I(t, U-) = y$, and $I(t + U, 0) = y - D(t, t + U)$. In equilibrium, therefore, $I(U-) = y$, and $I(0) = y - \tilde{D}(U)$. The average cost as a function of $y$ is $c_0(U, y|y)$. Now, we can approximate this function by another one, replacing $\bar{D}(U)$ by $\tilde{D}(U)$. We expect this approximation to be accurate, precisely when the central limit theorem applies, namely, when $\bar{D}(U)$ is the sum of many (say $M$) independent components. If so, optimization over the approximate function will produce good results. Abundant numerical evidence supports this notion (see, e.g., Zipkin [45], Section 6.4).
We apply the same idea for each \( k \). The equilibrium distribution is described by (4), and we approximate the increments \( \Delta_u \bar{D}(u) \) there and in the equivalent cost functions of (6) and (8) by \( \Delta_u \tilde{D}(u) \). Again, we expect the approximation to be accurate under the conditions of the central limit theorem. However, as \( N \) increases, we divide \( \bar{D}(U) \) into smaller and smaller pieces. To continue to invoke the central limit theorem at each step, therefore, we must assume that \( M \) increases along with \( N \).

Most of the development of Section 2 remains valid with this approximation. For instance, the functions \( c^*_k \) are convex in \( x \), and they converge to a limit \( c^* \). There are some differences, however. It is no longer necessarily true that \( y_k^* \geq 0 \) and \( y^* \geq 0 \). Also, neither \( c^*_k(u,x) \) nor \( c^*(u,x) \) equals \( J(u,x) \) for \( x \leq 0 \). (However, as \( x \to -\infty \), the difference goes to 0.) The smoothness properties of \( c^*_k \) and \( c^* \) are also different; we shall consider those below. And of course, Proposition 2 does not apply—we are not discussing the original control problem, only the class of echelon base-stock policies in equilibrium.

We performed the procedure for the parameters shown in Figures 1-3 (\( \sigma^2 = \mu = 16 \)). We compared the results to those for Poisson demand, making no adjustments for increasing \( N \). The resulting curves for \( y^* \) and \( y_0^- \) are indistinguishable by eye from the exact ones in the figures, and the optimal costs are nearly identical to those of Table 1. Of course, there are small differences, since the policies are integer valued for Poisson demand. Still, the overall approximation is quite accurate.

Several authors, starting with Bather [2], have studied single-stage models, assuming Brownian-motion demand \( D \) at the outset. These models are different from ours, and they lead to different results. For example, in the special case of Bather’s model with no fixed order cost, his policy reduces to a base-stock policy. The analogue of \( I(U-\) is not the constant \( y \), but instead \( y \) plus an exponential random variable, representing the negative excursions of \( D \). The average cost of any such policy, and the optimal one, are also different.

DeCroix et al. [13] consider a discrete-stage system with returns as well as demands. This leads to net demand increments that can be negative. They show how to do an exact analysis. Suffice it to say that this requires far more intricate calculations than the model here.

Our approach differs also from that of Harrison [25] and others to the approximation of control problems. There, the starting point is an intractable problem. One considers a sequence of such problems which, when properly rescaled, converges to a limiting control problem, whose dynamics are driven by a Brownian motion. This problem often is much easier to analyze than the original. Our problem, in contrast, is quite tractable, at least qualitatively. Here, Brownian motion enters the analysis only in equilibrium and only for a class of policies known to be optimal. The purpose is mainly to simplify the calculations and sensitivity analysis. (Recall, however, that our system
is uncapacitated. The approach of Harrison is most powerful in dealing with capacitated systems. We suspect that, to extend our model to finite capacities, something like that approach will prove essential.

4.2 Transformations

We now apply a sequence of transformations to $c_k^*$. The aim is to obtain a direct characterization of the first-passage times $\tau_k$ and their limit as $k \to \infty$.

Denote by $\phi$ the standard normal density function and $\Phi_0$ its complementary cdf. The interpolation formula (12), with the approximate increments $\tilde{D}(n\varepsilon, u)$ in place of $\bar{D}(n\varepsilon, u)$, implies that, for all $u > 0$ and $x \in \mathbb{R}$, $c_k^*$ is twice continuously differentiable ($C^2$) in $x$. This (and more – see Cannon [4]) follows from the smoothness of $\phi$.

First, define

$$p_k^*(u, x) = c_k^*(u, x) - J(u, x).$$

In these terms, the recursion (8) with the approximate demand increments $\Delta D(u)$ becomes

$$\bar{p}_k^*(0, x) = [x]^+, \quad p_k^*(u + \varepsilon, x) = E[p_k^*(u, x - \Delta D(u))], \quad u \geq 0$$

$$\frac{\partial p_k^*(u, y_k^*(u))}{\partial x} = 1 - h(u)$$

$$\bar{p}_k^*(u, x) = p_k^*(u, y_k^*(u) \wedge x) + [1 - h(u)] [x - y_k^*(u)]^+, \quad u \geq \varepsilon.$$

Next, standardize the variable $x$, the policy $y_k^*$ and $D$. Denote $w = \frac{x - \mu u}{\sigma}$, $z_k^*(u) = \frac{y_k^*(u) - \mu u}{\sigma}$, $W(u) = \frac{\tilde{D}(u) - \mu u}{\sigma}$, and define

$$q_k^*(u, w) = p_k^*(u, \mu u + \sigma w)/\sigma.$$

Now, (8) becomes

$$\bar{q}_k^*(0, w) = [w]^+$$

$$q_k^*(u + \varepsilon, w) = E[q_k^*(u, w - \Delta W(u))], \quad u \geq 0$$

$$\frac{\partial q_k^*(u, z_k^*(u))}{\partial w} = 1 - h(u)$$

$$\bar{q}_k^*(u, w) = q_k^*(u, z_k^*(u) \wedge w) + [1 - h(u)] [w - z_k^*(u)]^+, \quad u \geq \varepsilon.$$

This construction shows that, in the original recursion (8), up to the normal approximation, the parameters $\mu$ and $\sigma$ affect the solution as linear translation and scale factors. Specifically, $y_k^*(u) = \mu u + \sigma z_k^*(u)$, and

$$c_k^{**} = c_k^*(U, y_k^*(U)) = J(U, \mu U + \sigma z_k^*(U)) + \sigma q_k^*(U, z_k^*(U)).$$

(28)
The same is true of the limits \( y^* \) and \( c^{**} \). This effect is precisely analogous to the familiar result for single-stage systems mentioned above. Note that the effect of \( \sigma^2 \) is correctly described as a square-root law. However, recalling the discussion in Subsection 2.4, here we are comparing two entire systems, not the solutions for different \( u \) within one system.

Now, since \( c^*_k \) is \( C^2 \) in \( x \), \( q^*_k \) is \( C^2 \) in \( w \). So, we can define

\[
G^*_k(u, w) = \frac{\partial q^*_k(u, w)}{\partial w} \quad \text{and} \quad g^*_k(u, w) = \frac{\partial^2 q^*_k(u, w)}{\partial w^2}
\]

In terms of \( G^*_k \), (8) becomes

\[
\bar{G}^*_k(0, w) = 1 \{ w \geq 0 \} \quad (29)
\]

\[
G^*_k(u + \varepsilon, w) = E[\bar{G}^*_k(u, w - \Delta u W(u))], \quad u \geq 0
\]

\[
G^*_k(u, z^*_k(u)) = 1 - h(u)
\]

\[
G^*_k(u, w) = G^*_k(u, z^*_k(u) \wedge w), \quad u \geq \varepsilon.
\]

And, in terms of \( g^*_k \), denoting by \( \delta \) the Dirac delta centered at 0,

\[
\bar{g}^*_k(0, w) = \delta(w) \quad (30)
\]

\[
g^*_k(u + \varepsilon, w) = E[g^*_k(u, w - \Delta u W(u))], \quad u \geq 0
\]

\[
\int_{-\infty}^{z^*_k(u)} g^*_k(u, v) dv = 1 - h(u)
\]

\[
\bar{g}^*_k(u, w) = \begin{cases} g^*_k(u, w) & w < z^*_k(u) \\ 0 & w \geq z^*_k(u) \end{cases}, \quad u \geq \varepsilon.
\]

Now, a careful look at (29) and (30) reveals that they describe the first-passage time \( \tau_k \) of the process \( \tilde{D} \) observed at discrete points \( u = n\varepsilon \) to the boundary \( y^*_k \). (We do not need the condition \( y^*_k(u) > 0 \) here to get (10), due to the smoothness of \( c^*_k \).) First, note that \( \tau_k \) is also the first-passage time of the standardized process \( W \) to the standardized boundary \( z^*_k \).

**Proposition 3**

\[
G^*_k(u, w) = \Pr\{W(u) \leq w, \quad \tau_k > u - \varepsilon \}, \quad u = n\varepsilon > 0
\]

\[
G^*_k(u, w) = \Pr\{W(u) \leq w, \quad \tau_k > u \}, \quad u = n\varepsilon \geq 0.
\]

Consequently, for \( u > 0 \), \( \bar{g}^*_k(u, w) \) is the probability density of \([W(u) = w, \quad \tau_k > u]\) (that is, the conditional density of \([W(u) = w \mid \tau_k > u]\), times \( \Pr\{\tau_k > u\} \)).
4.3 Limiting First-Passage Time

Let us now consider the limit as $k \to \infty$. As in (36) above, let $\tau$ denote the first-passage time of $W$, observed at all $u > 0$, to the limiting $z^*$, the standardized version of $y^*$. By Proposition 1, we know that $G^*_k$ converges to a limit, say $G^*$, and $\bar{G}^*_k$ converges to the same limit. Indeed, the limit of (29) as $k \to \infty$ is one way to construct $\tau$, using a standard construction of $W$. It follows that

**Corollary 4**

$$G^*(u,w) = \Pr\{W(u) \leq w, \tau > u\}.$$  

The first-passage time $\tau$ of a standard Brownian motion $W$ to a boundary $z^*$ and its distribution $h$ have received much scrutiny. The original first-passage time problem takes the boundary $z^*$ as given and aims to determine $h$. Our problem (given $h$, determine $z^*$) is called the inverse first-passage time problem.

For the original problem, clearly, any càdlàg boundary $z^*$ leads to a well-defined distribution $h$. ($z^*(u)$ can even be infinite on certain intervals. If so, $h$ is constant on those intervals.) However, many boundaries lead to the trivial distribution $h(0) = \Pr\{\tau = 0\} = 1$. To avoid this, the boundary must be an upper function. The precise condition is rather technical (see, e.g., Peskir [32]). A sufficient condition is $z^*(0+) > 0$. For an upper $z^*$, $h(0) = \Pr\{\tau = 0\} = 0$. If an upper $z^*$ is continuous on $[0, \infty]$, then so is $h$. And, if $z^*$ is $C^1$ on $(0, \infty)$, then so is $h$, with $h'(u) > 0, u > 0$.

(See Peskir [33], Theorem 4.1, and the references therein.)

Cheng et al. [8] consider the inverse problem. (They consider the first passage of $W$ down to a lower boundary, instead of up to an upper boundary. We translate their results to the latter scenario.) Their setting is quite general. Their $W$ is a diffusion process. (Since we assume Brownian motion, however, we state only their results for this case.) They assume $h$ is any càdlàg cdf with $h(0) = 0$. Their approach begins with a variational inequality. They prove that this inequality has a unique solution, namely $G^*$, in the viscosity sense. (Under our stronger conditions on $h$, our Propositions 1 and 3 comprise an alternate proof that $G^*$ is indeed the solution, though not all of its properties.) The boundary $z^*$ is then defined as $z^*(u) = \sup\{w : G^*(u,w) < 1 - h(u)\}$. This is an upper function. When $h$ is continuous, then so is $G^*$. When $h$ is also $C^1$ on $(0, \infty)$, then $G^*$ is $C^1$ in $(u,w)$ for $u > 0$ and $w \in \mathbb{R}$. Also, if $h'(u) > 0$, then $z^*(u)$ is finite.

Most importantly for our case (where $h$ is $C^1$), for $u > 0$,

$$g^*(u,w) = \lim_{k \to \infty} \{g^*_k(u,w)\} = \lim_{k \to \infty} \{\bar{g}^*_k(u,w)\} = \frac{\partial G^*(u,w)}{\partial w}.$$  

This limit is continuous, and it is the density of $[W(u) = w, \tau > u]$. This implies that the limiting $c^*$ is $C^1$ in $(u,w)$ and $C^2$ in $w$. Moreover, $G^*$ itself is $C^2$ in $w$ for $w < z^*(u)$, $u > 0$, and $G^*$ and $z^*$
jointly satisfy a partial differential equation:

\[ G^*(0+, w) = 1 \{ w \geq 0 \}, \]

and for \( u > 0 \),

\[
\frac{\partial G^*}{\partial u} = \frac{1}{2} \frac{\partial^2 G^*}{\partial w^2}, \quad w < z^*(u) \\
G^*(u, z^*(u)) = 1 - h(u) \\
G^*(u, w) = G^*(u, z^*(u)), \quad w \geq z^*(u).
\]

(One can also derive this pde from (29) by a limiting argument, assuming \( G^* \) is smooth enough.) This in turn implies (see, e.g., [4]) that \( g^* \) is \( C^1 \) in \((u, w)\) and \( C^2 \) in \( w \) for \( w < z^*(u), \ u > 0 \), and \( g^* \) and \( z^* \) jointly satisfy the pde

\[ g^*(0+, w) = \delta(w), \]

and for \( u > 0 \),

\[
\frac{\partial g^*}{\partial u} = \frac{1}{2} \frac{\partial^2 g^*}{\partial w^2}, \quad w < z^*(u) \\
\frac{1}{2} \frac{\partial g^*}{\partial w}(u, z^*(u)) = -h'(u) \\
g^*(u, z^*(u)) = 0 \\
g^*(u, w) = 0, \quad w \geq z^*(u).
\]

The equation involving \( \partial g^*/\partial w \) follows from

\[ \frac{\partial G^*}{\partial u} = \frac{1}{2} \frac{\partial^2 G^*}{\partial w^2} = \frac{1}{2} \frac{\partial g^*}{\partial w}. \]

This pde is essentially the same one used to describe the original first-passage time problem. See, for example, Daniels [11], Lerche [29] and Durbin [17].

Let us restate this last pde in a slightly different form.

\[ g^*(0+, w) = \delta(w), \quad (31) \]

and for \( u > 0 \),

\[
\frac{\partial g^*}{\partial u} = \frac{1}{2} \frac{\partial^2 g^*}{\partial w^2}, \quad w \in \mathbb{R} \\
\frac{1}{2} \frac{\partial g^*}{\partial w}(u, z^*(u)) = -h'(u) \\
\tilde{g}^*(u, z^*(u)) = 0 \\
g^*(u, w) = \tilde{g}^*(u, z^*(u) \wedge w). \quad (35)
\]
In this formulation we seek two functions, $g^*$ and an auxiliary function $\tilde{g}^*$, linked by the last equation (35). The auxiliary function extends the pde past the boundary. In this sense it is simpler than $g^*$ itself. The key point is, we do not care at all about $\tilde{g}^*(u, w)$ for $w > z^*(u)$. This flexibility will prove useful later.

Let us briefly summarize some other results on first-passage times of Brownian motion:

- The exact solution is known in a few special cases, discussed below. But the general problem remains challenging. There are several numerical methods available, also discussed below.
- A good deal is known about the the asymptotic behavior (for small and large $u$) of $h$ and $z^*$. See, e.g., Peskir [32], Redner [36] and Zipkin [46].
- The exact relation between $h$ and $z^*$ can be expressed by certain integral equations. See, e.g., Peskir [33] and Jaimungal et al. [26].
- The latter paper also shows the effects on $z^*$ of certain transformations of $h$ and vice versa. The simplest of these is a change of scale: For any positive $\gamma$, if we change $h(u)$ to $h^\gamma(u) = h(u/\gamma)$, then the corresponding boundary is $z^\gamma(u) = \sqrt{\gamma} z^*(u/\gamma)$. The reader may recognize this as a diffusion scaling, which rescales both the $u$ and the $w$ dimensions in a way that preserves the law of $W$. The effects on the other parts of the supply-stream model are equally simple: $U^\gamma = \gamma U$, $H^\gamma(u) = \gamma H(u/\gamma)$, $J^\gamma(u, x) = \gamma J(u/\gamma, x/\gamma)$, $q^\gamma(u, w) = \sqrt{\gamma} q^*(u/\gamma, w/\sqrt{\gamma})$, $y^\gamma(u) = \gamma \mu(u/\gamma) + \sqrt{\gamma} \sigma z^*(u/\gamma)$, and, as in (28),

$$c^\gamma = -[1 - h(U)] [\gamma \mu U + \sqrt{\gamma} \sigma z^*(U)] + \gamma \mu [U - H(U)] + \sqrt{\gamma} \sigma q^*(U, z^*(U)).$$

The effect of $\gamma$, then, is similar to that of changing $\mu$ and $\sigma^2$, keeping $\sigma^2/\mu$ constant.

### 4.4 The Method of Images and Closed-Form Solutions

The method of images, a venerable technique in physics, was introduced to first-passage problems by Daniels [11] and elaborated by Lerche [29]. It is a synthetic method – it enables one to construct both a problem and its solution by fairly simple means. In terms of the system (31)-(35) above, the idea is as follows: Pick any initial condition for $\tilde{g}^*$ that does not interfere with the condition (31) for $g^*$. That is, freely choose $\tilde{g}^*(0+, w)$ for $w > 0$, leaving $\tilde{g}^*(0+, w) = g^*(0+, w) = \delta(w)$ for $w \leq 0$. Solve the pde (32) with no additional conditions. Given the solution $\tilde{g}^*$, use equation (34) to define $z^*(u)$ . Then, use equation (33) to determine $h(u)$. In this way we find explicit solutions to certain cases, albeit artfully constructed ones.

One fairly general form for $\tilde{g}^*(0+, w)$ is $-A(dw)$, where $A$ is a positive, $\sigma$-finite measure on the positive real numbers with

$$\int_0^\infty \phi(w) A(dw) < \infty.$$
The solution to (32) can then be written as
\[
\tilde{g}^*(u, w) = \frac{1}{\sqrt{u}} \phi \left( \frac{w}{\sqrt{u}} \right) - \int_0^\infty \frac{1}{\sqrt{u}} \phi \left( \frac{\xi - w}{\sqrt{u}} \right) A(d\xi).
\]

For each \( u > 0 \) separately, \( z^*(u) \) solves
\[
\tilde{g}^*(u, w) = 0.
\]

It turns out (see Lerche [29]) that there is a unique solution, and the resulting \( z^*(u) \) is concave and smooth (infinitely differentiable). Also, if the support of \( A \) starts at \( \eta \geq 0 \) (that is, \( \eta = \inf\{w : A(0, w) > 0\} \)), then \( z^*(0+) = \frac{1}{2} \eta \). So, the solution is consistent with (31). Then,
\[
h(u) = \Phi_0 \left( \frac{z^*(u)}{\sqrt{u}} \right) + \int_0^\infty \Phi_0 \left( \frac{\xi - z^*(u)}{\sqrt{u}} \right) A(d\xi),
\]
and
\[
h'(u) = \frac{1}{2u^{3/2}} \int_0^\infty \xi \phi \left( \frac{\xi - z^*(u)}{\sqrt{u}} \right) A(d\xi).
\]

At worst, these integrals can be computed numerically. In some cases, they are even easier.

**One point**

Suppose \( A \) is concentrated at a single point, \( \xi > 0 \). Specifically, \( A = a\delta(\xi - w) \), where \( a > 0 \). For this case \( z^*(u) \) solves
\[
0 = \tilde{g}^*(u, w) = \frac{1}{\sqrt{u}} \phi \left( \frac{w}{\sqrt{u}} \right) - a \frac{1}{\sqrt{u}} \phi \left( \frac{\xi - w}{\sqrt{u}} \right),
\]
or
\[
\ln(a)u = \frac{1}{2} [(\xi - w)^2 - w^2] = \frac{1}{2} \xi^2 - \xi w.
\]

Thus, setting \( \rho = -\ln(a)/\xi \),
\[
z^*(u) = \frac{1}{2} \xi + \rho u.
\]

Also,
\[
h(u) = \Phi_0 \left( \frac{\frac{1}{2} \xi + \rho u}{\sqrt{u}} \right) + a\Phi_0 \left( \frac{\frac{1}{2} \xi - \rho u}{\sqrt{u}} \right)
\]
and
\[
h'(u) = \frac{\xi}{2u^{3/2}} \phi \left( \frac{\frac{1}{2} \xi + \rho u}{\sqrt{u}} \right).
\]

This is an inverse Gaussian (or Wald, or Bachelier-Lévy) distribution. It is indeed the distribution of the first-passage time to the affine boundary \( z^*(u) \).

To summarize:
Proposition 5 For the $h$ in (36), the optimal policy is the affine function
\[ y^*(u) = \mu u + \sigma z^*(u) = \frac{1}{2} \xi \sigma + \left(\mu + \rho \sigma\right) u. \]

Clearly, by selecting the parameters correctly, we can recover any affine function with positive intercept.

From an inventory-theoretic viewpoint, it is surprising to learn that the cost functions $h$ of this form are the most basic, canonical ones, in the sense of leading to the simplest possible policies $y^*$. We cannot think of an intuitive story why this should be so. We don’t mean that these $h$ are odd or implausible, only that we can discern nothing obviously special about them. In contrast, for linear $h$, as in Figures 1-3, there appears to be no closed-form solution. These facts suggest that supply streams are indeed complicated systems. The interplay between their dynamics and economics leads to fairly subtle behavior. Apart from the transformation discussed above and a few examples presented in [22], we lack a solid understanding of the impacts of different forms of $h$.

This case certainly indicates the limits of the rule that safety stock should be proportional to the standard deviation of leadtime demand, namely $\sigma \sqrt{u}$. That rule, up to the normal approximation, correctly compares several single-stage systems. A single multi-stage system is different. The distribution of stock within it depends on the holding-cost structure as well as the demand parameters. For the particular structure (36), the safety stock grows linearly. (In the case $a = 1$, it is constant.) This echoes the point in Subsection 2.4 about monotonicity. Here, $y^*$ may increase (when $\mu + \rho \sigma > 0$), but not as $\sqrt{u}$.

Several points

Next, suppose $A$ is concentrated at several equally spaced points, specifically, $A = \sum_{j=1}^{J} a_j \delta(j \xi - w)$, where the $a_j > 0$. Here, $z^*(u)$ solves
\[ 0 = \tilde{g}^*(u, w) = \frac{1}{\sqrt{u}} \phi \left( \frac{w}{\sqrt{u}} \right) - \sum_{j} a_j \frac{1}{\sqrt{u}} \phi \left( \frac{j \xi - w}{\sqrt{u}} \right), \]

or
\[ \sum_{k=1}^{J} a_k \exp \left( -\frac{(k \xi)^2}{2u} - k \xi w \right) = 1. \]  
(37)

Denoting $\theta = \exp(\xi w/t)$, this is equivalent to
\[ \sum_{k=1}^{J} a_k \exp \left( -\frac{(k \xi)^2}{2u} \right) \theta^k - 1 = 0. \]
This polynomial has exactly one positive root for all \( u > 0 \). Then, set
\[
h(u) = \Phi_0 \left( \frac{z^*(u)}{\sqrt{u}} \right) + \sum_j a_j \Phi_0 \left( \frac{j\xi - z^*(u)}{\sqrt{u}} \right),
\]
\[
h'(u) = \frac{1}{2u^{3/2}} \sum_j a_j (j\xi) \phi \left( \frac{j\xi - z^*(u)}{\sqrt{u}} \right).
\]

This looks something like a mixture of inverse Gaussian distributions, but not quite, because \( z^*(u) \) is not linear, and it depends on all the \( a_j \).

The solution \( z^*(u) \) is not as simple as in the one-point model. However, we can obtain a fairly simple upper bound on it. For each \( j \), let \( \theta_j \) denote the positive solution of the polynomial equation
\[
\sum_{k=1}^j a_k \theta^k = 1.
\]

Set \( \rho_j = \ln(\theta_j)/\xi \) and
\[
\pi_j = \left( \frac{\xi}{2} \right) \frac{\sum_{k=1}^j k^2 a_k \theta_{jk}}{\sum_{k=1}^j k a_k \theta^k_{jk}},
\]
and define
\[
z^+(u) = \min_j \{\pi_j + \rho_j u\}.
\]

This is a piecewise-linear, concave function. Clearly, \( \theta_j \) and \( \rho_j \) are decreasing in \( j \), and \( \pi_j \) is increasing.

**Proposition 6** \( z^*(u) \leq z^+(u) \).

There are reasons to expect that \( z^+(u) \) approximates \( z^*(u) \) fairly well. For \( J = 1 \), they are identical. For \( J \geq 1 \), both are concave, and they are asymptotically equal for both small and large \( u \) (see [46]).

**Exponential weights**

Finally, suppose \( A \) has a density, that is, \( A(d\xi) = a(\xi)d\xi \). As discussed in Lerche [29] and Zipkin [46], there are several cases where the required calculations can be performed in closed form. One is \( a(\xi) = \alpha \beta e^{-\beta \xi} \) for positive constants \( \alpha \) and \( \beta \), a multiple of an exponential density.

One can obtain fairly simple expressions also for related densities \( a(\xi) \). For example, one can use a linear combination of exponentials with different values of \( \beta \). Or, one can use an Erlang density \( \beta(\beta \xi)^k e^{-\beta \xi}/k! \) for \( k > 0 \), or a linear combination of them. One can even use \( \beta e^{-\beta \xi} [1 + \gamma \sin(\delta \xi)] \) with \( |\gamma| \leq 1 \).

Besides the method of images, there are a few other explicit (but not simple) solutions to first-passage time problems. One such result is the distribution of \( \tau \) for a square-root boundary. The
formula is an infinite series involving parabolic cylinder functions. See Sato [38] and Novikov et al. [31]. So, notwithstanding the warnings above, there do exist cost functions \( h \) that lead to square-root optimal policies.

### 4.5 Numerical Methods

Apart from special cases like those above, one must use numerical methods to solve both original and inverse first-passage time problems. Approximations for the original problem are discussed by Daniels [12], and Song and Zipkin [42] develop one for the inverse problem.

Several exact methods are available for the original problem. Daniels [12] and Lo et al. [30] use the method of images within an approximation scheme. Durbin [18] develops a series representation. Wang and Pötzelberger [44] discretize \( u \) and approximate the boundary by a piecewise-linear function. This approach requires the calculation of certain expectations of a multivariate normal, which they perform by Monte-Carlo simulation. Novikov et al. [31] point out that these multivariate expectations can be expressed recursively as sequences of univariate expectations and performed by direct numerical integration. They also provide a detailed error analysis.

Zucca and Sacerdote [47] adapt the approach of Wang and Pötzelberger to the inverse problem. (They also develop an alternative method, based on one of the integral equations mentioned above.) With the recursive formulation of Novikov et al., this algorithm is similar in broad outline to (8) above. Of course, the details are different. Our (8) optimizes a cost function \( c^* \), while their algorithm works with the density \( g^* \). Also, they use a piecewise-linear approximation of the boundary, whereas we use a step function.

These observations suggest that (8) may be an effective method for the inverse first-passage time problem. (It is noteworthy that this algorithm, versions of which date to 1960, does solve the problem.) The main effort lies in the convolutions required to update \( c_k^*(u + \varepsilon, x) \) from \( \bar{c}_k^*(u, x) \). Such calculations are well suited for the fast Gauss transform, a technique which can speed them up considerably. (See Greengard and Strain [24] and Broadie and Yamamoto [3].) Also, the cost-minimization framework enjoys the advantage of monotone convergence, as in Proposition 1. Equivalently, one could solve (29), which also converges monotonically. (We have not attempted to compare this approach to the other available techniques, however.)

We should also mention that there are various methods for solving more general moving-boundary and stochastic-control problems. See, for example, Crank [10], Kushner and Dupuis [28], Fleming and Soner [20], and Kumar and Muthuraman [27].

### References


