When Gray Markets Have Silver Linings: All-Unit Discounts, Gray Markets and Channel Management

Ming Hu, Michael Pavlin, Mengze Shi
Rotman School of Management, University of Toronto, Toronto, Ontario, Canada M5S 3E6
{ming.hu@rotman.utoronto.ca, michael.pavlin06@rotman.utoronto.ca, mshi@rotman.utoronto.ca}

Gray markets are unauthorized channels of distribution for a supplier’s authentic products. This paper studies a distribution channel that consists of a supplier who offers all-unit quantity discounts for batch orders to enjoy cost savings, and a reseller who may divert some goods to the gray markets. Our analysis shows the impact of gray markets depends on the reseller’s batch inventory holding cost. When the reseller’s batch inventory holding cost is high, diversion to the gray markets improves the channel performance by enabling the reseller to make batch orders. Since the reseller’s order costs decrease through quantity discounts, diversion to the gray markets reduces the resale price and expands sales to the authorized channel. On the other hand, when the reseller’s batch inventory holding cost is low, the reseller would make the batch orders even without the gray markets. In this case the diversion to the gray markets may improve the reseller’s performance by shortening the order cycles and reducing the inventory holding costs. Interestingly, since diversion to the gray markets decreases the reseller’s cycle inventory volume, the reseller has the reduced incentive to push its inventory and, consequently, the resale price rises and sales volume decreases in the authorized channel. These results are shown to be robust when gray market overlaps with the authorized channel or when the gray market price is sensitive to reseller’s diversion rate.

Key words: gray markets, channel management, inventory, quantity discounts, supplier pricing

1. Introduction

The diversion of branded goods to unauthorized channels, also known as gray markets, is of substantial strategic interest to manufacturers. Industry reports show that gray market channels account for a significant portion of markets in a broad set of industries ranging from pharmaceuticals to consumer electronics. According to a KPMG study, the gray market accounted for as much as $58 billion in the 2008 global information technology (IT) hardware industry (KPMG LLP 2008). Gray market activity persisted in the IT hardware industry mainly because resellers exploited incentive programs offered by IT original equipment manufacturers (OEMs). The study shows that
90 percent of the OEM respondents offered incentives to their channel partners and customers. The two most popular incentive programs identified are promotional deals and quantity discounts, respectively adopted by 72 percent and 59 percent of the surveyed OEMs. The study indicates that channel partners, in order to obtain deeply discounted products, may deceive the OEMs into offering incentives on products for non-existent customers and then divert those products to the gray market. While a precise breakdown of the volume diverted through these means is not available, the survey identifies 24 percent of resellers as admitting to selling to unauthorized channels and gray market brokers.

The focus of this paper is on the operational and marketing issues associated with the gray market diversions induced by quantity discounts. Suppliers often implement quantity discounts with the intent of promoting operational goals (Munson and Rosenblatt 1998). Quantity discounts, in particular all-unit discounts, provide an incentive for the reseller to order in a manner consistent with efficient shipping and manufacturing. From this operational perspective, gray market and quantity discounts can be beneficial by allowing the supply chain to achieve economies of scale. The reseller can order up to an efficient threshold and then use the gray market to divert excess inventory. Our research goals are to understand how operational factors (e.g., inventory costs for the reseller and scale economies for the supplier) interact with the marketing problems (e.g., reseller pricing and supplier all-unit discounts) in markets where the resellers can divert to the gray market. We also investigate how the presence of the gray market affects the profits of the resellers and sales in the authorized channel.

The practice of offering all-unit discounts is consistent with both the operational perspective taken in this paper and the “situation on the ground”. Munson and Rosenblatt (1998) reported that economies of scale in production and transportation are the predominant drivers of quantity discounts. They also identified all-unit discounts as the most prevalent quantity discount offered by suppliers (utilized by 94 percent versus 34 percent for incremental discounts). An incremental discount applies only to the units above the volume-threshold. In contrast, an all-unit discount applies to each and every unit purchased as soon as the volume-threshold is achieved. Like Lal and
Staelin (1984), our model builds on the assumption that the supplier enjoys cost benefits when the resellers order in batches. Such cost savings can result from operational efficiency gained in batch handling and transportation. Our analysis indeed shows that only a schedule of all-unit discounts will lead resellers to order beyond the market demand to enjoy the quantity discount and divert the excess to the unauthorized gray market channel. In addition to the KPMG LLP (2008) survey, a wide variety of industrial literature has indicated reseller over-ordering in response to supplier’s discount pricing practices as a key driver for a persistent gray market (Lowe and McCrohan 1988, Jorgenson 1999, Gilroy 2004, Eban 2005). An early example which is commonly cited took place in the personal computer market when resellers’ responses to quantity discount became a concern for IBM (Ramirez 1985). The following quotation from Gilroy (2004) on the subject of car audio equipment suppliers illustrates that this behavior is not limited to the IT sector.

*Suppliers place an enormous pressure on resellers to make product quotas in order to receive volume discounts. Sometimes resellers dump overstock on the Internet to make their quotas, admitted suppliers. “A lot of times [dealers] say that the pressure placed on them by the supplier is very great”.*

Despite its prevalence in practice, to the best of our knowledge, the quantity-discount induced gray market has not been rigorously studied in the literature. Our paper addresses this gap by performing an economic analysis of a rich operational model with a reseller responding to a supplier’s all-unit quantity discount offerings. We take the premise that a supplier manages the gray market through tolerance of violation (Dutta et al. 1994, Bergen et al. 1998, Antia et al. 2004). In other words, the supplier chooses not to pursue enforcement through monitoring and legal action. Instead, the supplier anticipates the reseller’s access to the gray market and formulates the pricing strategy accordingly. Both the supplier and the reseller anticipate that the gray market, unlike other salvage channels, can cannibalize their sales in the authorized channels. The tolerance of violation is consistent with the estimate by KPMG LLP (2003, 2008) that, from 2003 to 2008, the IT gray market experienced a 45 percent increase from $40 billion to $58 billion. Even in ostensibly well-regulated industries such as pharmaceuticals, abuse of promotions has proved difficult
to counteract. Eban (2005) reports that up to four fifths of American nursing homes and similar healthcare institutions take advantage of wholesaler discounting practices to profit off sales of prescription medications to gray market channels. With this in mind, it is not surprising that in the IT sector, 42 percent of OEMs still have no process to identify or monitor gray market activity (KPMG LLP 2003).

Gray market supply may also result from excess stock due to demand uncertainty (Ahmadi et al. 2010, Altug and van Ryzin 2009, Xia and Bassok 2005). When realized demand is below expectations, the gray market provides a channel for resellers to dispose of the overstock. This driver is not incompatible with diversion resulting from all-unit discounts discussed in this paper. In practice we expect both drivers to contribute to facilitating the gray market. The setup leading to gray market diversion driven by demand uncertainty can be modeled by a newsvendor model and is particularly appropriate when goods are perishable and demand is periodic. This setup lends itself to seasonal and fashion goods. Both drivers may contribute to domestic gray market supplies as well as parallel imports. In the latter case, the drivers will act to reinforce price differences between countries leading to arbitrage opportunities resulting in parallel imports (Duhan and Sheffet 1988, Ahmadi and Yang 2000). Authorized resellers operating in the same channel and/or location as the gray market will compete with the gray market for customers. The extent of cannibalization will depend on the price sensitivity of customers, the degree of differentiation between the gray market and authorized market good (e.g. the gray market often lacks a warranty), and the trust in the gray market (e.g. the customer may worry that the product is counterfeit). Cannibalization with authorized market demand is what differentiates the gray market from a typical model of a salvage channel where the primary and secondary markets are often treated independently. Finally, while both the domestic gray market and parallel imports deal with authentic products sold through unauthorized channels, counterfeit products do not originate from the trademark owner and therefore are not authentic (Duhan and Sheffet 1988).
1.1. Contributions to the Literature

Our model and analysis integrate the operational and marketing decisions. First, our model considers a reseller optimizing over lot-sizing and resale price decisions when facing a gray market and all-unit quantity discounts. The closed-form analysis of the reseller’s dynamic lot-sizing problem yields a novel solution that links the cost of holding inventory to the supply of goods to the gray market. Specifically, we find gray market diversion occurs only in a middle range of the batch inventory holding cost. Within this range, the reseller finds it beneficial to use the gray markets to reduce its inventory holding costs. The gray market may allow the reseller to improve operational efficiency while enjoying the batch quantity discounts. Second, we examine the impact of gray market diversion on the resale price and sales in the authorized channel. Interestingly, the effect depends on the reseller’s batch inventory holding costs. When the batch inventory holding cost is sufficiently high such that the reseller would not order in batches without the gray market, diversion allows the reseller to enjoy the quantity discount and in turn reduce the resale price. As a result, the presence of gray markets expands sales in the authorized channel. However, when the batch inventory holding cost is low enough that the reseller would order in batches even without the gray market, the diversion reduces the reseller’s peak cycle inventory and expedites the ordering cycles. As the reseller faces reduced pressure to push out the inventory, the resale price increases and sales decrease in the authorized channel. Third, we study the effect of gray market diversion on the performance of the distribution channel. When gray market diversion enables the reseller to take advantage of the quantity discount, the supplier enjoys increased operational efficiency and total channel performance improves. Due to the Stackelberg structure of our model, the supplier captures all or most of the efficiency gains.

Our paper contributes to both the marketing and operations management (OM) literature. The marketing literature typically neglects inventory costs and limits attention to single-period models (Howell et al. 1986, Wilcox et al. 1987, Banerji 1990). These restrictions on the analysis do not permit operational characteristics, in particular inventory holding costs, to influence the reseller’s strategies or the diversion to a gray market. Though Lal and Staelin (1984) indeed consider the
effect of the reseller’s holding cost on the supplier’s quantity discount design in an economic-order-quantity (EOQ) setting, the authors assume that all products are sold to the authorized channel at an exogenous resale price. In this paper we incorporate the reseller’s operational decisions and derive the resale price as a function of the operational cost in the presence of a gray market. As a result, the supplier can take a reseller’s operating environment into account when making the channel management decisions.

The OM literature tends to emphasize the algorithmic issues determining optimal lot sizes (reseller ordering policy) in response to offered discounts due to the lack of tractability of the general multi-period problem. An overview of the area is covered by a pair of surveys from Benton and Park (1996) and Munson and Rosenblatt (1998). Despite an impressive breadth of work including many extensions of the lot-sizing problem under all-unit discounts, gray markets and, more generally, the ability to salvage surplus inventory has been discussed in only a couple of instances. Sethi (1984) and Arcelus and Rowcroft (1992) have examined the optimal lot-sizing problem with an all-unit discount and a fixed value for salvaged inventory. They both develop algorithms to numerically solve the reseller’s lot-sizing problem. This paper contributes to this literature by developing an explicit solution to a representative dynamic lot-sizing problem where the salvage channel is replaced by a gray market which may cannibalize some portion of authorized channel demand. The stylized elements in our model include no fixed order costs and deterministic demand. Without fixed order costs, the reseller is able to order as demand arrives and eliminate holding costs. Adding a fixed order cost would replace this strategy with a less profitable EOQ style policy and lead resellers to be more likely to use batch strategies, possibly including gray market diversion. Deterministic demand allows for better elucidation of the interaction of the gray market with the suppliers discount policy. More importantly, the closed-form solution allows optimal analysis of the reseller’s pricing and inventory decisions, and the resulting profit implications to the supplier and welfare consequences to the consumers.

We organize the rest of the paper as follows. In §2 we describe the main model, followed by a preliminary analysis to demonstrate the quantity discount-induced gray market. In §3 we follow
the conventional operational approach and assume an exogenously determined resale price. In this section we focus on the effect of a gray market on the reseller’s operational decisions and the subsequent effect on the supplier’s profit. In §4 we incorporate the reseller’s pricing decision and investigate the interaction between operational and marketing decisions in managing a channel with a gray market. We then extend the model in §5 to allow for an endogenous gray market wholesale price which responds to the reseller’s supply. Finally, we conclude with a summary of results and managerial implications in §6. All proofs can be found in the Online Appendix.

2. Model and Preliminary Analysis

We consider a market where a monopoly supplier sells its products to end consumers through a single reseller. The reseller can sell the goods through an authorized channel and a gray market. We consider a Stackelberg game: the supplier first sets an all-unit discount schedule. Given the discount policy and gray market condition, the reseller responds with decisions on ordering, inventory holding, gray market diversion, and possibly resale pricing based on deterministic demand and a constant lead time (see Lal and Staelin 1984 for a similar setting). For simplicity, we assume there is no fixed order costs because the all-unit discount has already incorporated an incentive for batch ordering to enjoy economies of scale. The obtained insights from the main model on gray market diversion remain when there is a positive fixed order cost.

Profits for both reseller and supplier will be considered over an infinite horizon with long-run average criterion. We will verify later that with diversion, EOQ-type cyclic policies can be the optimal inventory policies even without fixed order costs. This EOQ type of lot-sizing policy remains important for products with low demand seasonality and moderate life cycles. These are characteristics of the prescription drug and audio electronics markets cited in the introduction. The IT products cited also fall under this umbrella. The subjects of the KPMG study are mainly commoditized IT products such as hard drives and network routers. Such products are sold over very long timelines through periodic incremental specification improvements (e.g. a 1GB hard drive is replaced by a 1.5GB hard drive at the same price point).
Following the convention in the OM literature, we represent the supplier’s lot size-based all-unit quantity discount by the reseller’s order cost function, denoted by \( C(q) \) where \( q \) is the order size, as follows:

\[
C(q) = \begin{cases} 
  w_o q & \text{if } 0 \leq q < \eta, \\
  w_\eta q & \text{if } q \geq \eta,
\end{cases}
\]

where \( w_o \) is the list price before the discount, \( w_\eta \) is the discounted price and \( \eta \) is the threshold order quantity defining the change in the unit cost. For simplicity, we assume there is only one price breakpoint as it is sufficient to demonstrate the incentive for overbuying and gray-market diversion. If \( w_o = w_\eta \), the discount schedule reduces to the trivial case of no quantity discount. We assume the list price \( w_o \) is determined before the supplier optimizes the quantity discount by setting \( w_\eta \). The list price may be determined either exogenously\(^1\) or by an optimization on top of the Stackelberg game. As we have mentioned, all-unit discounts are the most popular type of quantity discounts and provide a unique incentive to order up to the threshold (Benton and Park 1996). In practice, all-unit discounts shift the batch-breaking decision from the supplier to the reseller, who shares the cost savings through the quantity discount.\(^2\) These discounts are most commonly motivated by operational advantages (Munson and Rosenblatt 1998) and are typically associated with an exogenous batch size. For instance, this batch size may correspond to a pallet, production lot size or the capacity of a truck. In this vein, we assume that the batch size \( \eta \) is exogenously determined.

We assume there exists economies of scale in production and distribution when the reseller orders in batches. Specifically, we let \( c_o \) represent the per unit supply cost when the reseller orders one unit of the product each time, and \( c_\eta (c_o) \) be the per unit supply cost when the reseller orders in batches of \( \eta \).

After the supplier announces the wholesale pricing policy, the reseller makes lot-sizing and, possibly, resale pricing decision \( (p) \). The reseller in our model may be viewed as either a retailer

\(^1\) Fixing the regular wholesale price is not uncommon in studies of quantity discount practices, e.g., Lal and Staelin (1984).

\(^2\) In addition, this interpretation is consistent with the Robinson-Patman act which forbids discriminatory pricing in the US through a quantity discount unless justified by underlying costs (Coughlan et al. 2001).
selling directly to consumers or an intermediate distributor selling to authorized retailers. At any
time \( t \), the reseller’s lot-sizing decisions include ordering \( q(t) \) units from the supplier at a cost of
\( C(q(t)) \) and diverting \( g(t) \) units into the gray market at a per unit diversion wholesale price \( s \). We
assume away demand uncertainty and allow replenishment to be instantaneous which is equivalent
to assuming a deterministic lead time. We denote the reseller’s inventory level at time \( t \) by \( I(t) \),
which is the sum of all orders minus all sales through the authorized channel and the gray market
up to time \( t \). For each unit of goods in inventory we let the holding cost be \( h \) per unit of goods per
unit of time.

The reseller can sell through the authorized channel and gray market. We focus on all-unit
discount induced gray market diversion and assume that demand is deterministic in both markets.
Specifically, in the authorized channel, the market demand (or order) arrives continuously with a
deterministic rate determined by a modified iso-elastic demand structure that takes into account
the cannibalization between the authorized channel and gray market:

\[
\lambda(p, p_s) = m/(p - \gamma p_s)^\alpha,
\]

where \( p_s \) is the gray market resale price that is offered to consumers for the gray market product.
We assume \( s < p_s \). We do not model the detailed gray market pricing mechanism, as emergence and
pricing of the gray market can be attributed to many factors. The gray market prices are treated as
exogenously determined. This is likely the case also when there are a large number of resellers and
one reseller’s diversion contributes only a tiny proportion of the entire supply to the gray market.
In §5, we will examine the robustness of our results by extending the model to scenarios where
the gray market prices are negatively affected by the focal reseller’s diversion volume and a fixed
mark-up mechanism relates the gray market wholesale price \( s \) and resale price \( p_s \). The parameter \( m \)
denotes the size of the market, \( \gamma \) parameterizes the sensitivity of the authorized channel’s demand
to the gray market resale price and \(-\alpha \) measures the demand elasticity to the adjusted market
price difference. The higher the value of \( \gamma \), the more sensitive the authorized channel’s demand to
changes in the gray market resale price. In addition to measuring demand sensitivity to changes
in $p_s$, $\gamma$ has a secondary and counterintuitive effect that an increase in this parameter increases demand. Thus, to study changes in cannibalization comparative statics should be taken on $p_s$ rather than on $\gamma$.

The degree of cannibalization depends on the structure of the gray market, for instance, whether these goods are sold in the same geography as the reseller. If the quantity discount is providing impetus for parallel importation into a separate marketplace $\gamma$ will be small due to minimal cannibalization. Also, $\gamma$ will be small when goods are well differentiated from their gray market counterparts. For instance, ancillary services such as warranties are lacking in the gray market channel. In the markets where counterfeiting is prevalent, the consumer is concerned that the gray market product is not genuine.

Furthermore, the gray market wholesale price is assumed to be below the batch supply cost, i.e. $s < c_n$, to eliminate the suppliers indirect arbitrage opportunity. This is a common assumption in the OM literature, for example, similar assumptions are made in the newsvendor problem. The conventional OM literature often considers an exogenous resale price and focuses on cost-minimizing lot-sizing decisions. We will follow this approach in §3 and assume that the resale price $p$ is fixed and that the reseller adjusts only his ordering and diversion behavior to minimize costs. We then incorporate the reseller’s pricing decisions in §4 but assume $\alpha = 2$ that lends tractability of obtaining closed form solutions. We will check the robustness of obtained insights for general values of $\alpha$ by numerical experiments. We summarize all notation in Table 1.

Finally, we summarize the sequence of events in Figure 1: First, the supplier sets the discount price $w_n$. After observing the all-unit discount, the reseller chooses the order size, inventory level and diversion volume. In the endogenous resale price model (§4), the reseller also selects the resale price. We assume that the supplier has perfect knowledge of the reseller’s holding costs as well as the demand structure of the authorized and gray markets. Therefore, the supplier can anticipate the reseller’s order, diversion and pricing decisions, and manage the authorized channel and possible diversion through the offering of the wholesale price schedule.
Table 1 Description of Principal Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\eta$</td>
<td>Order quantity discount threshold</td>
</tr>
<tr>
<td>$w_o$</td>
<td>List wholesale price per unit</td>
</tr>
<tr>
<td>$w_\eta$</td>
<td>Discounted wholesale price per unit</td>
</tr>
<tr>
<td>$c_o$</td>
<td>Supply cost per unit without economies of scale</td>
</tr>
<tr>
<td>$c_\eta$</td>
<td>Supply cost per unit with economies of scale</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Demand rate of the authorized channel</td>
</tr>
<tr>
<td>$p$</td>
<td>Resale price</td>
</tr>
<tr>
<td>$s$</td>
<td>Gray market wholesale price</td>
</tr>
<tr>
<td>$p_s$</td>
<td>Gray market resale price</td>
</tr>
<tr>
<td>$m$</td>
<td>Market size of the authorized channel</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Price elasticity</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Demand sensitivity of the authorized channel to the gray market resale price</td>
</tr>
<tr>
<td>$h$</td>
<td>Reseller’s inventory holding cost per unit per unit time</td>
</tr>
<tr>
<td>$I$</td>
<td>Reseller’s cycle inventory level</td>
</tr>
<tr>
<td>$I^o$</td>
<td>Reseller’s cycle inventory level in the diversion strategy</td>
</tr>
<tr>
<td>$H$</td>
<td>Reseller’s average inventory holding cost per unit per cycle when $I - \eta$ and $\lambda - m$</td>
</tr>
<tr>
<td>$G$</td>
<td>Gray market diversion per cycle</td>
</tr>
<tr>
<td>$g$</td>
<td>Gray market diversion per unit of time</td>
</tr>
<tr>
<td>$\Pi$</td>
<td>Supplier’s profits</td>
</tr>
<tr>
<td>$\pi$</td>
<td>Reseller’s profits</td>
</tr>
</tbody>
</table>

2.1. Preliminary Analysis: Quantity-Discount Induced Gray Market Diversion

We first analyze a simple one-period model to investigate the possible incentive for overbuying and diverting to the gray market under the all-unit quantity discount. Such a static model can be considered as a snapshot of our dynamic problem. Suppose the reseller orders $q$ units to fulfill demand $d$ in the authorized channel and diverts the remaining $g = q - d$ units to the gray market. Expressed in terms of $d$ and $q$ the reseller’s total effective cost associated with such a policy is $c(q, d) = C(q) - s(q - d)$.

If the resale price is exogenous, then maximizing the reseller’s profit becomes equivalent to minimizing the reseller’s cost. We calculate the optimal order quantity as a function of replenishment need $d$, $q^*(d) := \max\{\arg\min_{q\geq d} c(q, d)\}$. We let $c^*(d) := c(q^*(d), d)$ and $g^*(d) := q^*(d) - d$ denote the reseller’s optimal order cost and the optimal size of diversion to the gray market, respectively. We summarize the reseller’s optimal order and diversion strategy in Proposition 1.
Proposition 1. (Optimal One-Shot Order and Diversion Strategy). Under the all-unit discount $C(q)$, the optimal ordering and diversion strategy of the reseller with a one-shot demand size $d \geq 0$ is

$$q^*(d) = \begin{cases} \eta & \text{if } \hat{q} \leq d < \eta, \\ d & \text{otherwise,} \end{cases}$$

where $\hat{q} = \frac{w_\eta - s}{w_o - s} \eta < \eta$ is the threshold demand size above which the reseller orders up to the discount break point $\eta$. The optimal effective order cost is a continuous function

$$c^*(d) = \begin{cases} w_\eta d & \text{if } 0 \leq d < \hat{q}, \\ (w_\eta - s)\eta + sd & \text{if } \hat{q} \leq d < \eta, \\ w_\eta d & \text{otherwise.} \end{cases}$$

The diversion to the gray market is

$$G^*(d) = \begin{cases} \eta - d & \text{if } \hat{q} \leq d < \eta, \\ 0 & \text{otherwise.} \end{cases}$$
Remark 1. Incremental discounts and two-part tariffs do not generate incentive for overbuying and diversion to the gray market. See the Online Appendix for proofs.

The diversion to the gray market occurs only in the middle range of demand defined by $\hat{q} \leq d < \eta$. In this range, the benefit from receiving the quantity discount outweighs the loss from diverting the excess purchases to the gray market. If $d = \hat{q}$, the reseller is indifferent between 1) making an order $\hat{q}$ at the list price $w_o$ and 2) ordering up to $\eta$ at the discounted unit price $w_\eta$ and then diverting $\eta - \hat{q}$ units to the gray market; for consistency, we designate the reseller’s behavior when $d = \hat{q}$ to be the latter case of ordering and diversion. We illustrate this optimal order cost $c^*(d)$ in Figure 2. The optimal order cost function is a continuous piecewise linear function with three segments with respective slopes of $w_o$, $s$, and $w_\eta$. The reseller contributes to the gray market only over the middle segment where $\hat{q} \leq d < \eta$. In this case, the reseller orders $\eta$ units regardless of the size of demand; an incremental demand means a unit reduction of the reseller’s supply to the gray market, and hence an increase in the reseller’s total cost by $s$.

The above analysis has clearly demonstrated the cause and consequence of a quantity discount-induced gray market at the reseller’s level. Next we analyze the full model where both the reseller and supplier consider their decisions over an infinite horizon.

3. Model Analysis: Exogenous Resale Price

In this section we study the case where the resale price $p$ is exogenous. As a result, the reseller faces an exogenous and deterministic demand $\lambda = \frac{m}{(p - \gamma p_\lambda)\alpha}$ per unit of time, where the gray market
wholesale and resale prices \((s,p_s)\) are exogenously determined. We will solve the Stackelberg game backward by first analyzing the reseller’s optimal inventory decisions in response to the supplier’s quantity discount schedule, and then examining the supplier’s optimal quantity discount.

3.1. Reseller’s Inventory Policy

With an exogenous resale price, the reseller’s profit-maximization problem can be solved by minimizing the total of the reseller’s order cost and inventory holding cost. In the following lemma, we prove that the optimal policy is of stationary type among all dynamic policies. The stationary inventory policy includes a target inventory level \(S\), an order size \(q\), a diversion amount \(g\) to the gray market, and the timing of orders and diversion. Our analysis shows that it is optimal for the reseller to follow the zero-inventory policy described in Lemma 1.

**Lemma 1. (Reseller’s Zero-Inventory Policy).** The reseller’s optimal inventory policy consists of a target cycle inventory level \(I\). The reseller orders \(q^*(I)\) and diverts \(G^*(I)\) to the gray market at precisely the times when the inventory drops to zero.

The zero-inventory property in Lemma 1 is typically associated with EOQ models with constant demand and fixed order cost (see Zipkin 2000, Simchi-Levi et al. 2005). Under such an inventory policy, a reseller who diverts to the gray market will repeatedly place orders of volume \(\eta\) whenever the inventory level hits zero, immediately divert a quantity of \(G\) to the gray market, and serves the authorized channel from inventory afterwards until the inventory level reaches zero again. In the optimal inventory policy, the inventory level reaches a peak of \(I = \eta - G\) at the beginning of each order cycle. Since the demand arrives at a constant rate \(\lambda\), the length of each cycle is \(I/\lambda = (\eta - G)(p - \gamma p_s)/\alpha/m\). For simplicity of notation, we define the average unit holding cost per cycle for a full batch order \(I = \eta\) without diversion when the demand rate \(\lambda = m\) as follows:

\[
H := \frac{h\eta}{2m},
\]

where \(h\) is the unit holding cost for each unit of time. As an aggregate measurement, the holding cost \(H\) contains information not only about \(h\) but also about the cycle length, and is a more relevant measure of inventory costs in this paper.
Proposition 2. (Reseller’s Optimal Inventory Policy). The optimal inventory policy of the reseller is a zero-inventory policy with the cycle inventory defined as follows:

\[ I^* = \begin{cases} \eta & \text{if } H \leq \frac{w_o - s}{(p - \gamma p_s)^\alpha}, \\ I^o := \eta \frac{w_o - s}{H(p - \gamma p_s)^\alpha} & \frac{w_o - s}{(p - \gamma p_s)^\alpha} < H \leq \frac{(w_o - s)^2}{4(w_o - s)(p - \gamma p_s)^\alpha}, \\ 0 & \text{otherwise}; \end{cases} \]

if \( s < w_o < (w_o + s)/2 \),

\[ I^* = \begin{cases} \eta & \text{if } H \leq \frac{w_o - w_o}{(p - \gamma p_s)^\alpha}, \\ 0 & \text{otherwise}. \end{cases} \]

if \( (w_o + s)/2 \leq w_o \leq w_o \),

We define formally three inventory policies obtained in Proposition 2 and later may refer to them in their short names.

Definition 1 (Order-as-You-Go Strategy). The reseller does not carry any inventory \( (I = 0 \text{ in short}) \).

Definition 2 (Batch Strategy). At the beginning of each cycle the reseller orders \( \eta \) and does not divert any quantity to the gray market \( (I = \eta \text{ in short}) \).

Definition 3 (Diversion Strategy). At the beginning of each cycle the reseller orders \( \eta \), diverts \( \eta - I^o \) into the gray market immediately and does not divert any order to the gray market within the cycle \( (I = I^o \text{ in short}) \).

Proposition 2 shows that if the holding cost is sufficiently high, the reseller follows the order-as-you-go strategy; and if the holding cost is sufficiently low, the reseller follows the batch strategy. As the reseller increases the order size to \( \eta \) to enjoy the discount, it may incur an additional unit holding cost or a negative profit margin for those units diverted to the gray market. Only when the quantity discount is attractive enough and the holding cost is in an intermediate range, does the reseller adopt the diversion strategy. The first necessary condition for an optimal diversion strategy is \( w_o - s < w_o - w_o \), namely, the benefit from the quantity discount must outweigh the margin loss on diverting a unit. This is essential to ensure a window of profit opportunity for the reseller. Moreover, in order for the reseller to be better off diverting goods to the gray market
rather than keeping all products in inventory, the immediate diversion loss \( w_{\eta} - s \) needs to be smaller than the average unit holding cost \( H(p - \gamma p_s)^\alpha \) the reseller would have to incur in a cycle without diversion. Finally, in order for the reseller to be better off adopting the diversion strategy than the order-as-you-go strategy, costs associated with the optimal diversion strategy should be lower than those under the order-as-you-go strategy. This condition leads to an upper bound on the holding cost in order for the diversion strategy to be optimal.

The gray market has effects on both supply and demand in the authorized channel. On the supply side, the existence of the gray market provides an opportunity for the reseller to buy in excess to enjoy economies of scale under an all-unit discount. We call this the “diversion effect”. On the demand side, the gray market poses a threat to the authorized channel, possibly cannibalizing its demand. We call this the “cannibalization effect”. Under the understanding that the gray market wholesale and resale prices are positively correlated, a higher gray market wholesale price will have two implications. First, as the per unit diversion loss becomes smaller, the diversion effect increases the reseller’s incentive to resort to the gray market. Second, the cannibalization effect makes the gray market purchase less attractive and increases demand in the authorized channel which reduces the reseller’s incentive to divert at a loss. These two effects clearly work in opposition. Depending
on the relative magnitude of each effect, an increase in the gray market prices may either reinforce or reduce the reseller’s incentive for gray market diversion.

We illustrate the key results of Proposition 2 with Figure 3. The figures show the parameter spaces under which each of these three inventory policies are optimal. We assume that the gray market resale price is marked up by 40% over the gray market wholesale price. First, we let \( \alpha = 0 \) which eliminates the cannibalization effect and use Figure 3(a) to illustrate the diversion effect. When the gray market wholesale price increases from \( s_1 = 2 \) to \( s_2 = 3 \), the boundary lines for the diversion strategy versus batch and order-as-you-go strategies both shift towards the right, indicating that the diversion strategy becomes more attractive. At sufficiently low values of \( H \), the boundary between diversion and batch strategies shifts in parallel from \( H = w_\eta - s_1 \) to \( H = w_\eta - s_2 \). The size of this shift is equal to the increase in the gray market wholesale price. In this case, the gray market provides an opportunity for the reseller to reduce cycle inventory in addition to enjoying the discounted price. Such a benefit is independent of the holding cost within the range of low values of \( H \). At sufficiently high values of \( H \), the boundary between diversion and order-as-you-go strategies shifts disproportionally further to the right at higher values of \( H \). In this case, the gray market offers the reseller a chance to order additional units to enjoy the discount. This benefit to the reseller is larger when the holding cost is higher.

To illustrate the cannibalization effect, we increase the value of \( \alpha \) to 2 in Figure 3(b). Clearly the diversion effect still exists: as in Figure 3(a), as the gray market wholesale price increases from 2 to 3, the boundary lines shifts rightwards and a diversion strategy is more likely to be optimal. In Figure 3(b), positive values for parameters \( \alpha \) and \( \gamma \) bring about the cannibalization effect and consumer demand in the authorized channel decreases. As a result, in order to be optimal, the batch strategy requires a much smaller \( H \). (Note that, while the scale for \( w_\eta \) is the same in three figures, the scale for \( H \) is very different, with much smaller units in Figure 3(b) and 3(c).) Overall, as the value of \( \alpha \) increases from 0 to 2, the parametric spaces for both batch and diversion strategies become much smaller. Correspondingly, in Proposition 2, a positive \( \alpha \) simply increases the values
of denominators and hence reduces the size of intervals for holding cost $H$ associated with diversion and batch strategies.

Figure 3(b) also illustrates the interaction between the diversion and cannibalization effects on the reseller’s strategies. With holding costs ($H$) are low and $s$ is increased from 2 to 3, in contrast to Figure 3(a) where the boundary line between diversion and batch strategies shifts in parallel, here the slope of the boundary line increases. Thus, while a larger $s$ makes the diversion strategy more efficient, the effect is smaller with relatively larger values of $w_\eta$. This is because in our numerical setup, when $s$ increases, the gray market resale price $p_s$ also increases (with a fixed mark up of 40%). The increased $p_s$ reduces the cannibalization effect, increasing market demand in the authorized channel and making the batch strategy more profitable. The effect of reduced cannibalization is even stronger when $\alpha$ increases to 4 in Figure 3(c): the cannibalization effect eclipses the diversion effect when the value of $w_\eta$ is close to 6. At high holding costs the effects differ. When $H$ is large, the border between diversion and order-as-you-go strategies moves upward with a larger $s$ due to the diversion effect. The reduced cannibalization effect from a higher $p_s$ pushes the border further upwards, making the diversion strategy even more attractive. Figure 3(c) illustrates that the reduction in the cannibalization effect is stronger at a larger value of $\alpha$.

3.2. Supplier’s Discount Policy

When deciding on the discounted price, the supplier needs to anticipate the best response from the reseller. Based on Proposition 2, we can derive the supplier’s profit function and optimize its all-unit discount policy. We summarize the results as follows.

**Proposition 3. (Supplier’s Optimal Discount with Exogenous Resale Price).** Given that the reseller employs the optimal inventory policy, the profit-maximizing supplier’s optimal discounted wholesale price is:

$$w^*_\eta = \begin{cases} w_o - H(p - \gamma p_s)^\alpha & \text{if } H \leq \frac{w_o - s}{2(p - \gamma p_s)^\alpha}, \\ s + \frac{(w_o - s)^2}{4H(p - \gamma p_s)^\alpha} & \text{if } \frac{w_o - s}{2(p - \gamma p_s)^\alpha} < H \leq \frac{(2c_o - w_o - s)(w_o - s)}{4(c_\eta - s)(p - \gamma p_s)^\alpha}, \\ w_o & \text{otherwise;} \end{cases}$$
if \( c_o - c_\eta \leq (w_o - s)/2 \),

\[
w^*_\eta = \begin{cases} 
   w_o - H(p - \gamma p_s)^\alpha & \text{if } H \leq \frac{c_o - c_\eta}{(p - \gamma p_s)^\alpha}, \\
   w_o & \text{otherwise}.
\end{cases}
\]

Proposition 3 shows that the supplier’s decision to offer a quantity discount largely depends on the supplier’s benefit from economies of scale \( c_o - c_\eta \) and the reseller’s holding cost \( H \). When the benefit from economies of scale is sufficiently small and the holding cost is sufficiently large, the supplier does not offer a discount, i.e., \( w^*_\eta = w_o \). Inadequate economies of scale allow only a small window of profitable discounts, which is not enough to make the reseller hold inventory. On the other hand, a significant benefit from economies of scale may lead the supplier to provide a quantity discount. The size of the discount also depends on the inventory holding cost.

When the holding cost is low, the supplier offers a discount of \( H(p - \gamma p_s)^\alpha \) off the list wholesale price \( w_o \). The discount exactly accounts for the reseller’s incremental holding cost from ordering in batches without diversion. In this case, \( w^*_\eta = w_o - H(p - \gamma p_s)^\alpha > (w_o + s)/2 \), which is equivalent to \( H(p - \gamma p_s)^\alpha < w^*_\eta - s \). Thus, the supplier can induce batch orders with a discounted price \( w^*_\eta \) sufficiently high in comparison to the gray market wholesale price to discourage any diversion. The supplier enjoys all the net benefits for the entire channel, \( c_o - c_\eta - H(p - \gamma p_s)^\alpha \) per unit, resulting from economies of scale.

Only when the holding cost is in an intermediate range does gray market diversion occur. Proposition 3 clearly indicates the importance of considering the reseller’s operational parameters when investigating gray market diversion. Such an intermediate range exists when the condition \( c_o - c_\eta > (w_o - s)/2 \) holds. In this diversion range, the optimal discounted price \( w^*_\eta = s + (w_o - s)^2/[4H(p - \gamma p_s)^\alpha] \), which implies a discount size less than \( H(p - \gamma p_s)^\alpha \). This discount size perfectly offsets the reseller’s holding cost for a less-than-full-cycle-inventory as well as the loss incurred in gray market diversion. The supplier enjoys economies of scale by making it just incentive-compatible for the reseller to order in batches followed by an immediate diversion.

**Corollary 1. (Benefit Allocation under Exogenous Resale Price).** In the case of an exogenous resale price, when it is optimal for the supplier to offer an all-unit quantity discount
to the reseller to enjoy economies of scale, the supplier takes all the net benefits.

In Lal and Staelin (1984), the motivation behind the supplier’s quantity discount is increased channel efficiency resulting from economies of scale. The size of the discount is just to offset the reseller’s extra inventory holding costs. However, in this paper the possible channel efficiency from economies of scale is also affected by the gray market. When the gray market prices (both wholesale and resale prices) become more attractive, the diversion and cannibalization effects are complementary in helping the supplier to achieve economies of scale. First, the diversion loss becomes smaller. Second, the higher gray market resale price leads to less cannibalization and hence increases the authorized channel demand.

4. Model Analysis: Endogenous Resale Price

In this section we extend the previous analysis to allow the resale price to be an endogenous decision variable of the reseller. For tractability, we fix the elasticity by setting $\alpha = 2$, which leads to a demand rate function $\lambda(p) = m/(p - \gamma p_s)^2$, where the gray market resale price $p_s$ is exogenously given. Recall that we also assume $\gamma < 1$. As in the previous section, we will solve the Stackelberg game backward by first examining the reseller’s optimal pricing and inventory decisions in response to the supplier’s quantity discount schedule, and then solving the supplier’s optimal quantity discount.

4.1. Reseller’s Pricing and Inventory Policies

The reseller jointly determines the optimal resale price and inventory policy by balancing revenues with ordering, diversion and inventory costs. We solve for the optimal resale price and inventory policy given the supplier’s discount schedule and summarize the results in the following proposition.

**Proposition 4. (Reseller’s Optimal Pricing and Inventory Policy).** The optimal pricing and inventory policy for the reseller with the demand function $\lambda(p) = m/(p - \gamma p_s)^2$ is:

if $s < w_o < \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s}$,
The above proposition describes the impact of market and operational parameters on the resale price in the authorized market. The interesting case occurs when the reseller diverts a portion of its orders to the gray market. As in the case when the resale price is exogenous, only when the discounted wholesale price is sufficiently low and the inventory holding cost is moderate does diversion to the gray market occur. We illustrate this result with Figure 4, which depicts the parameter spaces under which diversion strategy becomes optimal. Overall the patterns look very similar to the case under exogenous resale price illustrated by Figure 3. Specifically, Figure 4(a) shows the diversion effect in the case without demand cannibalization ($\gamma = 0$). When the gray market wholesale price increases from $s = 2$ to $s = 4$, the diversion strategy becomes more attractive. All else equal, when the holding cost $H$ is small, there is a range of $w_\eta$ where the reseller’s optimal response switches from a batch to a diversion strategy in order to reduce the inventory cost. When the holding cost $H$ is large, there is a range of $w_\eta$ where the reseller’s optimal response changes from order-as-you-go to a diversion strategy in order to take advantage of the quantity discount. Figure 4(b) and (c) show the interaction of diversion and cannibalization effects. With positive $\gamma$ the diversion effect remains, such that when $s$ is increased from 2 to 4 the diversion strategy is more likely to outperform the order-as-you-go strategy at large $H$ and outperform the batch strategy at small $H$. In our examples, as $s$ increases the gray market resale price $p_s$ increases according to a 20% markup over the wholesale gray market wholesale price $s$. As a result, market demand in the
Figure 4  Reseller’s Best Responses with Endogenous Resale Price

Note. The solid (resp. dashed) line indicates policy boundaries, $s = 2$ (resp. $s = 4$). $m = 10, \eta = 20, w_\alpha = 10, \alpha = 2$, gray market resale markup= 20%.

Figure 5  Reseller’s Optimal Endogenous Resale Price

Note. The solid (resp. dashed) line indicates policy boundaries, $s = 0$ (resp. $s = 2$). $m = 10, \eta = 20, w_\alpha = 10, w_\eta = 2.5, \alpha = 2$, gray market resale markup= 20%.

authorized channel increases as cannibalization is reduced. This lowers inventory costs associated with higher inventory strategies. Consequently, at high $H$, the benefits to the diversion strategy are greater than to the order-as-you-go strategy. However, at low $H$, the cannibalization effect promotes the batch strategy at the expense of the diversion strategy.

It is useful to note that lowering the resale price and diverting to the gray market are two alternatives for the reseller to order enough to enjoy the quantity discount. Lowering the resale
price generates additional demand in the authorized channel. This yields a shorter reorder cycle which reduces cycle holding costs. Diversion to the gray market, by reducing the cycle inventory, decreases cycle length and holding costs of the remaining inventory. In solving its optimal resale price and inventory policy, the reseller balances between lowering the price to increase demand and diverting to the gray market to lower the cycle inventory. When a batch strategy is the optimal inventory policy \( I^* = \eta \), the reseller can sufficiently stimulate demand from the authorized channel to enjoy the quantity discount without resorting to the gray market. Adoption of the diversion strategy occurs when the difference between the discounted wholesale price and the gray market wholesale price is small resulting in correspondingly small diversion loss and an effective wholesale price sufficiently close to the discounted wholesale price. Alternatively, at higher \( H \), the diversion strategy may be used at more substantial diversion losses because attempting to achieve the same goal by lowering the price would result in even greater profit loss.

We use Figure 5 to illustrate the effect of a gray market on the optimal resale price. In Figure 5(a), without a gray market \( (s=0 \text{ and } \gamma = 0) \), the optimal resale price (solid line) is low when the firm enjoys the discounted wholesale price \( w_{\eta} \) due to low holding cost \( (H) \), and is high when the firm incurs high holding cost \( (H) \) and pays regular wholesale price \( w_{0} \). With a positive gray market wholesale price \( (s = 2) \) but still no demand cannibalization \( \gamma = 0 \), the resale price increases at medium-low range of \( H \) but decreases at medium-high range of \( H \). First, when the holding cost is relatively low, the reseller would order in batches regardless of the gray market wholesale price. An increase in the gray market wholesale price makes diversion more attractive which reduces the reseller’s inventory. Now that the reseller has a reduced incentive to use a lower price to attract consumers the resale price increases. As a consequence, sales in the authorized channel would decrease. Second, when the holding cost is relatively high, without the gray market the reseller would follow the order-as-you-go strategy. With an increased gray market wholesale price, the reseller would overbuy and benefit from the discount while diverting part of the order to the gray market. The reduction in order costs decreases the resale price which expands the market coverage in the authorized channel.
Figures 5(b) and 5(c) show the interaction of diversion effect in Figure 5(a) and demand cannibalization effect. With positive $\gamma$, a higher gray market resale price leads to a weaker cannibalization effect and higher demand in authorized channel. Given the assumed demand function, the price elasticity increases with the cannibalization parameter $\gamma$. Thus, as shown in Proposition 4, given the reseller’s inventory strategy, the optimal resale price decreases as $\gamma$ and/or $p_s$ increases. In our numerical example featuring a larger value of $\gamma$ and increased price elasticity, the optimal resale price in the authorized channel is always lower when the gray market resale price is increased from $s = 0$ to $s = 2$. The reduced resale price is expected to further expand the market coverage in the authorized channel.

**Corollary 2. (Optimal Bulk Inventory Strategy).** The optimal reseller bulk strategy is:

$$I^0(w_n) = \begin{cases} \eta & \text{if } H < \frac{w_n - s}{4(w_n - \gamma p_s)^2} \\ I^o & \text{otherwise} \end{cases}$$

where $I^o$ is the optimal reseller bulk inventory strategy for all $w_n$ if $H \geq 1/[16(s - \gamma p_s)]$.

The above corollary highlights the tradeoff between the batch and diversion strategies. It is consistent with Figures 5(a)-(c) which show the optimal bulk order strategy transitions from batch to diversion as holding costs increase. The corollary highlights that this tradeoff is independent of the wholesale price but depends strongly on the gray market prices. There are two contributing factors in this tradeoff. (i) As competition from the gray market increases ($p_s$ decreases), the diversion strategy becomes more valuable owing to the increased holding costs from a reduction in authorized channel sales. (ii) The diversion strategy is more valuable when the per unit cost from diversion decreases ($s$ increases). From the supplier’s point of view, this corollary implies that irrespective of $w_0$, if the holding costs are high, the supplier can only achieve economies of scale with prices which lead to diversion.

**4.2. Supplier’s Discount Policy**

In this section we examine the supplier’s optimal discount price, taking into consideration the reseller’s best response to the discount schedule. The supplier’s profit depends on the effective
wholesale price with or without discount, the reseller’s best-response resale pricing and inventory policies, and the effective supply cost with or without economies of scale. By Proposition 4, it is easy to derive the supplier’s profit function given the reseller’s best response. Then we can proceed to solve the supplier’s profit maximization problem. In addition to the previous assumptions \( s < c \) and \( \gamma < 1 \), we make an additional assumption \( w_o < 4s - 3\gamma p_s \) for tractability. This is not restrictive when the two markets are relatively differentiated (i.e. \( \gamma \) is not too close to 1).

**Proposition 5. (Supplier’s Optimal Discount with Endogenous Resale Price).** We assume \( w_o < 4s - 3\gamma p_s \). Given that the reseller employs the optimal inventory policy, the profit-maximizing supplier’s optimal discounted price is:

**case (i).** if \( 0 \leq 4H < \max\{0, 1/(2(c - \gamma p_s)) - 1/(w_o - \gamma p_s)\} \),

\[
\hat{w}_\eta = \begin{cases} 
2(c - \gamma p_s) & \text{if } (w_o - \gamma p_s)^2 \geq 4(c - \gamma p_s)(w_o - c_o), \\
\frac{w_o - c_o}{w_o} & \text{otherwise};
\end{cases}
\]

**case (ii).** if \( \max\{0, 1/(2(c - \gamma p_s)) - 1/(w_o - \gamma p_s)\} \leq 4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s) - 1/(w_o - \gamma p_s)} \),

\[
\hat{w}_\eta = \begin{cases} 
\frac{1}{2(c - \gamma p_s)} - \frac{1}{(w_o - \gamma p_s)} + \sqrt{\frac{1}{4(c - \gamma p_s)^2} - \frac{w_o - c_o}{(c - \gamma p_s)(w_o - \gamma p_s)^2}}, \\
\frac{w_o - c_o}{w_o} & \text{otherwise};
\end{cases}
\]

**case (iii).** if \( 4H \geq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s) - 1/(w_o - \gamma p_s)} \),

\[
\hat{w}_\eta = \begin{cases} 
\frac{1}{c - \gamma p_s} - s \left( \frac{(s - \gamma p_s)}{w_o - \gamma p_s} - \frac{(s - \gamma p_s)}{w_o - \gamma p_s} \right) \left( \frac{w_o - \gamma p_s}{(s - \gamma p_s)} - \frac{w_o - c_o}{w_o - \gamma p_s} - 1 \right), \\
\frac{w_o - c_o}{w_o} & \text{otherwise};
\end{cases}
\]

where \( \tilde{w} := 1/[4H + 1/(w_o - \gamma p_s)] + \gamma p_s \) and \( \hat{w} := s + \left( 1 - \sqrt{(s - \gamma p_s)/(w_o - \gamma p_s)} \right)^2/(4H) \).

If the holding cost \( H \) is relatively low, the supplier sets the optimal all-unit discount such that, in response, the reseller sets a low resale price to drive up demand and achieve the minimum batch threshold without diversion (see the first sub-cases in cases (i) and (ii)). However, if the holding cost \( H \) is high enough, the high inventory cost may lead to an optimal discount policy under which the reseller simultaneously resorts to gray market diversion and a lower resale price (see the first sub-case in case (iii)). Consistent with its exogenous resale price counterpart, Proposition 5 shows
that gray market diversion occurs if and only if the economies of scale are sufficiently large and the reseller’s inventory holding cost falls in an intermediate range of values. Unlike the case with exogenous resale prices, here the reseller can also use a lower resale price to increase demand and reduce inventory costs. If the supplier jointly optimizes the wholesale price \( w_0 \) as well as \( w_\eta \) the results will have a similar structure: if reseller holding costs are moderate or greater, the supplier will select between a diversion and zero inventory policy depending on their scale economies. \(^3\)

**Corollary 3. (Benefit Allocation under Endogenous Resale Price).** In the case of an endogenous resale price, when it is optimal for the supplier to offer an all-unit quantity discount to the reseller to enjoy economies of scale, the reseller shares either none or part of the net benefits of economies of scale. Consumers in the authorized channel are always better off with economies of scale than without.

Unlike the case of exogenous resale prices, when the resale price is endogenous the reseller may share part of the benefits from batch orders induced by the supplier’s optimal quantity discount. When resale prices are exogenous, the supplier’s optimal wholesale price discount just offsets the reseller’s increased inventory holding cost in the induced batch strategy. In the case of an endogenous resale price, the supplier can benefit from additional authorized market demand when the reseller uses the discounted wholesale price and lowers the resale price. Note that in a typical Stackelberg game with the supplier as the leader and the reseller as the follower facing the downward-sloping demand function \( \lambda(p) = m/(p - \gamma p_s)^2 \), the supplier’s optimal wholesale price with an effective supply cost \( c \) is \( 2c - \gamma p_s \). When \( 2c_\eta - \gamma p_s < w_o \), the supplier may enjoy a reduced supply cost \( c_\eta \) through batch orders and charge a discounted wholesale price \( 2c_\eta - \gamma p_s \). When \( 2c_\eta - \gamma p_s < w_o \) and the holding cost is sufficiently low as in case (i), the net benefit of charging the

\(^3\) The joint optimization problem can be solved easily if resellers are homogenous because only one of the prices, either the base wholesale or discounted price, will be active as the effective purchase price, and the other price is within certain range to ensure that purchases at the effective price are indeed optimal for the reseller. The standard rationale for offering multiple price levels is the presence of heterogeneous resellers with varying holding costs. When the supplier is facing a reseller network consisting of heterogeneous resellers, the problem is analytically much less tractable. As the two supplier prices are used to target different resellers, there is a complex interaction resulting from incentive constraints. However, our analytic modeling is appropriate when the base wholesale price is separately targeted to and optimized for one segment of small buyers who do not order in bulk.
discounted price at $2c_\eta - \gamma p_s$ is greater than the benefit of charging a wholesale price just offsetting the increased inventory holding cost. In this case, the benefit of increasing sales in the authorized channel dominates the additional subsidy sufficiently to offset the inventory holding cost. As a result, the benefits from the supplier’s economies of scale can trickle down the supply chain from the supplier to the reseller and to the end users.

The above reseller benefits do not carry through to case (ii) because $2c_\eta - \gamma p_s$ is no longer a feasible discounted price. When the holding cost is at the low end of the holding cost interval corresponding to this case, the supplier provides an optimal quantity discount which is just sufficient to compensate for the reseller’s additional inventory holding costs. The supplier anticipates that the reseller diverts no inventory to the gray market.

The most interesting situation is case (iii) where the reseller follows the diversion strategy if their holding costs are at the low end of the corresponding interval. In Figure 4 these moderate levels of $H$ are where the diversion strategy region borders the zero inventory strategy region. In this case, the reseller again shares none of the economic benefits resulting from batch orders. When the holding cost falls into the bottom end of the range corresponding to this case, the supplier provides the precise quantity discount sufficient to make up for the reseller’s higher inventory and diversion costs. Compared with the case where the supplier does not offer batch discounts, all else equal, when the channel members are able to order in batches, the resale price is always lower and consumers in the authorized channel are always better off.

5. Model Extension: Endogenous Gray Market Prices

So far we have assumed the gray market prices are exogeneous. This is reasonable when a large number of resellers contribute to the gray market such that no one reseller’s diversion decision will change gray market wholesale price. In this section we extend the model by allowing the gray market wholesale price to decrease with the amount of diversion from the focal reseller. This section serves to examine the extent to which the main insights generated in previous sections are robust to endogenous gray market prices. Due to analytical complexity, we limit this investigation to a computational exercise.
To allow the gray market wholesale price to change with the reseller’s diversion, we assume the gray market wholesale price obeys a decreasing linear function, \( S(g) = b - ag \). Parameter \( b \) represents the gray market wholesale price when the focal reseller does not divert. Parameter \( a \) measures sensitivity of the gray market wholesale price to the rate of diversion from the focal reseller. If \( a = 0 \), we revert to cases already analyzed in previous sections. Intuitively, with all else equal, the value of \( a \) should decrease with the size of a company’s resale network. Under an intensive resale network, the diversion from each reseller accounts for only a small proportion of the entire gray market supply and consequently the gray market wholesale price will be less sensitive to the diversion from an individual reseller. In this exercise, we further assume that the gray market employs a fixed percentage markup \( u \) in its sale to the end users. The markup \( u \) should be smaller for competitive markets with commodity-type products. After incorporating the above assumptions, the primary market demand in the authorized channel is equal to \( \lambda(p, s) = m/(p - (1 + u)\gamma S(g))^{\eta} \). Finally, we assume that the cannibalization parameter \( \gamma \) is sufficiently small so that peak inventory decreases with gray market wholesale price \( s \). That is, \( I^\circ_s < 0 \). Then, with a lower gray market wholesale price, the reseller has less incentive to divert and consequently holds a larger peak inventory in equilibrium.

An endogenous gray market wholesale price leads to substantial additional complexity in the analysis. When the reseller diverts to the gray market, in equilibrium the reseller needs to optimize the amount of diversion to the gray market conditional on the gray market wholesale price \( s \). We have already solved for the optimal rate \( G_s(s, p) = (\eta - I^\circ(p, s))\lambda(p, s)/I^\circ(p, s) \). With an endogenous gray market wholesale price, gray market demand is also a function of the market price \( s \), \( G_s(s) = (b - s)/a \). For the equilibrium to be consistent, we equate the demand with supply and solve for the gray market wholesale price, \( S(p) = \{ s : G_r(s, p) = G_s(s) \} \). Such a \( S(p) \) exists and is unique as long as \( I^\circ_s < 0 \) holds. The gray market policy selects the profit-maximizing price \( p(w_\eta) = \arg \max_p (\Pi(p, I^\circ(p, s), S(p))) \). We use standard numerical optimization techniques (e.g. bisection) to perform root finding and maximization.
Figure 6  Reseller’s Response to Endogenous Gray Market Prices

(a) Reseller inv. policy ($\alpha = 0.5$)  (b) Reseller inv. policy ($\alpha = 2$)  (c) Existence of GM region

Note. $b = 2, m = 10, \eta = 20, w_o = 10$. Panels (a) and (b) the solid (resp. dashed) refers to $\alpha = 0$ (resp. $\alpha = 2$), $\gamma = 0.5$ and $u = 20\%$. In panel (c) $\alpha = 1$ (resp. $\alpha = 2$) and $u = 0\%$

Figure 7  Diversion Rate in Response to Endogenous Gray Market Prices

(a) Low $H$ ($H = 0.5, \alpha = 0.5$)  (b) High $H$ ($H = 1.3, \alpha = 0.5$)

Note. The solid (resp. dashed) refers to $\alpha = 0$ (resp. $\alpha = 2$), $b = 2, m = 10, \eta = 20, w_o = 10, \gamma = 0.5$ and $u = 20\%$.

Our numerical analysis covers a wide range of parameter values. Overall, as illustrated in Figures 6(a) and 6(b), our results and qualitative insights uncovered in previous sections remain intact when gray market prices are endogenous and dependent on the reseller’s diversion volume. Figure 6(a) shows the reseller’s optimal inventory policy under a low elasticity parameter $\alpha = 0.5$ and Figure 6(b) shows the case of a moderate elasticity parameter $\alpha = 2$. In both cases, once again the reseller follows the diversion strategy when $w_{\eta}$ is sufficiently small and holding cost $H$ falls in
an intermediate range of values. Alternatively, the reseller will follow the batch strategy if $H$ is too small, and follow the order-as-you-go strategy if $H$ is too large. When $a$ is increased from 0 to 2, the gray market prices become more sensitive to the reseller’s diversion rate. Since diversion becomes a less profitable option, there is a range in the parameter space where an increased $a$ leads the reseller to switch from diversion to order-as-you-go strategies. Lower gray market prices can dampen the reseller’s incentive because, first, more diversion leads to a lower gray market wholesale price and hence a higher unit diversion loss. Second, a lower gray market resale price will lead to increased cannibalization and hence lower profit in the authorized channel. Thus, all else being equal, with a larger value of $a$, it is less likely for the reseller to divert to the gray market. Figure 6(c) shows the boundary defined by $\alpha$ and $\gamma$, beyond which the cannibalization effect is too strong for the reseller to divert to gray market.

At lower demand elasticities such as $\alpha = 0.5$ shown in Figure 6(a), gray market diversion has a smaller effect on the demand in the authorized channel. As a result, the diversion decision is determined primarily by diversion loss, the difference between $w_\eta$ and $s$. As shown in Figure 6(a), if $H$ is low the reseller decides between the diversion and batch strategies. As shown in detail in Figure 7(a), when $H = 3$ the reseller prefers the diversion strategy as long as the discounted price $w_\eta$ is less than 5. In this case, when sensitivity parameter $a$ increases from 0 to 2, the amount of diversion $g$ decreases. Interestingly, the decision on the adoption of a diversion strategy remains the same. In contrast, if $H$ is high, the reseller decides between diversion and order-as-you-go strategies. As shown in Figure 7(b), when $H = 6$ the reseller selects the diversion strategy as long as the discounted price $w_\eta$ is less than 4.7. At this higher $H$, when the gray market wholesale price sensitivity parameter $a$ increases from 0 to 2, not only does the rate of diversion decrease, but, the reseller requires a greater discount in order to adopt the diversion strategy. Overall, when the gray market wholesale price is more sensitive to the reseller’s diversion, we expect the optimal diversion rate to be lower.

These results have further implications when we consider $a$ as a proxy for the impact of the
distribution network size. When a regional gray market is supplied by a small number of distributors, we expect gray market wholesale price to be more sensitive to a focal reseller. In such cases, our results indicate high holding cost resellers will be less likely to utilize a diversion strategy and diversion overall will be decreased.

6. Conclusions

This paper examines the impact of a gray market on authorized channel members and their decision making, specifically, the reseller’s inventory and pricing decisions, the supplier’s all-unit quantity discount policy, and the welfare of consumers in the authorized channel. We study a Stackelberg game and develop closed-form solutions for the reseller’s and the supplier’s subgame perfect equilibrium decisions. When the gray market wholesale price is sufficiently high, the resellers can use gray market diversion to reduce inventory costs as efficiently as resale price reduction. From an operational perspective, the presence of a gray market, by reducing the reseller’s inventory cost, can help the supplier to take advantage of economies of scale in batch processing. Overall, gray market diversion occurs only in markets where the inventory holding costs fall in an intermediate range of values. In the lower part of this range, the reseller would order in batches even without the option of gray market diversion. The gray market diversion reduces effective inventory size and shortens order cycles. Since the reseller faces reduced pressure to push out the inventory, the resale price goes up and sales volume goes down in the authorized channel. On the other hand, in the higher part of the intermediate range of holding cost, the reseller would follow the order-as-you-go strategy without the option of using the gray market. The gray market diversion reduces inventory holding costs and induces the reseller to order in batches. Now that the reseller enjoys a discounted price through batch orders, the resale price goes down and sales increase in the authorized channel. Given the monopoly nature of the model, the supplier is able to extract all the benefits resulting from gray market diversion. Finally, we extend the model and allow the gray market prices to depend on the amount of diversion from the reseller. We find our results regarding the effect of a gray market to be robust.
These results yield several useful implications on how firms may manage their distribution channels with potential gray market leakage. First, in industries with sufficiently attractive gray market prices, only those resellers in the intermediate range of inventory holding costs may divert goods to the gray market. This may help managers to identify the resellers prone to gray market diversion. Managers can easily monitor the gray market wholesale price and examine its gap from the discounted price $w_γ$ to assess the attractiveness of the gray market. Although our results cannot prescribe specific parameter ranges to predict a reseller’s gray market activities, the managers could use the values of parameters $η, h$, and $m$ to estimate the gray market activities. Consider a given industry with a constant batch size $η$. One may expect a reseller’s unit holding cost and market demand to depend on the reseller’s geographical location. For instance, a downtown location may have higher unit holding cost than a suburban location, a young and educated city may have a higher consumption rate for IT products than a rural and less tech-savvy market.

Second, managers should be aware of the complex effect of gray market diversion on the authorized channel. Among resellers engaging in gray market activities, those with relatively smaller holding costs use gray market diversion to reduce inventory costs and increase the prices in the authorized channel. This would reduce sales in the authorized channel which typically serve more valuable customers. The managers may consider treating these resellers differently from those with relatively large inventory holding costs. For the latter, the gray market diversion enables the resellers to enjoy batch discounts and pass a part of these discounts on to consumers in the authorized channel. Such incidences of gray market diversions could be treated with more tolerance because of the positive effect on the sales in the authorized channel.

Finally, this paper shifts from the usual focus on legal issues in gray markets and studies the strategic impact to the channel members. Gray markets have existed for a very long time and have gained momentum, growing in recent years through the online channel. The emergence of online markets makes it easier to establish independent resale operations and to reach a much larger geographic market. Since the gray market will remain for the foreseeable future, firms should fully consider its operational and marketing impacts on their authorized resellers. This paper
demonstrates that simply looking at one aspect (e.g., the pricing decisions) but ignoring another (e.g., the inventory decisions) could lead to erroneous conclusions on gray markets.

References


This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.
Online Appendix to  
“When Gray Markets Have Silver Linings: All-Unit Discounts, Gray Markets and Channel Management”  
PROOFS.

Proof of Proposition 1. To satisfy the demand \( d \geq 0 \), the order size is \( q \geq d \) and the diversion size is \( q - d \geq 0 \). If \( d \geq \eta \), then \( q \geq d \geq \eta \); the order cost \( c(q, d) = w_o q - s (q - d) \) that is linear in \( q \) with a slope \( w_o - s > 0 \), thus it is minimized at \( q^*(d) = d \). If \( d < \eta \), the order cost is

\[
c(q, d) = \begin{cases} 
w_o q - s (q - d) & \text{if } d \leq q < \eta, \\
w_q q - s (q - d) & \text{if } q \geq \eta. 
\end{cases}
\]

On the above two regions the cost-minimizing solutions are respectively \( q = d \) and \( q = \eta \). To find the optimum \( c^*(d) \) when \( d < \eta \), it suffices to compare the cost at the two solutions, i.e., \( c^*(d) = \min\{w_o d, (w_o - s) \eta + sd\} \): if \( 0 \leq d \leq \hat{q} \), \( c^*(d) = w_o d \) and if \( \hat{q} < d < \eta \), \( c^*(d) = (w_o - s) \eta + sd \).

Proof of Remark 1. First, consider an incremental discount:

\[
C(q) = \begin{cases} 
w_o q & \text{if } 0 \leq q < \eta, \\
w_q (q - \eta) + w_o \eta & \text{if } q \geq \eta.
\end{cases}
\]

To satisfy the demand \( d \geq 0 \), the order size is \( q \geq d \) and the diversion size is \( q - d \geq 0 \). If \( d \geq \eta \), then \( q \geq d \geq \eta \); the order cost \( c(q, d) = w_q (q - \eta) + w_o \eta - s (q - d) \) that is linear in \( q \) with a slope \( w_q - s > 0 \), thus it is minimized at \( q^*(d) = d \). If \( d < \eta \), the order cost is

\[
c(q, d) = \begin{cases} 
w_o q - s (q - d) & \text{if } d \leq q < \eta, \\
w_q (q - \eta) + w_o \eta - s (q - d) & \text{if } q \geq \eta.
\end{cases}
\]

On the above two regions the cost-minimizing solutions are respectively \( q = d \) and \( q = \eta \). To find the optimum \( c^*(d) \) when \( d < \eta \), it suffices to compare the cost at the two solutions, i.e., \( c^*(d) = \min\{w_o d, w_o \eta - s (\eta - d)\} = w_o d \). For both cases of \( d \geq \eta \) and \( d < \eta \), the optimal order size is \( q^*(d) = d \).

Second, consider a two-part tariff: \( C(q) = F + w_o q \), \( q > 0 \). To satisfy the demand \( d \geq 0 \), the order size is \( q \geq d \) and the diversion size is \( q - d \geq 0 \). The order cost is \( c(q, d) = F + w_o q - s (q - d) \) for \( q \geq d \) and is linear in \( q \) with a slope \( w_o - s > 0 \). Thus it is minimized at \( q^*(d) = d \).
Proof of Lemma 1. At any time $t$ the reseller can change his inventory position from the current position $I(t)$ to a new position $I(t) + \Delta I(t)$ by a combination of order $q(t)$ from the supplier and gray market diversion $g(t)$. Fix any time $t$. We first argue that since replenishment is instantaneous, it is suboptimal for $\Delta I(t) = i > 0$ if $I(t) > 0$. The action of changing the current inventory position $I(t)$ to $I(t) + i$ could be profitably delayed to the time $t_0 := \inf\{x > t : I(x) = 0\}$. Letting $\Delta I(t) = 0$ and $\Delta I(t_0) = \Delta I(t_0) + i$ has an improvement $h(t_0 - t)i > 0$ in the holding costs up to time $t_0$. Recursively applying this process results in an improved set of orders where $\Delta I(t) > 0$ only when $I(t) = 0$. We now argue that it is suboptimal for disposal of goods $\Delta I(t) = j < 0$ if $I(t) > 0$ since this action could have been profitably performed at an earlier time $t_{-1} := \sup\{x < t : I(x) = 0\}$, which has a holding cost improvement $h(t - t_{-1}) |j| > 0$. Therefore, in an efficient inventory policy any ordering or gray market diversion occurs only at times when $I(t) = 0$. The set of times when $I(t) = 0$ represents a set of renewal points. Since the demand rate is stationary, the optimal action is identical at each of these times. To complete the proof let $I$ be equal to the optimal inventory adjustment when $I(t) = 0$ and then $q^*(I)$ and $g^*(I)$ correspond to the optimal order and gray market diversion quantities respectively.

Proof of Proposition 2. To solve for the optimal inventory policy, the reseller selects the cycle inventory level $I$ that minimizes the total costs. Given the optimal zero-inventory policy characterized by Lemma 1, the total costs for each order cycle of length $I/\lambda$ consist of order cost $c^*(I)$ given by Proposition 1 and holding cost $h I^2/(2\lambda)$. We can then calculate the long-run average cost per unit time, denoted by $G(I,p,s)$ with dependence on $p$ and $s$ suppressed in this proof, as $c^*(I)\lambda/I + h I/2$. By substituting the reseller’s optimal cost function given by Proposition 1 (where $d = I$), we obtain the expression for $G(I)$ as follows:

$$G(I) = \begin{cases} w_n m / (p^\alpha) + h I/2 & \text{if } 0 \leq I < \hat{q}, \\ (w_n - s) \eta m / [I (p - \gamma p_s)^\alpha] + s m / (p - \gamma p_s)^\alpha + h I/2 & \text{if } \hat{q} \leq I < \eta, \\ w_n m / (p - \gamma p_s)^\alpha + h I/2 & \text{otherwise.} \end{cases}$$

Recall that in the first and third cases, the reseller orders up to the desired cycle inventory level $I$ and sells the entire order through the authorized channel over time; no goods are diverted to the gray market in these two cases. However, in the second case, the reseller orders up to the
quantity of \( \eta \) to enjoy the quantity discount and sells the excess amount \( \eta - I \) to the gray market.

Within this range \( \hat{q} \leq I < \eta \), the reseller will choose a locally optimal cycle inventory level \( I^o \) that minimizes the cost \( G(I) \), where \( I^o := \sqrt{\frac{2(w_\eta - s)}{\eta H(p - \gamma p_s)}} = \eta \sqrt{\frac{w_\eta - s}{H(p - \gamma p_s)}} \).

The reseller selects the optimal cycle inventory and gray market diversion by comparing the minimum cost \( G(I) \) in each of the three regions.

Since the demand and resale price is fixed for the reseller, the reseller’s revenue is fixed. The reseller is aiming at minimizing cost \( G(I) \). To find the minimum of \( G(I) \) we compare the optimal solutions for each region of \( I \in [0, \hat{q}] \), \( I \in (\hat{q}, \eta) \) and \( I \in [\eta, \infty) \). Over the first and third regions, \( G(I) \) is a linearly increasing function and is minimized at \( I = 0 \) and \( I = \eta \) respectively. Over the second region \( I \in (\hat{q}, \eta) \), \( G(I) \) is convex and minimized at an interior point \( I^o \) if it is indeed in \( (\hat{q}, \eta) \).

Otherwise, \( G(I) \) is minimized at one of the boundary points \( \hat{q} \) or \( \eta \).

The necessary and sufficient condition for \( I^* = I^o \) is \( I^o \in (\hat{q}, \eta) \). \( G(I^o) < G(0) \) and \( G(I^o) < G(\eta) \).

The feasibility condition \( I^o > \hat{q} \) holds if and only if \( H(p - \gamma p_s)^o < \frac{(w_o - s)^2}{(w_o - s)} \); The other feasibility condition \( I^o < \eta \) holds if and only if \( H(p - \gamma p_s)^o > (w_\eta - s) \). The optimality condition \( G(I^o) \leq G(\eta) \) always holds since \( G(I) \) is continuous at \( I = \eta \). Finally, the other optimality condition \( G(I^o) < G(0) \) holds if and only if \( H(p - \gamma p_s)^o < \frac{(w_o - s)^2}{[4(w_\eta - s)]} \). Taking the intersection of regions defined by the feasibility and optimality conditions yields that \( I^* = I^o \) if and only if \( (w_\eta - s) < H(p - \gamma p_s)^o < (w_o - s)^2/[4(w_\eta - s)] \). Such holding costs exist only if \( w_o \in (s, (w_o + s)/2) \).

\( I^* = \eta \) if and only if \( I^* \neq I^o \) and \( G(\eta) \leq G(0) \). \( G(\eta) \leq G(0) \) holds if and only if \( H(p - \gamma p_s)^o \leq w_o - w_\eta \) with equality holding at \( H = w_o - w_\eta \). Intersecting \( H(p - \gamma p_s)^o \leq w_o - w_\eta \) with the region where \( I^* \neq I^o \), i.e., \( H(p - \gamma p_s)^o \notin (w_\eta - s, (w_o - s)^2/[4(w_\eta - s)]) \), results in \( H(p - \gamma p_s)^o \leq \min\{w_\eta - s, w_o - w_\eta\} \). Note that \( w_o - s < w_o - w_\eta \) if and only if \( w_o < (w_o + s)/2 \). The necessary and sufficient condition for \( I^* = \eta \) follows immediately.

The remaining possible holding cost regions are \( H(p - \gamma p_s)^o > (w_o - s)^2/[4(w_\eta - s)] \) if \( w_\eta < (w_o + s)/2 \), \( H(p - \gamma p_s)^o > w_o - w_\eta \) otherwise, which correspond to \( I^* = 0 \).

**Lemma 2.** (Suppliers Profit Function under Exogenous Resale Price). *Given that the reseller employs the optimal inventory policy in response to a discounted wholesale price \( w_\eta \),*
the supplier receives the following profit per unit of time:

\[ \Pi(w) = \begin{cases} 
  m(w - c) \sqrt{H} & \text{if } s < w \leq s + H(p - \gamma p_s)^\alpha, \\
  m(w - c) \frac{H}{|w - s|} \left( \frac{p - \gamma p_s}{(p - \gamma p_s)^\alpha} \right) & \text{if } s + H(p - \gamma p_s)^\alpha < w < w_0 - H(p - \gamma p_s)^\alpha, \\
  m(w - c) \frac{H}{|w - s|} \left( \frac{p - \gamma p_s}{(p - \gamma p_s)^\alpha} \right) & \text{if } w_0 - H(p - \gamma p_s)^\alpha < w \leq w_0; 
\end{cases} \]

if \( H \geq (w_o - s)/[2(p - \gamma p_s)^\alpha], \)

\[ \Pi(w) = \begin{cases} 
  m(w - c) \sqrt{H} & \text{if } s < w \leq s + \frac{(w_o - s)^2}{4H(p - \gamma p_s)^\alpha}, \\
  m(w - c) \frac{H}{|w - s|} \left( \frac{p - \gamma p_s}{(p - \gamma p_s)^\alpha} \right) & \text{if } s + \frac{(w_o - s)^2}{4H(p - \gamma p_s)^\alpha} < w \leq w_0. 
\end{cases} \]

**Proof of Lemma 2.** The supplier’s profit depends on the inventory strategy the reseller optimally selects according to conditions shown in Proposition 2. The supplier’s profit per unit of time is equal to the rate the supplier supplies the reseller multiplied by the profit margin per unit. The margin is \( w_o - c, \) per unit if \( I^* = 0 \) and \( w_o - c, \) per unit otherwise. If no goods are sold to the gray market (i.e. \( I = 0 \) or \( I = \eta \)), the supplier’s demand rate is equivalent to the reseller’s demand rate \( \lambda. \) However, if \( I = I^o \) then the supplier’s demand rate is \( \eta \lambda / I^o = \lambda \sqrt{H(p - \gamma p_s)^\alpha/(w - s)}. \) The supplier’s profit function is therefore

\[ \Pi(w) = \begin{cases} 
  \lambda(w_o - c) & \text{if } I^* = 0, \\
  \lambda(w - c) \sqrt{H(p - \gamma p_s)^\alpha} & \text{if } I^* = I^o, \\
  \lambda(w - c) & \text{if } I^* = \eta. 
\end{cases} \]

It remains to show that the conditions denoted in the proposition are sufficient to entail the appropriate reseller’s best response.

We begin by considering the conditions on the discount which imply \( I^* = \eta. \) The conditions where \( I^* = \eta \) denoted in Proposition 2 are equivalent to \( w \in (s + H(p - \gamma p_s)^\alpha, (w_o + s)/2] \cup [(w_o + s)/2, w_o - H(p - \gamma p_s)^\alpha). \) Hence \( I^* = \eta \) if and only if \( w \in (s + H(p - \gamma p_s)^\alpha, w_o - H(p - \gamma p_s)^\alpha) \) and such a \( w \) exists if and only if \( w_o - s > 2H(p - \gamma p_s)^\alpha. \) Therefore, as stated in the proposition, \( \Pi(w) = \lambda(w - c) \) if and only if \( w \in (s + H(p - \gamma p_s)^\alpha, w_o - H(p - \gamma p_s)^\alpha) \) which is non-empty only if \( w_o - s > 2H(p - \gamma p_s)^\alpha. \)

Similarly the conditions from Proposition 2 which imply \( I^* = I^o \) can be summarized as \( w \in (s, s + H(p - \gamma p_s)^\alpha) \cap (s, s + (w_o - s)^2/(4H(p - \gamma p_s)^\alpha)) \cap (s, (w_o + s)/2), \) which can be simplified by considering whether \( w_o - s \leq \eta H/\lambda. \) If \( w_o - s \leq 2H(p - \gamma p_s)^\alpha \) then \( I^* = I^o \iff w \in \)
generating profit \( \lambda \) for the supplier. We can now compare supplier’s profits given the reseller’s best response of \( w \) and is therefore maximized at 

\[ \Pi = \Pi^* \Leftrightarrow w_\eta \in (s, s + H(p - \gamma p_s)^\alpha) . \]

The remaining scenarios of \( w_\eta \) are attributed to when the reseller selects \( I^* = 0 \). Instantiating the regions corresponding to each of the reseller’s inventory policy into \( \Pi(w_\eta) \) is sufficient to verify that Lemma 2 holds. \( \Box \)

**Proof of Proposition 3.** Given the profit function in Lemma 2, we solve the problem of optimizing \( \Pi(w_\eta) \) over \( w_\eta \in (s, w_o] \). For those discounted prices \( w_\eta \) which are close enough to \( w_o \) and elicit \( I^* = 0 \), the profit function \( \Pi(w_\eta) = \lambda(w_o - c_o) \) remains a constant. It is sufficient to set \( w_\eta = w_o \) to generate this profit. Consider \( H \geq (w_o - s)/[2(p - \gamma p_s)^\alpha] \). By Lemma 2, if \( w_\eta < s + (w_o - s)^2/(4H(p - \gamma p_s)^\alpha) \), then \( I^* > 0 \). Under the assumption that \( s < c_o \), \( \Pi(w_\eta) \) is increasing over \( w_\eta \in (s, s + (w_o - s)^2/(4H(p - \gamma p_s)^\alpha)] \), and is maximized at \( w_\eta = s + (w_o - s)^2/(4H(p - \gamma p_s)^\alpha) \) generating profit \( \lambda[(w_o - s)/2 - 2H(p - \gamma p_s)^\alpha(c_o - s)/(w_o - s)] \) for the supplier. Consider \( H < (w_o - s)/[2(p - \gamma p_s)^\alpha] \). If \( w_\eta \leq w_o - H(p - \gamma p_s)^\alpha \), then \( I^* > 0 \). Again since \( s < c_o \), it is easy to see that \( \Pi(w_\eta) \) is continuous at \( w_\eta = s + H(p - \gamma p_s)^\alpha \) and increasing over \( w_\eta \in (s, w_o - H(p - \gamma p_s)^\alpha] \), and is therefore maximized at \( w_\eta = w_o - H(p - \gamma p_s)^\alpha \) generating profit \( \lambda(w_o - H(p - \gamma p_s)^\alpha - c_o) \) for the supplier. We can now compare supplier’s profits given the reseller’s best response of \( I^* > 0 \) or \( I^* = 0 \) depending on whether \( (w_o - s)/2 > H(p - \gamma p_s)^\alpha \). When \( H(p - \gamma p_s)^\alpha \geq (w_o - s)/2 \), setting \( w_\eta = w_o \) generates greater profit for the supplier than \( w_\eta = s + (w_o - s)^2/(4H(p - \gamma p_s)^\alpha) \) if and only if \( H(p - \gamma p_s)^\alpha > (2c_o - w_o - s)(w_o - s)/[4(c_o - s)] \). When \( H(p - \gamma p_s)^\alpha < (w_o - s)/2 \), setting \( w_\eta = w_o \) generates greater profit for the supplier than \( w_\eta = w_o - H(p - \gamma p_s)^\alpha \) if and only if \( H(p - \gamma p_s)^\alpha > c_o - c_\eta \). \( \Box \)

**Proof of Corollary 1.** By Proposition 3, if \( H(p - \gamma p_s)^\alpha < \min\{(w_o - s)/2, c_o - c_\eta\} \), then \( w_\eta^* = w_o - H(p - \gamma p_s)^\alpha \) and the reseller’s best response is \( I^* = \eta \). The quantity discount \( H(p - \gamma p_s)^\alpha \) per unit off \( w_o \) is just to offset the increased holding cost in induced strategy \( I^* = \eta \) compared to the inventory strategy \( I^* = 0 \) when no quantity discount is offered. Again by Proposition 3, if \( (w_o - s)/2 \leq H(p - \gamma p_s)^\alpha \leq (2c_o - w_o - s)(w_o - s)/[4(c_o - s)] \), then \( w_\eta^* = s + (w_o - s)^2/(4H(p - \gamma p_s)^\alpha) \) and the best response is \( I^* = I^\alpha = \sqrt{(w_\eta^* - s)/H(p - \gamma p_s)^\alpha} \eta = (w_o -
For the reseller, the increased holding cost per cycle with length $I^o/\lambda$ is $(I^o/2)(I^o/\lambda)h = (w_o - s)^2\eta/(4H(p - \gamma p_s)^{\alpha})$ as compared to the inventory strategy $I^* = 0$; the loss in gray market diversion within the same cycle is $(w^*_\eta - s)(\eta - I^o) = (w_o - s)^2\eta[1 - (w_o - s)/(2H(p - \gamma p_s)^{\alpha})]/(4H(p - \gamma p_s)^{\alpha})$. The reseller’s gain from the quantity discount for the same cycle is $(w_o - w^*_\eta)I^o = [w_o - s - (w_o - s)^2/(4H(p - \gamma p_s)^{\alpha})][w_o - s]\eta/(2H(p - \gamma p_s)^{\alpha})$, which is exactly equal to the sum of the increased holding cost and loss in gray market diversion for the same cycle. Therefore, we can conclude that the supplier’s optimal all-unit quantity discount leaves the reseller with zero profits. □

Proof of Proposition 4. The reseller selects his resale price $p$ and inventory policy $I$ to maximize the expected profit per unit of time. We write the reseller’s profit in terms of the endogenously determined demand rate $\lambda$: $\pi(I, \lambda) = \sqrt{m\lambda} + \gamma p_s \lambda - G(I, \lambda)$, where $\sqrt{m\lambda} + \gamma p_s \lambda$ is the reseller’s revenue per unit of time and $G(I, \lambda)$ is the total costs per unit of time among all zero-inventory policies characterized by the initial cycle inventory level $I$. By Proposition 2, for any given resale price $p$ and its corresponding demand rate $\lambda$, the reseller will choose the unique inventory policy that minimizes $G(I, \lambda)$. Using the reseller’s optimal inventory response to an arrival rate, we derive the minimum inventory cost function $G^*(\lambda) = G(I^*(\lambda))$ as follows:

if $s < w_\eta < (w_o + s)/2$,

$$G^*(\lambda) = \begin{cases} 
\frac{w_o}{\eta} \lambda 
& \text{if } 0 < \lambda < \frac{4(w_o - s)mH}{(w_o - s)^2}, \\
\sqrt{4(\eta - s)mH} + s\lambda 
& \text{if } \frac{4(w_o - s)mH}{(w_o - s)^2} \leq \lambda < \frac{mH}{w_\eta - s}, \\
w_\eta \lambda + mH 
& \text{otherwise};
\end{cases}$$

if $(w_o + s)/2 \leq w_\eta \leq w_o$,

$$G^*(\lambda) = \begin{cases} 
w_o \lambda 
& \text{if } 0 < \lambda < \lim sup_{c \to w_\eta} - \frac{mH}{w_\eta - c}, \\
w_\eta \lambda + mH 
& \text{otherwise}.
\end{cases}$$

The above cost function includes the cost of ordering, diversion and holding inventory. When the order size is $\eta$ and there is no gray market diversion, the reseller enjoys the low unit cost $w_\eta$ but suffers an average holding cost of $mH$ per unit. In the case where the reseller diverts to the gray market, the reseller optimizes the diversion quantity $\eta - I$ by comparing the holding cost $hI/2$ with the diversion cost $w_\eta - s$. The reseller profit function with demand function $\lambda(p) = m/(p - \gamma p_s)^2$ is
\[ \pi(\lambda) = \begin{cases} 
\pi_0(\lambda) := \sqrt{m \lambda + (\gamma p_s - w_o)\lambda}, & \text{if } 0 < \lambda \leq \frac{4}{mH}, \\
\pi^o(\lambda) := \sqrt{m \lambda + (\gamma p_s - s)\lambda - \sqrt{4(w_o - s)mH}}, & \text{if } \frac{4}{mH} < \lambda \leq \frac{mH}{w_o - s}, \\
\pi_\eta(\lambda) := \sqrt{m \lambda + (\gamma p_s - w_o)\lambda - mH}, & \text{otherwise,}
\end{cases} \]

where \(\pi_0\), \(\pi^o\) and \(\pi_\eta\) correspond to when the reseller adopts the inventory policy \(I^* = 0\), \(I^* = I^o\) and \(I^* = \eta\) respectively. Note that \(\pi_0(\lambda)\) and \(\pi_\eta(\lambda)\) are concave since \(m > 0\) and \(\pi^o(\lambda)\) is concave if \(\sqrt{m} - \sqrt{4(w_o - s)mH} > 0\). We take the derivative of \(\pi_0(\lambda)\), \(\pi^o(\lambda)\) and \(\pi_\eta(\lambda)\) with respect to \(\lambda\) as 
\[
\frac{\partial \pi_0(\lambda)}{\partial \lambda} = \sqrt{m}/(2\sqrt{\lambda}) + \gamma p_s - w_o, \\
\frac{\partial \pi^o(\lambda)}{\partial \lambda} = \sqrt{m}/(2\sqrt{\lambda}) + (\gamma p_s - s) - \sqrt{(w_o - s)mH}/(2\lambda), \\
\frac{\partial \pi_\eta(\lambda)}{\partial \lambda} = \sqrt{m}/(2\sqrt{\lambda}) + \gamma p_s - w_o.
\]

The local optima satisfying the first-order conditions are \(\lambda^*_1 = m/(4(w_o - \gamma p_s)^2)\), \(\lambda^*_2 = (\sqrt{m} - \sqrt{4(w_o - s)mH})/4(s - \gamma p_s)^2\) and \(\lambda^*_3 = m/(4(w_o - \gamma p_s)^2)\) respectively, and the corresponding profits are \(\pi_0(\lambda^*_1) = m/(4(w_o - \gamma p_s))\), \(\pi^o(\lambda^*_2) = (\sqrt{m} - \sqrt{4(w_o - s)mH})^2/4(s - \gamma p_s)\), \(\pi_\eta(\lambda^*_3) = m/(4(w_o - \gamma p_s)) - mH\). The continuity of \(\pi(\lambda)\) is easily verified by checking at the two breakpoints \(\lambda_A = 4(w_o - s)mH/[4(w_o - s)^2]\) and \(\lambda_B = mH/[4(w_o - s)]\). Since \(\lim_{\lambda \to \lambda_A^+} \frac{\partial \pi_0(\lambda)}{\partial \lambda} < \lim_{\lambda \to \lambda_A^-} \frac{\partial \pi^o(\lambda)}{\partial \lambda}\), we eliminate the breakpoint \(\lambda_A\) as a global optimum. Since \(\lim_{\lambda \to \lambda_B^+} \frac{\partial \pi^o(\lambda)}{\partial \lambda} = \lim_{\lambda \to \lambda_B^-} \frac{\partial \pi_\eta(\lambda)}{\partial \lambda}\), the global optimum \(\lambda^* = \lambda_B\) only if \(\lambda^*_2 = \lambda^*_3 = \lambda_B\). Hence, we conclude that the global optimum \(\lambda^*\) must be one of the local optima \(\lambda^*_1\), \(\lambda^*_2\) and \(\lambda^*_3\).

It remains to check under what conditions each local optimum dominates. First, note that 
\[
\lim_{\lambda \to \lambda_B^-} \frac{\partial \pi^o(\lambda)}{\partial \lambda} = \lim_{\lambda \to \lambda_B^+} \frac{\partial \pi_\eta(\lambda)}{\partial \lambda} \leq 0\]

if and only if \(H \geq (w_o - s)/[4(w_o - \gamma p_s)^2]\). Hence, a necessary condition for \(\lambda^*_3\) to be a global optimum is \(H \geq (w_o - s)/[4(w_o - \gamma p_s)^2]\) and a necessary condition for \(\lambda^*_2\) to be a global optimum is \(H \leq (w_o - s)/[4(w_o - \gamma p_s)^2]\). Second, we compare the profit of each of the batch order policies \(I^* = I^o\) or \(I^* = \eta\) to the profit of the order-as-you-go policy \(I^* = 0\): 
\[
\pi_0(\lambda^*_1) > \pi^o(\lambda^*_2) \iff H > (1 - \sqrt{(s - \gamma p_s)/(w_o - \gamma p_s)^2})/[4(w_o - s)]
\]

and 
\[
\pi_0(\lambda^*_3) > \pi_\eta(\lambda^*_2) \iff H > 1/[4(w_o - \gamma p_s)] - 1/[4(w_o - \gamma p_s)].
\]

Lastly, conditioned on whether \(w_o - \gamma p_s < \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)}\), the breakpoints on \(H\) can be ordered as follows: if \(w_o - \gamma p_s < \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)}, (1 - \sqrt{(s - \gamma p_s)/(w_o - \gamma p_s)^2})/[4(w_o - s)] > (w_o - s)/[4(w_o - \gamma p_s)^2]\) and if \(w_o - \gamma p_s \geq \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)}, 1/[4(w_o - \gamma p_s)] - 1/[4(w_o - \gamma p_s)] \geq (w_o - s)/[4(w_o - \gamma p_s)^2]\).

Therefore, it is not hard to conclude that when \(w_o - \gamma p_s \geq \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)}\), the optimal
demand rate is \( \lambda^* = \lambda_2^* \) if \( H > (1 - \sqrt{(s - \gamma p_s)}/(w_o - \gamma p_s))^2/[4(w_o - \gamma p_s)^2] \), \( \lambda^* = \lambda_2^* \) if \( (w_o - s)/[4(w_o - \gamma p_s)^2] < H \leq (1 - \sqrt{(s - \gamma p_s)}/(w_o - \gamma p_s))^2/[4(w_o - s)] \) and \( \lambda^* = \lambda_3^* \) otherwise; when \( w_o \geq \sqrt{w_o s} \) the optimal demand rate is \( \lambda^* = \lambda_1^* \) if \( H > 1/[4(w_o - \gamma p_s)] - 1/[4(w_o - \gamma p_s)] \) and \( \lambda^* = \lambda_3^* \) otherwise. By the relationship between price and demand rate \( \lambda \), the corresponding reseller’s optimal pricing and inventory policy follows immediately.

**Proof of Corollary 2.** This corollary follows from the profit functions described in the proof of Proposition 4. Namely, a diversion strategy will be preferred to a batch strategy if \( \pi^*(\lambda_2^*) > \pi_0(\lambda_2^*) \). Solving for \( H \) shows the transition point occurs at \( H_0 = \frac{w_o - s}{4(w_o - \gamma p_s)^2} \) confirming the first part of the Corollary. The transition point is a unimodal function of \( w_o \) with a unique maximizer at \( w_o^* = 2s - \gamma p_s \). This is confirmed by observing that \( \partial H_0(s + \epsilon)/\partial w_o > 0 \) for small \( \epsilon \) and then determining that there is a unique critical point. The proof of the Corollary is completed by evaluating \( H_0(w_o^*) \). □

**Proof of Proposition 5.** It is readily apparent that when the reseller’s best response is to order in batches without any gray market diversion, the supplier enjoys economies of scale from batch processing at the same rate as the demand rate \( \lambda(p^* = 2w_o - \gamma p_s) = m/[4(w_o - \gamma p_s)^2] \) in the authorized channel. As a result, the supplier’s profit per unit of time is \( \Pi_0(w_o) := (w_o - c_o) m/[4(w_o - \gamma p_s)^2] \).

When the reseller’s best response is to order on demand and not to hold inventory at all, the supplier delivers the product at the list price and the same rate as the demand rate \( \lambda(p^* = 2w_o - \gamma p_s) = m/[4(w_o - \gamma p_s)^2] \) in the authorized channel, and hence the supplier’s profit per unit of time is \( \Pi_0 := (w_o - c_o) m/[4(w_o - \gamma p_s)^2] \) which is independent of the size of the all-unit discount. Finally, when the reseller’s best response is to order in batches with part of the order diverted to the gray market, the supplier enjoys economies of scale from orders of size \( \eta \) every \( I^*/\lambda(p^*) \) time units, and hence the supplier’s profit per unit of time can be shown to be

\[
\Pi^*(w_\eta) := \frac{m(w_o - c_o)}{2(s - \gamma p_s)} \left( \sqrt{\frac{H}{w_\eta - s}} - 2H \right).
\]

As a precursor to establishing this proposition we derive the supplier’s profit function. By Proposition 4, the supplier’s profit function given that the reseller employs the optimal pricing and inventory decisions can be described as
\[ \Pi(w_\eta) = \begin{cases} \\ \Pi^L(w_\eta) & \text{if } w_\eta \in R^L_\eta \cup R^H_\eta, \\ \Pi^0(w_\eta) & \text{if } w_\eta \in R^0_\eta, \\ \Pi_0 & \text{if } w_\eta \in R^L_\eta \cup R^H_\eta. 
\]

We let \( \bar{w} := 1/(4H + 1/(w_o - \gamma p_s)) + \gamma p_s \), \( \hat{w} := s + (1 - \sqrt{(s - \gamma p_s)/(w_o - \gamma p_s)})^2/4H \), \( w := \left(1 - \sqrt{1 - 16(s - \gamma p_s)H}/(8H) + \gamma p_s, \bar{w} := \left(1 + \sqrt{1 - 16(s - \gamma p_s)H}/(8H) + \gamma p_s, \right. \) where \( w \) and \( \bar{w} \) are the two real roots, if they exist, of the quadratic equation \( f(w) := 4H(w - \gamma p_s)^2 - w + s = 0 \). Then, the regions of quantity discount that induces different reseller pricing and inventory decisions are \( R^H_0 := \left\{ \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \leq w_\eta \leq w_o \mid w_\eta > \hat{w} \right\}, \) \( R^H_\eta := \left\{ \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \leq w_o \mid w_\eta \leq \bar{w} \right\}, \) \( R^L_\eta := \left\{ s < w_\eta < \sqrt{w_o s} \mid 4H(w_\eta - \gamma p_s)^2 - w_\eta + s \leq 0 \right\}, \) \( R^o := \left\{ s < w_\eta < \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \mid w_\eta \leq \hat{w}, 4H(w_\eta - \gamma p_s)^2 - w_\eta + s > 0 \right\} \) and \( R^L_0 := \left\{ s < w_\eta < \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \mid w_\eta > \hat{w} \right\}. \) Regions \( R^H_0 \) and \( R^H_\eta \) are mutually exclusive, with one of them possibly being an empty set. Regions \( R^L_0, R^o \) and \( R^H_0 \) are mutually exclusive, with no more than two of them possibly being an empty set.

To simplify the profit function, we condition on the magnitude of \( H \) according to the following three cases.

Case (i). Consider \( 4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s) - 1/(w_o - \gamma p_s)} \), which is equivalent to \( \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} < \hat{w} \), hence \( R^H_\eta = \left[ \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s}, \hat{w} \right] \) and \( R^H_0 = (\hat{w}, w_o]. \) It can be easily verified that \( 4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s) - 1/(w_o - \gamma p_s)} \) is also equivalent to \( \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} < \hat{w} \), hence \( R^L_\eta = \emptyset \). Note that \( f(w = s) = 4H(w - \gamma p_s)^2 \geq 0 \) and \( f(w = \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s}) < 0 \) when \( 4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s) - 1/(w_o - \gamma p_s)} \), hence the smaller root \( \bar{w} \) must be real-valued and exist between \( s \) and \( \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \), i.e., \( \bar{w} \in \left[ s, \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \right]. \) Hence \( R^L_\eta = \left[ \bar{w}, \sqrt{(w_o - \gamma p_s)(s - \gamma p_s) + \gamma p_s} \right] \) and \( R^o = (s, \bar{w}) \) \( (R^o \text{ degenerates to } \emptyset \text{ if } \bar{w} = s \text{ which is equivalent to } s4H = 0) \). In summary, \( R^L_\eta \cup R^H_0 = (\bar{w}, w_o], \) \( R^L_0 \cup R^H_\eta = [\bar{w}, w_o] \) and \( R^o = (s, w) \).

Case (ii). Consider \( 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s) - 1/(w_o - \gamma p_s)} \leq 4H \leq 1/[4(s - \gamma p_s)] \). Such an interval of \( H \) indeed exists since \( 1/[4(s - \gamma p_s)] + 1/(w_o - \gamma p_s) \leq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} \) with equality holding if and only if \( w_o - \gamma p_s = 4(s - \gamma p_s) \). Note that \( 4H \geq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - \)}
1/(w_o - \gamma p_s) is equivalent to \( \hat{w} \leq \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s \); hence \( R^H_0 = \emptyset \) and \( R^H_0 = \left[ \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s, w_o \right] \). Also note that \( 4H \geq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \) is equivalent to \( \hat{w} \leq \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s \); hence \( R^L_0 = \left( \hat{w}, \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s \right) \). Furthermore, when \( 4H \leq 1/[4(s - \gamma p_s)] \), the discriminant of the quadratic equation \( f(w) = 0 \) is non-negative, hence the roots \( \hat{w} \) and \( \overline{w} \) of equation \( f(w) = 0 \) must be real-valued. It is easy to check that \( 4H \geq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \) is equivalent to \( \overline{w} \leq \hat{w} \). Since \( f(w = s) \geq 0 \), \( f(w = \hat{w}) \geq 0 \) and \( f(w = \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)}) \geq 0 \) with the last two inequalities ensured by \( 4H \geq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \), the roots \( w \) and \( \overline{w} \) must exist between \( s \) and \( \hat{w} \), i.e., \( [w, \overline{w}] \subseteq [s, \hat{w}] \). Hence \( R^L_\eta = \left[ w, \overline{w} \right] \) and \( R^o = (s, w) \cup (\overline{w}, \hat{w}) \). In summary, \( R^H_0 \cup R^L_0 = (\hat{w}, w_o) \), \( R^o = (s, w) \cup (\overline{w}, \hat{w}) \) and \( R^L_\eta \cup R^H_\eta = [w, \overline{w}] \).

Case (iii). Consider \( 4H \leq 1/[4(s - \gamma p_s)] \). According to the case (ii), we know that \( 4H \leq 1/[4(s - \gamma p_s)] \) \( \geq 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \); \( 4H > 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \) leads to that \( R^H_\eta = \emptyset \), \( R^H_0 = \left[ \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s, w_o \right] \) and \( R^L_0 = \left( \hat{w}, \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} + \gamma p_s \right) \). Moreover, \( 4H > 1/[4(s - \gamma p_s)] \) guarantees that \( f(w) > 0 \) for any \( w \), hence \( R^L_\eta = \emptyset \) and \( R^o = (s, \hat{w}) \). In summary, \( R^H_0 \cup R^L_0 = (\hat{w}, w_o) \), \( R^o = (s, \hat{w}) \) and \( R^L_\eta \cup R^H_\eta = \emptyset \).

Thus, the profit function can be expressed as follows completing the derivation:

case (i). if \( 4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \),
\[
\Pi(w_\eta) = \begin{cases} 
\Pi^o(w_\eta) & \text{if } s < w_\eta < w, \\
\Pi^o(w_\eta) & \text{if } \overline{w} \leq w_\eta \leq \hat{w}, \\
\Pi_0 & \text{if } \hat{w} < w_\eta \leq w_o; 
\end{cases} \quad (I^* = I^o)
\]

case (ii). if \( 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \leq 4H \leq 1/[4(s - \gamma p_s)] \),
\[
\Pi(w_\eta) = \begin{cases} 
\Pi_\eta(w_\eta) & \text{if } \overline{w} \leq w_\eta \leq \hat{w}, \\
\Pi^o(w_\eta) & \text{if } s < w_\eta < \overline{w} \text{ and } \overline{w} < w_\eta \leq \hat{w}, \\
\Pi_0 & \text{if } \hat{w} < w_\eta \leq w_o; 
\end{cases} \quad (I^* = I^o)
\]

case (iii). if \( 4H \leq 1/[4(s - \gamma p_s)] \),
\[
\Pi(w_\eta) = \begin{cases} 
\Pi^o(w_\eta) & \text{if } s < w_\eta \leq \hat{w}, \\
\Pi_0 & \text{if } \hat{w} < w_\eta \leq w_o. 
\end{cases} \quad (I^* = I^o)
\]

Given that the reseller employs the optimal pricing and inventory policy in response to a discounted wholesale price \( w^*_\eta \), the supplier earns the following profit per unit of time:
Taking the first-order derivative of $\Pi_\eta(w_\eta)$ with respect to $w_\eta$, we have $\partial \Pi_\eta(w_\eta)/\partial w_\eta = m/(4(w_\eta - \gamma p_s)^2)(1 - 2(w_\eta - c_\eta)/(w_\eta - \gamma p_s))$, hence the function $\Pi_\eta(w_\eta)$ is increasing on $(0, 2c_\eta - \gamma p_s]$ and decreasing on $[2c_\eta - \gamma p_s, \infty)$. Note that under the assumption that $(w_o - \gamma p_s)/4 < (s - \gamma p_s) < (c_\eta - \gamma p_s)$, we have $\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} < 2c_\eta - \gamma p_s$. Taking the first-order derivative of $\Pi_\eta(w_\eta)$ with respect to $w_\eta$, we have

$$\frac{\partial \Pi_\eta(w_\eta)}{\partial w_\eta} = \frac{m\sqrt{H}}{4(s - \gamma p_s)(w_\eta - s)\frac{3}{2}} \left[ -4\sqrt{H(w_\eta - s)}\frac{3}{2} + w_\eta + c_\eta - 2s \right].$$

Taking the second-order derivative of $\Pi_\eta(w_\eta)$ with respect to $w_\eta$, we have

$$\frac{\partial^2 \Pi_\eta(w_\eta)}{\partial w_\eta^2} = \frac{m\sqrt{H}(4s - 3c_\eta - w_\eta)}{8(s - \gamma p_s)(w_\eta - s)\frac{3}{2}} < 0.$$

Under the assumption that $s < c_\eta$, then $\partial^2 \Pi_\eta(w_\eta)/\partial w_\eta^2 < 0$ for $w_\eta > s$, namely, $\Pi_\eta(w_\eta)$ is strictly concave on $(s, \infty)$. Furthermore, since $s < c_\eta$, $\lim_{w_\eta \to s^+} \partial \Pi_\eta(w_\eta)/\partial w_\eta = \infty$. Under the assumption that $w_o - \gamma p_s)/4 < (s - \gamma p_s) < (c_\eta - \gamma p_s)$,

$$\frac{\partial \Pi_\eta(w_\eta)}{\partial w_\eta} \bigg|_{w_\eta = \bar{w}} = \frac{m\sqrt{H}}{4(s - \gamma p_s)(\bar{w} - s)\frac{3}{2}} \left[ \frac{(1 - \sqrt{(s - \gamma p_s)/w_o - \gamma p_s})^2}{4H} \left( 2\sqrt{(s - \gamma p_s)/(w_o - \gamma p_s)} - 1 \right) + c_\eta - s \right] > 0.$$

Therefore, $\Pi_\eta(w_\eta)$ is strictly increasing on $(s, \bar{w}]$.

In cases (i) and (ii). In both cases, $4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)$ with the form of the supplier's profit function corresponding to case (i) of the supplier profit function. If $4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)$, by the derivation of the supplier profit function, $w < \bar{w}$. Hence, $\Pi_\eta(w_\eta)$ is increasing on $(s, \bar{w}]$. Note that $\Pi_\eta(w_\eta)$ and $\Pi_\eta(w_\eta)$ are continuous at $w$. Furthermore, by the derivation of the supplier profit function, $w < \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)}$ and we know under the assumption $(w_o - \gamma p_s)/4 < (s - \gamma p_s) < (c_\eta - \gamma p_s)$, $\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} < 2c_\eta - \gamma p_s$, hence $w < 2c_\eta - \gamma p_s$. Recall that $\Pi_\eta(w_\eta)$ is increasing on $(0, 2c_\eta - \gamma p_s]$ and decreasing on $[2c_\eta - \gamma p_s, \infty)$. Therefore, if $2c_\eta - \gamma p_s < \bar{w}$, $\Pi_\eta(w_\eta)$ is increasing on $[w, 2c_\eta - \gamma p_s]$ and decreasing on $[2c_\eta - \gamma p_s, \bar{w}]$; otherwise, $\Pi_\eta(w_\eta)$ is increasing on $[w, \bar{w}]$. Finally, it remains to compare the supplier’s profit at the list price $w_o$ and the one at the discounted price $\min\{2c_\eta - \gamma p_s, \bar{w}\}$ that maximizes the profit when a discount is offered.
case (iii). We consider $4H < 1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)$. First, we consider $1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s) \leq 4H \leq 1/(4(s - \gamma p_s))$ with the form of the supplier’s profit function corresponding to case (ii) of the supplier profit function. Note that if $4H \leq 1/(4(s - \gamma p_s))$, $w \leq \bar{w} \leq \sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} < 2c_\eta - \gamma p_s$, where the second inequality is shown in the derivation of the supplier profit function and the last is due to the assumption $(w_o - \gamma p_s)/4 < (s - \gamma p_s) < (c_\eta - \gamma p_s)$. Hence $\Pi_\eta(w_\eta)$ is increasing on $[w, \bar{w}]$ and maximized at $\bar{w}$. We also know that $\Pi^o(\eta_\eta)$ is increasing on $(s, w]$ and $[\bar{w}, \hat{w}]$, and obtains its maximum at $\hat{w}$. Note that $\Pi_\eta(w_\eta)$ and $\Pi^o(\eta_\eta)$ are continuous at $w$ and $\bar{w}$. Therefore, if $1/\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - 1/(w_o - \gamma p_s)) \leq 4H \leq 1/(4(s - \gamma p_s))$, the discounted price $\hat{w}$ maximizes the supplier’s profit when a discount is offered. Second, we consider $4H > 1/(4(s - \gamma p_s))$ with the form of the supplier’s profit function corresponding to case (iii) of the supplier profit function. Recall that $\Pi^o(\eta_\eta)$ is strictly increasing on $(s, \hat{w}]$. Therefore, if $4H > 1/(4(s - \gamma p_s))$, the discounted price $\hat{w}$ maximizes the supplier’s profit when a discount is offered. Finally, combining the two subcases, it remains to compare the supplier’s profit at the list price $w_o$ and the one at the optimal discounted price $\hat{w}$. $\square$

Proof of Corollary 3. First we check case by case how in Proposition 5 the benefits from economies of scale are allocated between the supplier and reseller. When it is optimal for the supplier to offer a discounted price $w_\eta^s < w_o$, the supplier enjoys economies of scale. By Proposition 4, the reseller’s best response $(p^*(w_\eta^s), I^*(w_\eta^s))$ of pricing and inventory is either to take the $I^* = I^o$ strategy, namely, $(p^*(w_\eta^s), I^*(w_\eta^s)) = (2(w_\eta^s - \gamma p_s, \eta)/(s - \gamma p_s) - (w_\eta^s - s)\gamma p_s, s)\) or the $I^* = I^o$ strategy, namely, $(p^*(w_\eta^s), I^*(w_\eta^s)) = (2w_\eta^s - \gamma p_s, \eta)\) if the best response of the reseller $(p^*(w_\eta^s), I^*(w_\eta^s)) = (2w_\eta^s - \gamma p_s, \eta)$, then the reseller’s profit per unit of time is $\pi(\lambda(p^*(w_\eta^s))) = m/(4(w_\eta - \gamma p_s)) - mH$. When the supplier sets $w_\eta = w_o$ and does not enjoy economies of scale, the reseller’s profit per unit of time in the best response is $\pi(\lambda(p^*(w_o))) = m/(4(w_o - \gamma p_s))^2$. It is easy to see that if $w_\eta^s \leq (\text{resp.} <) \hat{w}$, the reseller is (resp. strictly) better off when the supplier offers a discounted price $w_\eta = w_\eta^s$ as compared to when the supplier does not, i.e., $\pi(\lambda(p^*(w_\eta^s))) \geq (\text{resp.} >) \pi(\lambda(p^*(w_o)))$. By the proof of Proposition 5, we can verify that to elicit $I^* = \eta$, in cases (i) and (ii) of Proposition 5, it is optimal for the supplier to
set \( w^*_\eta = \min\{2c_\eta - \gamma p_s, \tilde{w}\} \leq \tilde{w} \). Hence, in case (i), \( w^*_\eta = 2c_\eta - \gamma p_s < \tilde{w} \) and the reseller shares part of the benefits from economies of scale; in case (ii), \( w^*_\eta = \tilde{w} \) and the reseller shares no benefits.

By Proposition 5, the other scenario is that the supplier sets \( w^*_\eta = \tilde{w} \) to induce the reseller to take the \( I^* = I^o \) strategy. By Proposition 4, the corresponding best response of the reseller is \( (p^*(\tilde{w}), I^*(\tilde{w})) = \left(2\sqrt{(w_o - \gamma p_s)(s - \gamma p_s)} - \gamma p_s, \left(\sqrt{\frac{s - \gamma p_s}{w_o - \gamma p_s}} - \frac{s - \gamma p_s}{w_o - \gamma p_s}\right) \eta/(s - \gamma p_s)4H\right) \) and the reseller’s profit per unit of time is \( \pi(\lambda(p^*(\tilde{w}))) = \left(\sqrt{m} - \sqrt{4(\tilde{w} - s)mH}\right)^2/(4(s - \gamma p_s), = m/(4(w_o - \gamma p_s)) = \pi(\lambda(p^*(w_o))) \). Hence in this scenario, the reseller earns the same profit as without the quantity discount and shares no benefits from economies of scale.

From the analysis of all three cases, we can see that the resale price in equilibrium when the supplier enjoys economies of scale is always strictly smaller than the resale price \( 2w_o - \gamma p_s \) that the reseller will charge if the supplier does not enjoy economies of scale and offers the wholesale price at \( w_o \). Hence, consumers will always enjoy a lower resale price with the supplier’s economies of scale than without. \( \Box \)