Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Research Framework for Business Models: What Is Common Among Fast Fashion, E-Tailing, and Ride Sharing?

Gérard P. Cachon

informs
http://pubsonline.informs.org/journal/mnsc
MANAGEMENT SCIENCE
Vol. 66, No. 3, March 2020, pp. 1172–1192
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

A Research Framework for Business Models: What Is Common Among Fast Fashion, E-Tailing, and Ride Sharing?

Gérard P. Cachon[a]

[a] The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104
Contact: cachon@wharton.upenn.edu, http://orcid.org/0000-0002-7046-8835 (GPC)

Received: November 19, 2017
Revised: September 16, 2018
Accepted: November 6, 2018
Published Online in Articles in Advance: August 29, 2019

https://doi.org/10.1287/mnsc.2018.3275

Abstract. Every firm has a business model, which is the collection of strategic decisions that determine how the firm generates a sustainable enterprise through the creation of enough value (its supply model) and the extraction of a sufficient portion of that value (its revenue model). Innovative business models—for example, fast fashion (e.g., Zara), e-tailing (e.g., Amazon), and ride-sharing (e.g., Uber)—are capable of offering new products and services that generate considerable consumer utility and transform industries. This paper develops a research framework for understanding business models and how business models have evolved over time. Links are made to the existing literature (primarily in pricing and operations), and simple models are developed to unify and clarify existing research findings. Through this framework, it is possible (i) to identify the few design decisions that explain the success of these diverse firms with otherwise seemingly disparate models, and (ii) to speculate on potential future business-model innovations.

History: Accepted by Teck Ho, operations management.

Keywords: business models • theory • modeling • economics • innovation

1. Introduction

"Business model" and its offspring "business-model innovation" are buzzwords that have emerged to describe the upheaval that has occurred in many markets. Examples of firms with transformative business models include Zara with apparel (fast fashion), Amazon with e-commerce retailing (e-tailing), and Uber with transportation (ride sharing), among many others. These changes have led to novel sources of consumer value, remarkable financial opportunities for entrepreneurs, and a bounty of new phenomena and practices for academics to explore.

The first goal of this paper is to construct a framework for conceptualizing business models in general, and innovations in business models in particular. Although it might appear that business models come in a vast array of different "shapes, sizes, and colors," it is argued that all business models are defined by a relatively small set of high-level strategic decisions, each with its own basic trade-offs. The second goal of this paper is to use this business-model framework to highlight the linkages that exist among the growing set of research papers on various facets of business models—academics (appropriately) focus on narrow issues, but combining the various pieces forms a larger picture. By understanding the common threads among seemingly disparate business models, we hopefully are better able to evaluate the potential of a model and possibly even predict the emergence of successful new models.

2. Business-Model Definition and Framework

A business model is the set of strategic decisions that determine how a firm generates a sustainable and successful enterprise. There are three points worth emphasizing regarding this definition. First, a business model is not a single decision, but rather a *set of decisions* that must be considered as a collection. Second, a business model is based on *strategic decisions*—that is, decisions that have broad implications for the firm, are difficult to change, and often are made when the firm is established, thereby defining the essence of the firm and endowing it with its distinctive character. For example, the lack of a physical storefront has been a defining feature of Amazon's business model. Third, the emphasis with a business model is on *how* the firm does its business rather than *what* the firm offers its customers. The emphasis on "how," rather than "what," might seem misplaced because the success of many firms can be traced to what they offer their customers. For instance, Microsoft's Windows long dominated the operating-system market for personal computers, and Pfizer's cholesterol-reducing product, Lipitor, can safely be described as the blockbuster of blockbusters in the pharmaceutical industry. Firms do indeed thrive if they are lucky enough to have a distinctive and valuable product, especially if they have some legally sanctioned monopoly protection (e.g., patents and trademarks). But a great "what" is not the only way to succeed. Zara is not known for any one particular product or style. Its substantial and remarkable revenue growth has come more from how it delivers the products it sells rather than what specific products it offers. Similarly, Uber's innovation is not in what it offers (transportation) but, rather, how it offers it (independent contractors and dynamic pricing).

To add another layer of detail, the firm's set of strategic decisions in its business model can be divided into two parts, as illustrated in Figure 1: a revenue model and a supply model. The revenue model is how the firm earns the revenue used to procure resources (e.g., labor and capital). The supply model is how the firm manages and utilizes its resources to generate customer value. Customers value having their needs satisfied and doing so with transactional efficiency (e.g., a simple, fast, and easy process). For the firm to be sustainable over time, the amount of customer value created must be sufficient to enable enough revenue to be generated to cover the cost of the resources needed to produce the value.[1]

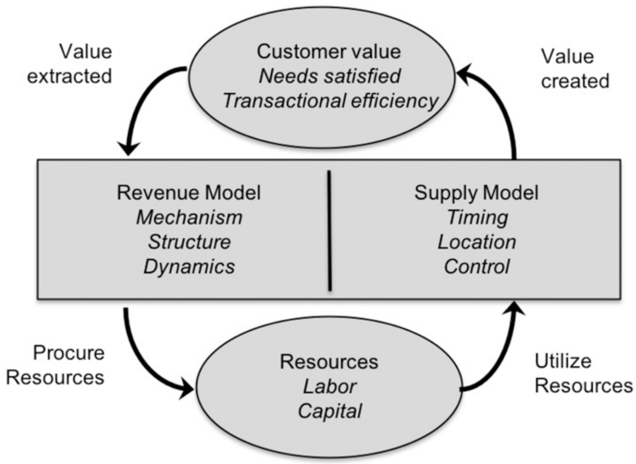
2.1. Customer Value and Resources

Customer value is generated when the firm effectively satisfies a need, with as little hassle as possible, either through a physical product or a service, or both. (For simplicity, "product" is used to describe physical goods as well as services.) For example, a customer's need can be as mundane as a box of pasta to make a meal, or as specific as transportation to move from one part of a city to another, or as elaborate as medical care. Needs are usually multidimensional, as in tasty gluten-free pasta, or quick and safe transportation, or effective and friendly medical care. The quality of a product, both in terms of performance quality (an absolute measure) and conformance quality (consistency in meeting specifications), influences how effectively the need is satisfied.

Although it is important to satisfy a need, customers also want their interaction with the firm to be as simple, convenient, and pleasant as possible. In other words, customers value transactional efficiency. This includes the time and effort to search, evaluate, select, and receive the firm's product. For example, online transactions are easier than phone transactions, nearby locations are more convenient than faraway locations, and one-day shipping is preferred over five-day shipping. A need well satisfied, but provided with a transactionally inefficient process, is of little value to a consumer. As such, transactional efficiency is a critical element in the customer-value proposition and one that plays an important role in many recent business-model innovations.

The actual creation of value is done through the coordination of an ensemble of resources. The two

Figure 1. The Components of a Business Model



obviously needed resources are labor and capital—firms generally need employees and some physical plant and equipment to function. Less obvious is the ownership of these resources. For example, labor could be freely provided by a firm's customers. The actual ownership structure is an important element of the supply model.

2.2. Revenue Model

The firm's "revenue model" describes how it earns revenue. Presuming customers value a firm's product, the primary role of the revenue model is to extract some of that value for the firm. For the firm to be sustainable, the revenue extracted must be sufficient to justify the cost of the firm's resources.

In its most basic form, a revenue model is quite simple—the firm sets a price for its product, and if a customer decides to accept the take-it-or-leave-it offer, then the customer consumes the product. This basic model involves a single price chosen by the firm for a single product at a single moment in time, which is contemporaneous with consumption. Each of those dimensions provides an opportunity for the firm to develop a new and more complex revenue model. To be specific, a revenue model specifies the transaction terms that occur between the firm and customers. There are three key three components in the revenue model: pricing mechanism (i.e., how are prices chosen?), payment structure (i.e., a fee per use or fixed fees?), and dynamics (i.e., how are terms adjusted over time?).

Some mechanism is always used to decide the amount that is exchanged between the customer and the firm. With a basic mechanism, the firm chooses a price to offer customers. But the firm could let customers decide on the price, such as through an auction mechanism or via their own take-it-or-leave-it offers to the firm. For instance, consumers could submit bids to a hotel for a room, which the hotel either accepts or not. A firm posted price is transactionally efficient—it is simple to communicate and can be executed immediately—but it puts the burden on the firm to choose the "right" price based on what the firm knows about consumer preferences and its own supply. More elaborate mechanisms, such as an auction and haggling, are transactionally more burdensome, but facilitate preference discovery and potentially more revenue extracted.

The payment structure refers to how payments are made in the transaction. For example, a firm selling a physical product, like a photocopier, could sell the copier with a single transaction to a customer. Or the firm could lease/rent the photocopier to a customer for its use over an interval of time. And with that option, the customer could pay a fixed fee for the use of the copier or a per-use fee that is proportional to the number of copies made. Firms selling services have a similar decision. A music retailer could offer customers the option to purchase individual songs in a library, or customers could be offered access to all of the songs in the library. If access is granted to the library, then the duration of this access must be established, along with possible quantity restrictions.

Although at a particular moment in time, the firm and a customer might agree on the terms of their transaction, there is no requirement that these offered terms remain constant over time or that both parties stick with the initial agreement. The dynamics of a revenue model are defined by the frequency and magnitude of changes in offered transaction terms as well as the extent to which initial agreements can be modified. For example, does an airline increase or decrease seat prices as the day of departure approaches? Does an apparel retailer maintain its prices or offer a discount near the end of the selling season? Does a seller of tickets to a sporting event allow the initial buyer to resell the ticket to another customer? These are challenging questions to analyze theoretically and equally challenging to implement effectively in practice for a host of reasons, ranging from mathematical complexity (i.e., can good solutions to these hard optimization problems be found?), to data quality (i.e., what data can be obtained, and are the data reliable?), to consumer behavior and perceptions (i.e., do consumers strategize, and how are they influenced by their views of fairness and limited cognitive capacity?).

2.3. Supply Model

Like the revenue model, the supply model comes in many varieties. Nevertheless, there are three crucial decisions in the supply model regarding how the firm manages its resources (labor and capital): timing, location, and control.

The timing of a supply model refers to when the firm activates its resources relative to customer demand. At the most basic level, there are two choices here—either the firm activates resources before customer demand or after customer demand. If before demand, which is often referred to as "make-to-stock," then the challenge for the firm is to correctly anticipate demand. If after demand, which is often referred to as "make-to-order," then the challenge for the firm is to respond to demand in a sufficiently timely manner. In both cases, the goal is to have high utilization of resources while also generating customer value.

Location refers to where the firm places its resources relative to its demand. One extreme is a limited number of locations far from demand, which allows the firm to exploit economies of scale and low procurement costs (for inputs and resources). However, long distances to customers require additional transportation costs and shipping time. The alternative is many locations close to consumers, which reduces transportation costs and reduces shipping times, but also decreases economies of scale and increases procurement costs.

Decisions on control reflect the degree to which the firm can determine the quantity, quality, and timing of the resources used to serve customers. The highest level of control is achieved when the firm owns the resource. For example, if the firm owns a factory, then it can use the factory to produce when it wants. If the firm "rents" a factory, then the firm must wait for when it is available and pay the prevailing price at that time. A spectrum of control also applies to labor—the firm has greater control over the actions of employees than contractors. In general, the key trade-off in the control decision is between flexibility and cost—the greater the firm's need to directly choose when a resource is used, how it is used, and the quantity used, the more likely the firm is to opt for greater control, albeit at greater cost (usually because of lower utilization).

2.4. Innovation

Although every firm operates with some type of business model (whether they can articulate it or not), not every firm creates a *business-model innovation*, which is a business model that satisfies a novel need, has a novel revenue model, or implements a novel supply model. More often than not, a business-model innovation involves a combination of multiple novelties in needs satisfied, revenue generation, or supply processes.

Novelty often comes from challenging implicit assumptions regarding what product is important for a firm to offer or the process by which the firm offers its products. In doing so, a business-model innovation generally involves what could be referred to as a "smart sacrifice"—the firm explicitly chooses to perform poorly along one dimension of customer needs (the sacrifice) so that it can dramatically and substantially outperform the competition along another dimension of customer need (the smart choice). For example, before Amazon, customers had the need for a wide selection of books that could be physically inspected before purchase and obtained quickly. Amazon sacrificed physical inspection and immediacy, but dramatically improved upon the selection of available books.

New business models emerge in part because entrepreneurs try many new ideas, some of which turn out to be both novel and effective. But the fact that so many new business models have emerged over the last few decades is not entirely an accident. There are other trends that contribute to the creation of new business models. For instance, reductions in transportation costs enable a firm to consider a broader set of location options in the supply model, enabling both the movement of goods from far away and the quick movement of goods nearby. The growth of cities concentrates demand, which builds economies of scale in operations that yield lower costs. Increased standardization of physical goods (e.g., through precise machinery and the ability to communicate detailed specifications electronically) enables expanded options in terms of control—assets that used to have been owned can now be rented. The Internet allows retailers to expand their product offering and to implement complex pricing dynamics. Mobile devices dramatically increase the amount of data that can be exchanged and the speed of that exchange. All of these enhancements to telecommunication greatly improved transactional efficiency and opened up new sources of customer value. Et cetera. In sum, changes in living patterns and technology lead to cheaper and faster movement of people, resources, and information, which enable new combinations of supply and revenue models to serve new customer needs.

3. Business-Model Research

Business-model innovations challenge implicit assumptions, thereby revealing new ways of doing business that were assumed to not be feasible. In effect, they provide a window to a new world that was not known to exist. Not surprisingly, successful business models garner substantial attention in industry and academia. This section discusses the academic literature on a wide range of business-model innovations from the past few decades, including sponsored-search advertising auctions (Google), "name your own price" and opaque selling (Priceline.com and Hotwire.com), "selling for free" (Facebook), bundles and subscriptions (Spotify and Netflix), everyday low pricing (Walmart), limited markdowns (Zara), surge pricing (Uber), reselling (StubHub), make-to-order production (Dell), e-tailing (Amazon), and room-sharing (Airbnb), to name a few.

3.1. Revenue Model

There is remarkable variation in how firms earn revenue. As discussed in Section 2, the primary goal of the revenue model is to extract rents from customers to support the resources needed to serve them. The high-level decisions associated with the revenue model can be divided into three components: mechanism (what prices are chosen and by whom?), structure (sell or rent, and charge per use or fixed fees?), and dynamics (how often are the offered transaction terms changed, and what options are available to modify initial agreements?).

3.1.1. Mechanism. The simple posted price is probably the most common mechanism for establishing the price a consumer pays. It is easy to communicate and enables a quick and immediate transaction. However, it also requires that the firm determine the best price based on what the firm knows about its own supply and consumer preferences. When such knowledge is limited, the firm might make a significant error. An alternative to the posted price is some type of auction in which the price is endogenously determined.[2] With a well-designed auction, the final price can reflect the true preferences, thereby allowing the firm to extract more value. That said, an auction imposes significant transactional costs on consumers, such as the delay needed to gather auction participants, the uncertainty of knowing whether a transaction can be completed, and the time needed to participate in and monitor the auction. Thus, the auction mechanism tends to be preferred over posted prices when consumer preferences are sufficiently dispersed (Wang 1993). For example, auctions have recently been used to find the lowest-cost supplier throughout the world (Mukhopadhyay and Kekre 2002) and to match consumers selling their own goods to other consumers (e.g., eBay; Simonsohn and Ariely 2008). Among the greatest successes for auctions is with sponsored-search advertising markets, such as those that Google operates (Hosanagar and Abhishek 2013). Thousands of search queries can be submitted per second, and advertisers vary considerably in how much they are willing to pay at any given moment to associate their ad with the query. Auctions can be implemented quickly with modern technology and are superior at generating revenue in these environments relative to posted prices.

Given that posted prices and auctions each have their own limitations, one solution is to offer both of them to consumers: Etzion et al. (2006) demonstrate that offering both mechanisms simultaneously can increase revenue because it enables additional consumer segmentation. Alternatively, Priceline.com was the first to develop a hybrid mechanism, called name your own price (NYOP), in which consumers submit bids to a firm that are either immediately accepted or rejected (Hann and Terwiesch 2003, Fay 2004, Terwiesch et al. 2005, Spann and Tellis 2006, Amaldoss and Jain 2008, Wang et al. 2009). The firm usually implements a relatively simple strategy—assign a fixed threshold, accept all bids above the threshold, and reject all bids lower than the threshold. Although this initially seems just like the traditional take-it-or-leave-it single price, there are important differences. First, because consumers don't know the firm's threshold, the firm earns additional revenue whenever the consumer bids above the threshold. Second, other firms cannot easily observe the firm's threshold, which changes competitive dynamics.

A surprising extension of NYOP is "pay-what-you-want," which is NYOP in which the firm publicly announces a zero threshold for accepting bids—that is, the firm accepts any bid (Gneezy et al. 2012, Schmidt et al. 2015). Although consumers are allowed to bid zero, they might actually bid more because they want to be (or want to be viewed to be) prosocial individuals. As a result, pay-what-you-want can (surprisingly) generate revenue, and in some cases even more revenue than other mechanisms.

In NYOP, the firm keeps the price it will accept hidden from consumers, but consumers are told exactly what product they are bidding on. Opaque selling reverses what is known to consumers: With opaque selling, which is also called "probabilistic selling," the consumer is given a posted price for a virtual product that is a lottery over several real products. For example, a consumer could be offered a 3-star hotel in center city on a particular evening for $125, and only if the consumer accepts this offer is the actual hotel revealed. Hence, with opaque selling, a consumer runs the risk of being assigned the least preferred product in the set of possible choices. Fay and Xie (2008) explain how selling an opaque product can work to a firm's advantage. Say the firm offers two products and consumers vary in their preferences across the two products: Some really like one of them but not the other, and some are relatively indifferent between the two. The firm has a pricing dilemma: Price each product high to cater to the consumers who like the products while not selling to the indifferent consumers, or price each product low to try to capture sales from the indifferents. Neither is ideal. However, suppose the firm can create a virtual third product with which a consumer receives a random choice between A and B (i.e., a probabilistic product). The indifferents would rather choose the product than receive a random choice, but the other consumers really do not like the opaque product. Thus, the firm can set a high price to sell to the picky consumers and also sell to the indifferents via a lower (but not too low) price assigned to the opaque product. As a bonus, because an opaque product is not a real product, the firm does not have to invest in the development of another real product to cater to the indifferents. (See Section A.1 for a simple numerical illustration.)

In Fay and Xie (2008), there is a single firm that offers the third/probabilistic product. But it can also be helpful for an intermediary firm (e.g., Hotwire .com) to offer an opaque product that is a random choice between products from competing providers (Jerath et al. 2010). To explain, say an airline's customers can be divided into two types: loyal customers who only fly with that airline and bargain hunters who shop between different carriers. The airline might want to offer an opaque product choosing among its flights, but if its loyal customers know that they will fly with the airline even if they choose the opaque product, then it is hard to convince them to pay the higher, regular fare. If, instead, the opaque product is offered by a third party, who mixes in flights from other airlines, then the loyal customers are far less likely to consider that channel. This allows the airline to separate the segments: a high price to the regular customers and a lower price to cater to the bargain hunters via the opaque intermediary.

3.1.2. Structure. The structure of the revenue model is concerned with the terms of the transaction between the firm and the consumer. Two key dimensions of transaction terms are (a) ownership and (b) payment form.

Many transactions between firms and customers involve the issue of ownership. Ownership refers to the degree of control over a product or resource. For example, the owner of a product can use it when wanted and to the extent wanted. The owner is responsible for maintaining or upgrading the product, and the owner can choose to dispose of the product or to sell it to another owner. Many products are naturally sold to customers—that is, ownership is transferred. For example, a bakery only sells loaves of bread. It makes no sense for the bakery to lease or rent bread to customers. This is generally true with any disposable physical product. However, ownership need not be transferred with durable physical products. Say the firm makes mainframe computers. The firm could sell it to a customer, say, a bank. The bank would then be free to use the computer as much or as little as it wants. Over time, the computer depreciates, and it is up to the bank to decide when it might sell the mainframe on the secondary market or if it will use it to the point at which it has to be scrapped. Alternatively, the mainframe manufacturer could lease the computer to the bank. The upfront payment is far smaller, but then the bank makes periodic payments for the use of the mainframe. At the end of the lease, the mainframe is returned to the manufacturer, and it is the manufacturer's responsibility for salvaging any remaining value. This "sell versus lease" decision has been investigated extensively. Leasing can be better than selling for the manufacturer because it commits the firm to not flood the market in future periods, which would depress the residual value of the units—the bank will not pay a premium for a mainframe today if it knows that the price of the mainframe will fall steeply in the future because of additional production (Stokey 1981, Bulow 1982). However, there are situations in which firm may wish to both sell and lease (Desai and Purohit 1998).

In addition to ownership, the firm also must decide how to charge customers for the amount consumed. Although it is natural to charge a fee for each unit a customer uses, there is an alternative to per-use (or variable fee) pricing. A firm could offer a fixed price for a bundle of products from which the consumer can choose as many or as few as desired. For example, Spotify with a fixed monthly fee allows users to listen to any of the songs in their library. Given the breadth of genres in their collection, most consumers have little interest in the vast majority of their titles. Yet, despite the fact that consumers might be knowingly paying for many things that they never want, this extreme bundling strategy can be better for the firm than charging for each song (Bakos and Brynjolfsson 1999). To provide a simple illustration, say there are 10 products and consumers have independent uniform [0, 1] valuation for each of the products. If each product is sold individually, the optimal price would be 0.5; a consumer would purchase on average 1/2 of the products, and the expected revenue per customer is therefore $0.5 \times 1/2 \times 10 = 2.5$. Alternatively, the firm could offer the bundle of all 10 products at a price that matches its expected value, which is 5. In that case, half of the consumers would be willing to purchase the bundle, and the per-customer revenue is again 2.5—no improvement. But lowering the bundle's price to 3.9 has a dramatic effect—about 88.5% of customers are willing to pay 3.9 for the bundle of 10 products, thereby increasing the expected per-customer revenue to $0.885 \times 3.9 = 3.45$, which is 138% higher than selling the products individually. The firm can extract more revenue from customers via the bundle because consumers can be less price-sensitive regarding the bundle than they are for individual products.

Charging per product and charging for the entire bundle are two extreme strategies. The firm is, of course, not required to pick just one of those approaches—the firm can offer a menu of choices. In the mentioned example, the firm can do even better by giving consumers the option either to purchase a bundle for 3.9 or to purchase items for 0.88 each. Interestingly, the incremental revenue increases by a very small amount, from 3.44 to 3.47, which is only about 0.7%. This is not the only example in which a single, simple pricing option is not optimal, but nearly optimal relative to the more complex pricing menu (e.g., Cachon and Zhang 2006).

There are some limitations to bundling. For one, bundling can run into trouble if there is a nontrivial marginal cost to serve each customer. Continuing with the 10-product example, bundling and per-use yield the same profit when there is a cost of 0.216 per product served to a customer.[3] Higher marginal costs favor per-use pricing. Furthermore, the negative effect of marginal costs on bundling becomes more substantial as the number of products increases (i.e., the break-even marginal cost decreases). Nevertheless, although bundling requires low marginal costs to be the superior strategy, it does not require zero marginal costs. Next, bundling cannot perfectly segment across customers that differ in their breadth of preferences. Instead of the 10-product example, say, the number of products, n, is very large (e.g., $100 < n$). Half of the market has broad preferences in the sense that they have some value for all products. But the other half of the market is composed of consumers who only value a single product. In either case, when valuations are positive, they are uniformly distributed $[0, 1]$. If the firm could perfectly segment these two types of consumers, then the firm could charge the "broad" customers $p_n = 1/2n$ for the entire bundle of n products and earn on average 1/2 per product. To the "narrow" customers, the firm charges $p_1 = 1/2$ per product and earns on average 1/4 per product. Given that broad customers are half of the market, the average profit per product is 3/8. But this assumes the consumers can be perfectly segmented. Unfortunately, if the price is $p_1 = 1/2$ per product and $p_n = 1/2n$ for the bundle of n, then the broad customers do not buy the bundle. Instead, they choose to make their own bundle of half of the products, paying the per-product fee p_1, and the firm is back to earning 1/4 per product. Fortunately, there is a better solution for the firm—charge $p_1 = 2/3$ per product and $p_n = (4/9)n$ for the bundle. Raising the per-product price enables the firm to set a bundle price that earns some revenue and induces the broad customers to take it. Total revenue is now 1/3 per product, which is better than 1/4 per product, but less than the ideal of 3/8.

The airline industry illustrates that the choice between bundled and per-use pricing is not always straightforward. Traditionally, airlines bundled the price of the seat in the aircraft with the service of carrying a passenger's luggage. However, around 2007, some airlines began to offer separate fees for each product—that is, they started to charge baggage fees. Southwest Airlines resisted that trend (and still offers bundled pricing), but the majority of the other major airlines have transitioned to separate fees for each service (Nicolae et al. 2017).

Bundling can be applied across products, and it can be applied across time, in which case it is usually referred to as subscription pricing. For example, Blockbuster dominated the movie-rental business for a period of time with per-use pricing, but Netflix successfully introduced subscription pricing—customers could rent as many movie DVDs as they wished during the month for a single fixed subscription price. The advantage of bundling across time is analogous to the advantage of bundling across products—the user becomes less price-sensitive when considering a series of consumption opportunities (some of which are taken, some not) rather than considering them individually. For example, it can be better to sell a membership to use a gym over a period of time rather than to sell individual visits to the gym. The drawback of selling subscriptions is that consumers might overconsume the service—once a gym membership is paid and there is no fee for each use, a consumer is likely to use the gym even on occasions when the marginal value created is less than the cost of providing the service. This is particularly problematic when the offered product suffers from congestion—each additional user in the gym creates some inconvenience to the other people trying to use the gym. Even though in these situations it might seem that charging per-use fees would be better (to reduce excessive use), Cachon and Feldman (2011) demonstrate that subscription pricing still may be preferred. To explain, congestion can either be controlled via a per-use fee or through buying more capacity/resources. However, the latter requires additional revenue, and subscription pricing is capable of extracting more revenue from customers than per-use pricing. Hence, even in services subject to congestion (e.g., mobile telecommunication), it can be better to sell subscriptions than to charge per-use because it is better to buy more capacity than to regulate how consumers use capacity (see Section A.2 for details).

A surprising pricing strategy involves neither the per-use nor fixed fees of bundles nor subscriptions. Instead, the firm charges neither—that is, it offers its product for free. Clearly, firms cannot survive by selling *all* of their products for free. But when can a firm profit from selling some of its products for free? Alphabet (i.e., Google) and Facebook have been among the most valuable public companies (in terms of market capitalization), and yet the vast majority of their users pay nothing for their products. This remarkable achievement can be understood via the literature on two-sided markets (e.g., Rochet and Tirole 2006 and Parker and Van Alstyne 2005). For example, credit card companies charge merchants to process each transaction, but often earn no revenue at all from some customers. Even though value is created for both sides of the market, a positive price is charged to only one side. In some domains, this occurs because it is technically infeasible to charge one side—traditional radio and television stations broadcasting their content could not monitor (and therefore charge) consumers of their music or shows. The introduction of cable and Internet technology for distribution enabled consumers to be charged, and some companies adopted the new revenue model (e.g., HBO). However, even when it is feasible to charge both sides of the market, sometimes it is not done (as with Google and Facebook). The justification for the zero price in one market comes from the existence of externalities across the markets. A free price for using a credit card allows the firm to acquire a much larger base of credit card users than it would even if it charged a relatively modest price. The larger user base is far more valuable to merchants, thereby allowing the credit card company to charge merchants more than is lost from the absence of customer-generated revenue. Without the merchants' willingness to pay marginally more as the user base increases, there is no reason to sell for free to consumers merely to increase their numbers. That said, why stop at free? Why not pay consumers to join the user base? Once a platform jumps from "free" to "we pay you to sign up," it is likely that the incremental demand would be exclusively interested in the free money and of little value to the other side of the market.

The software industry has special versions of selling for free, one called "open source" and the other called "freemium." With open source, not only does a firm allow some customers to use their product for free, they allow the users to modify and enhance the product (Casadesus-Masanell and Llanes 2011). This strategy is successful when the code that is given away flourishes with independently developed enhancements, which creates a user base that is willing to pay for expertise on how to use the code effectively. In effect, like in the classic two-sided markets, one group of agents generates positive externalities that can be captured via another group of customers. With freemium, the firm allows consumers to use a limited-feature version of the software without time limit and also offers a premium version of the product for a positive fee (Cheng and Liu 2012). For example, Dropbox allows consumers free use of their service with a limited amount of storage for as long as a consumer wants. This is different than a two-sided market: There are few externalities between the two groups in Dropbox (free users and paying users). Instead, this model is justified by two assumptions: (i) Users will not try the product if they have to pay up front, nor will they be inclined to try a product that has a limited-time offer (why learn to use the product if you can use it for only a limited time?); and (ii) the desirability of the premium product may increase over time as familiarity with the product increases (Kumar 2014).

3.1.3. Dynamics. Part of a firm's revenue model involves its long-run policy toward the frequency and magnitude of adjustments to its transaction terms, which generally means its price. For example, does a firm develop a reputation for periodically offering temporary sales, or setting high prices that are later marked down, or offering advance-purchase discounts, or prices that swing up and down throughout the day? All of these cases have been implemented in practice and researched extensively.

Some retailers are known for periodic deep discounts, which has been called "hi–low" pricing. Theory suggests that this can increase revenue by segmenting consumers either in their willingness to wait for a discount or in their willingness to carry inventory (Blattberg et al. 1981, Ho et al. 1998).[4] To give a simple example, say a retailer sells cans of soup and has two types of customers. The "high types" are willing to pay \$4 per can, want one can per week, and do not want to carry inventory in their home. The "low types" pay at most \$2 per can, want one can per week, and are willing to buy as many as five cans at a time. Say there are 100 of the high types and 200 of the low types. If the retailer always charges \$4 per can, then the retailer earns \$400 in revenue per week selling only to the high types. If the retailer always charges \$2 per can, then the retailer's revenue is \$600 per week selling to all consumers. But if the retailer has a regular price of \$4 per can and a discount price of \$2 every 5 weeks, then the retailer earns \$400 per week from the high types when the regular price is offered and \$2,200 in the promotion weeks (because the low types buy five cans to carry them to the next promotion), for an average of \$760 per week. The hi–low strategy earns considerably more (27%) than the best nondynamic pricing policy.

Segmentation via hi–low pricing can work, but it runs the risk of not being popular with consumers, and it might add to operational costs due to higher demand variability (Lee et al. 1997). Some retailers have attempted to adopt a "value-pricing" or "everyday-low-pricing" (EDLP) approach that eliminates promotions or at least reduces their frequency and depth (e.g., Walmart). However, the empirical evidence tends to lean in favor of hi–low pricing, at least when the focus is on short-term revenue (Hoch et al. 1994).

With seasonal goods, especially with fashion items that have short-lived preferences due to shifting tastes, a standard pricing model begins with a high price, and then discounts are offered toward the end of the season if inventory remains that needs to be salvaged for something. Hi–low pricing in this context is sometimes referred to as a "price-skimming" strategy. The declining price path is used to try to sell to the consumers with a strong preference (if they buy at the high price), while also ensuring that some sales are made through the lower prices at the end of the season if it is necessary to sell to consumers with weaker preferences (e.g., Lazear 1986 and Besanko and Winston 1990). This strategy works very well if consumers myopically purchase whenever they see a price they are willing to pay. Unfortunately, consumers might be strategic, avoiding the high initial price because of the expectation that the item can be purchased later on at a lower price.[5] In fact, that behavior can be self-reinforcing—if all consumers refuse to purchase at the higher price because they expect a lower price to follow, then the sparse initial sales indeed force the firm to discount later on. Retailers tend to have an opinion on this behavior: A CEO of Best Buy once described the former (myopic) consumers as "angels" (because they purchase at full price) and the latter (strategics) as "devils" (because they tend to buy at discounted prices; see McWilliams 2004).

The consequences of strategic consumer behavior and the appropriate firm response have been debated extensively by both practitioners and academics. Some firms, such as Zara, have found success by severely curtailing discounting (Hansen 2012), whereas other firms (e.g., JC Penney) have attempted to avoid discount pricing but found disappointment (Clifford and Rampell 2013). A (somewhat) simple example highlights the potential benefits and pitfalls of these strategies (Section A.3). One firm buys q units before a selling season and pays \$50 for each unit. The product is sold over two periods. A random number of consumers arrive at the start of the first period. That demand can be described with a gamma distribution with mean 100 and standard deviation 100. All of those consumers are willing to pay up to \$100 for the product. Half of them are myopic—they buy in the first period in which they see a price less than \$100. The other half are strategic—they purchase in the period that gives them the better deal.[6] The strategics know that they are more likely to be able to find the item available for purchase in period 1, but they also know that the price is lower in period 2. In particular, if the firm has units remaining at the start of period 2, then, because there is an unlimited number of bargain shoppers in period 2 willing to pay \$30, the firm can lower the price to \$30 to ensure that all of the inventory is sold.

The firm has four strategic options. With hi–low pricing, the firm chooses the highest possible initial price, $p_1 = \$100$, and then discounts if inventory is remaining in period 2, $p_2 = \$30$. The strategics get no value from purchasing in period 1, so this approach focuses on selling to the myopics, and the strategics are relegated to discount shoppers. With value pricing, the firm chooses some intermediate price to try to get the strategics to purchase in period 1—that is, $\$50 < p_1 < \$100$—and then discounts in period 2 to $p_2 = \$30$ if inventory remains. The strategics can be willing to pay a bit more in period 1 because waiting involves the risk that the item might be out of stock in period 2. A third approach is "never discount": $p_1 = p_2 = \$100$. The strategics buy in period 1, but leftover inventory is totally scrapped.[7] Su and Zhang (2008) refer to this as price commitment. The fourth approach is a "refund" strategy (Lai et al. 2010): set a high initial price, $p_1 = \$100$, discount to $p_2 = \$30$ in period 2 if there is any inventory left over, and promise period 1 customers that they will receive a refund equal to the full price difference, $p_1 - p_2 = \$70$, if a discount is taken. The refund eliminates the risk from buying in period 1. Hence, all of the strategics are willing to buy in period 1, even though the initial price is very high.[8]

Table 1 displays the results of the four strategies. Value pricing is clearly inferior—it does attract the strategics to purchase at the regular price, but because the regular price is less than what the myopics will pay, the firm's profit suffers overall. Hi–low and refunds perform similarly, even though they are clearly very different strategies—with hi–low, the strategics don't buy at the regular price, but with refunds, the firm needs to be conservative with its order quantity to avoid the risk of having to give a discount to everyone. The best approach is the one that does not alter the price across the season, the never-discount strategy—the firm incurs the significant cost of being unable to salvage any leftover inventory, but it has the benefit of forcing the strategics to buy at the regular price.

It is dangerous to conclude too much from a single example, but in a broader set of scenarios (reported in Section A.3), it is found that never discount is most frequently the optimal strategy, and it is robust (i.e., even when it is not optimal, it is nearly optimal). Similarly, value pricing is never optimal, suggesting that it is ill-advised to pander to strategics. (That said, the story on value pricing is not complete until the supply model is considered.)

The strength of the no-discount strategy raises the question of whether a firm should even use dynamic pricing—there is nothing dynamic about the no-discount commitment that maintains a single price. However, this is not the only reason managers may hesitate to use dynamic pricing. Another reason arises from the general perception that dynamic pricing reduces consumer surplus. Given that dynamic pricing involves "high" and "low" prices, it seems intuitive that high prices are used to extract rents from consumers, thereby lowering their surplus. This is particularly salient if it is assumed that the total value in the system is a fixed amount—if the firm gains by a higher price, surely consumers must lose. In the short-term, this shouldn't be a direct concern for managers. But in the long-term, it could lead to a permanent loss of customers. And that is a concern. Interestingly, Chen and Gallego (2018) show that when demand is stationary but costs are stochastic, consumers are indeed generally better off when the firm implements dynamic pricing.

Table 1. Pricing Strategies for a Fashion Product with Uncertain Demand

Strategy	q	p_1	p_2	Profit
Hi–low	62.6	100	30	1,247
Value	81.7	75.3	30	893
Never discount	69.3	100	100	1,534
Refund	52.0	100	30	1,124

The questions of whether dynamic pricing should be used and who benefits or is harmed are central to the recent rise of ride-sharing platforms like Uber. Uber matches independent drivers to riders seeking local transportation. For Uber, the central challenge is not strategic consumer behavior—many consumers are unable to strategically time their need for transportation. Rather, the first-order challenge for Uber is to match supply with demand as it learns information about demand and supply conditions. Uber has a finite pool of drivers at any moment, but its demand can vary wildly from day to day and hour to hour, based on local events and weather, among other sources of variability. Too many drivers on the road is a problem because then they are idle more than they want (which destroys value). Too few drivers on the road is another problem because then customers wait too long and eventually become disenchanted with the service (which destroys value). Uber's innovation is to implement dynamic pricing, which they refer to as surge pricing, with its demand (customers) *and* with its supply (drivers). Relative to a pricing regime with fixed prices and wages (e.g., a traditional taxi model), the optimal surge pricing strategy is much better for Uber. Surprisingly, dynamic pricing can also be much better for the drivers (even though they get paid less during slow periods) and consumers (even though they have to pay much more during peak periods) because dynamic pricing enables an overall increase in supply, which increases the potential value generated in the system (Cachon et al. 2017).[9]

In many dynamic pricing settings (groceries, fashion apparel, and ride-sharing), transactions and delivery of the product occur approximately at the same time. However, some firms provide customers with the opportunity to commit to purchase their product well in advance of actually using the product. For example, this is customary, and expected, in the travel industry (reserving hotels, car rentals, airline flights, etc.) This selling strategy works for several reasons. To begin, some customers value knowing that they can plan to have access to the product at the designated time, and a portion of this value can be captured by the firm (Gale and Holmes 1993, Alexandrov and Lariviere 2012). Second, advance selling allows for segmentation across customers that differ in their value for the product based on when they arrive to the market. The classic story is that price-sensitive customers arrive early to the market (e.g., leisure travelers), and price-insensitive customers arrive late (e.g., business travelers). Advance selling allows the firm to sell to these segments at different prices (Su 2007). A third reason for advance selling is less intuitive—customers may be willing to purchase before they are certain of their value for the product (Xie and Shugan 2001, Chu and Zhang 2011). To illustrate with a simple model, suppose a customer can purchase the product either "in advance"—that is, well before consumption—or "on the spot"—that is, just before consumption. In advance, the customer is uncertain regarding her value for the product, but knows that her value is uniformly distributed between 0 and 1. On the spot, the consumer learns her value. On the spot, the firm sets $p = 1/2$, which maximizes revenue if spot consumers have valuations uniformly distributed $[0, 1]$. The consumer anticipates that she earns a surplus of 1/8 if she purchases on the spot and the firm earns 1/4.[10] But if the firm sets an advance price of 3/8, then the consumer prefers to purchase in advance (before knowing her value), and the firm earns 3/8, which is 50% higher profit than selling on the spot.

It is remarkable that advance selling can be favorable to the firm, even in the absence of availability issues (i.e., no capacity constraint) or segmentation opportunities (i.e., consumers are homogeneous). Advance selling works, despite the cost it imposes on consumers, because consumers are more homogenous in the advance period (they all have expected value of 1/2) than in the spot period (valuations are distributed $[0, 1]$). It can be better to sell to a population of homogeneously valued consumers, even if their valuation for the product is less than 1/2 of consumers in a heterogenous population. This is essentially the same reason for the effectiveness of bundles across products (e.g., a single fixed price for access to a library of music) or across time (e.g., subscriptions).

Although advance selling can work in the firm's favor, because of the gap in time between the transaction agreement and the transaction delivery/completion, it is possible that one or more of the parties in the transaction learns new information that motivates an adjustment to the transaction terms. When this possibility can be anticipated, the firm can explicitly incorporate a particular recourse into the transaction agreement. For example, the firm can offer partial refunds to consumers who later learn that their valuation for the product is low. If the return/refund is completed with enough time before consumption, the firm then has the opportunity to sell the returned product to another consumer (Xie and Gerstner 2007, Guo 2009, Gallego and Sahin 2010). The firm could also try to sell to another customer, even if consumers are not allowed to return the product. In effect, the firm tries to sell its capacity twice, a practice known as overbooking. If the firm does successfully sell the unit to another customer, then the initial customer is denied service. In the context of airlines, this is called "bumping" a passenger off of a flight. Although the initial customer might be disappointed, the possibility of being denied service and the resulting compensation are known in advance. Finally, rather than being involved in a possible ex-post transfer of capacity from one customer to another, the firm could allow customers to resell the product (Su 2010, Cui et al. 2019). Cachon and Feldman (2018) find that a seller can improve its revenue considerably with any of the three recourse strategies (refunds, overbooking, or reselling), but reselling is generally the best. Reselling is most effective because it helps to ensure that the consumer who values the product the most is the one who uses the product. Interestingly, this additional value is equally shared between the firm and consumers, so both benefit from the practice, which is consistent with the considerable growth in ticket exchanges like StubHub. However, this is predicated on the assumption that the reseller actually earns some value from reselling. If the reseller does not earn anything, then neither does the firm. For example, "reselling" in digital music is often called "piracy." It is intuitive that piracy can harm the firm (Rob and Waldfogel 2006), but there is also the possibility that sharing facilitates consumer search, which can be helpful (Zhang 2018).

3.2. Supply Model

The main function of the revenue model is to extract value, and the main function of the supply model is to generate value. There are three high-level strategic decisions that the firm must make with respect to how it designs its supply: timing (when are resources activated relative to the arrival of demand?), location (where are resources positioned relative to demand?), and control (to what extent is the firm able to decide when and how resources are utilized?).

3.2.1. Timing. A firm needs to decide if it begins work in anticipation of demand or if it waits for demand to arrive before starting service. The firm's positioning on this timing spectrum determines whether its resources wait for customers or whether its customers wait for resources. In the early years of the personal computer industry, the key insight was that it was better to make customers wait. That is not always the case.

Some form of waiting invariably occurs because resources are relatively rigid, whereas demand is volatile. Two examples illustrate this point, one in which demand is volatile in predictable ways (seasonality) and the other in which randomness is the cause.

Most firms face some form of seasonal demand. Say there are two seasons, a low season and a high season; the seasons alternate in occurrence; each season lasts for τ time; and demand in season $s \in \{l, h\}$ occurs at constant rate λ_s, $\lambda_l < \lambda_h$. Let $\lambda = (\lambda_l + \lambda_h)/2$ be the average demand rate, and let $\delta = (\lambda_h - \lambda_l)/2$ be a measure of the amplitude of seasonal variation. The firm chooses a constant production rate, μ, and incurs a constant cost rate of $c\mu$ to maintain that capacity, where (without loss of generality) let $c = 1$.

The firm's capacity should be selected in the range $[\lambda, \lambda + \delta]$ so that it has enough capacity to satisfy all demand, but not so much that a portion of it is never used. Hence, the firm has sufficient capacity during the low season to satisfy demand, but does not have enough capacity for the high season. There are essentially two extreme options for dealing with the high season. First, the firm can do some work in advance of the high season. For example, a toy manufacturer might build and inventory completed toys before the fourth quarter, or a restaurant might prepare its ingredients before the evening rush so that dinners can be cooked more quickly once orders are received. Either approach creates inventory-carrying costs: Let h be the cost incurred per unit of work per unit of time the work is completed before its demand arrives. Second, the firm can make some high-season demand wait—that is, create a queue of demand that is satisfied when capacity becomes available (i.e., during the low season). This, too, is costly: let w be the cost per unit of work that is in the queue per unit of time. And as in most cases when presented with extreme options, there exists an intermediate approach that combines the use of doing work before the season starts (inventory) with doing work after the season begins (queues).

Volatility is never helpful to the firm: Costs increase in both the amplitude of the seasonal variation, δ, as well as the duration of the seasons, τ. (See Section 3.2.1 for details.) To cope with that variability, the firm can either just use inventory or just make customers wait, or some blend of the two. Interestingly, one of the two extreme strategies tends to be optimal. To be specific, the intermediate strategy that blends both inventory and queues is optimal only if $1/3 < w/h < 3$, otherwise, one of the two extreme strategies is preferred. For example, the firm's best solution is to only use inventory whenever $3 \le w/h$: If the cost of waiting is more than three times the cost of holding inventory, the optimal solution is to only build inventory. Analogously, if $w/h < 1/3$, then it is best to only use a queue. Neither threshold is particularly extreme, which suggests that in practice a firm is likely to take one approach or the other, but not a mixture of both. For example, toy manufacturers build inventory before their high season, whereas aircraft manufacturers make their customers wait.

Now consider the timing question in an environment with stochastic demand. Say a firm sells N physical products that can be inventoried. The lead time to produce each product is l (with no capacity constraint). Total demand occurs at rate λ, and the i^{th} product demand rate is $\lambda_i = \lambda/N$. Interarrival times for demand are exponentially distributed. Again, let h be the cost to hold inventory per unit of time and w the cost to make a unit of demand wait one unit of time.

Each product is managed with a base stock policy. When demand is sufficiently slow, the boundary between inventory (make-to-stock) and queue (make-to-order) occurs between a base stock level of $S = 1$ (make-to-stock) and $S = 0$ (make-to-order). Making customers wait ($S = 0$) is better when

$$l\lambda/N < \ln(1 + h/w). \tag{1}$$

So, make-to-order is more likely to be the better strategy when the lead time is low (small l means customers don't have to wait too long), there are many products (large N), holding costs are high (large h), and customers are relatively patient (low w).

To get a sense of how likely (1) is to be satisfied, let's evaluate plausible values. First, consider h, which includes the opportunity cost of capital, some accounting for obsolescence or spoilage costs, and storage costs. To estimate the latter, consider 2016 data from Kroger (a large grocer in the United States). Assume retail space costs about \$200 per m^2 per year, which is approximately the cost of commercial real estate. Their annual cost of goods sold was about \$90 billion, their average inventory was about \$8 billion, and they operated with about 18 million m^2 of retail space. This implies they have on average \$8B/18M m^2 = \$444 of inventory per m^2 of retail space. Storage costs alone are thus \$200 per m^2 per year /\$444 per m^2 = \$0.45 per dollar of inventory per year, or 45% of the value of the inventory on an annual basis. Add in the opportunity cost of capital and other inventory storages and maintenance costs, and Kroger could easily have an annual holding cost of about 60% (i.e., to hold \$1 in inventory for 1 year costs them \$0.60). A similar evaluation for other retailers demonstrates that an annual holding cost of 60% is reasonable and often is conservative.[11] Now consider the waiting cost. Plausible values for w (cost per day) could range from 1% to 20% of the product cost: w = 5% implies that the daily cost to make a customer wait equals 5% of the cost of the item, which means that after $1/0.05 = 20$ days, the consumer's waiting cost equals the product's cost.

Using the plausible ranges for h and w, Table 2 evaluates the maximum demand rate that justifies make-to-order—that is, the λ that makes (1) bind with equality. The resulting demand rates appear small, ranging from 0.004 units per day to 0.305 units per day. But note that those rates are for the individual products. If a firm sells 100 units per day, and offers 10,000 different variants of the product, then the demand rate per product is 100/10,000 = 0.01, which is comparable to the thresholds in the table and lower than all of the thresholds with the two higher holding costs. One can conclude that make-to-order (i.e., queues of customers waiting for resources) combined with broad variety can be the best strategy, especially when waiting costs are not extreme. That insight led Dell Computer to become a dominant player in the personal computer industry.

Although the issue of timing usually involves the boundary between the firm and the customer (as in the previous examples), the firm makes strategic decisions regarding timing in other domains. For example, it has been established that a firm can reduce its investment in inventory, while not changing the assortment of products available to consumers, by delaying the process that adds variety to the product as late as possible in the supply chain. This strategy is called "delayed differentiation" or "postponement" or "localize remotely" (Lee and Tang 1997).

Timing is also important in the decision of which products to offer—the closer in time product-design decisions are made relative to when products are sold, the more likely for better product designs to be selected. This is a crucial component to Zara's business-model success—they are able to take a design concept and deliver a new product in a matter of weeks, rather than months (or many months), which means they are more likely to offer desirable fashions (Cachon and Swinney 2016). But even Zara still produces inventory before demand. The next level is to decide which products to produce only after demand is realized. The extreme version of this is called mass customization—customers request products customized to their preferences, and the firm then delivers on this quasiinfinite assortment. Customers receive the product they want, but they need to wait for it. Furthermore, this product strategy has implications for competitive pricing dynamics (Mendelson and Parlaktūrk 2008).

Table 2. Maximum Demand Threshold per Product to Support Make-to-Order

w	h = 20%	h = 60%	h = 100%
1%	0.069	0.194	0.305
5%	0.014	0.042	0.069
10%	0.007	0.021	0.035
20%	0.004	0.011	0.018

Note. h, annual holding cost rate; w, daily wait cost; l, 1; 280 days per year.

3.2.2. Location. Even in a world filled with "cloud" technology, most resources exist in some physical location. Resources close to consumers can provide faster service. But moving resources closer to consumers requires dispersing resources into more locations, with each location likely serving less demand. For retailers the basic trade-off is between scale and responsiveness—consumer demand can be more quickly satisfied with nearby resources, but then more numerous locations leads to a loss of scale, which reduces operating efficiency (Allon and Gurvich 2009).

Although a trade-off exists, firms have been successful at different ends of the location spectrum. Zipcar changed the car-rental industry by moving cars closer to consumers, so close that they could walk to the car. Aldi operates small grocery stores (e.g., 1,000 m^2), again, closer to consumers than most of its competitors. In both of those examples, and others, the more numerous/closer locations comes with a price—to somewhat mitigate the negative effects of lower scale, each of these firms reduces the variety of products offered (relative to competitors who operate with fewer locations serving greater demand).

Amazon took a different approach. The initial Amazon model eliminated the nearby bookstores and replaced them with a single location for the entire country. This eliminated the expense of storing inventory in costly retail locations and paying for store employees. But shipping books to customers is costly and takes time—whereas the neighborhood bookstore could provide a book within hours, Amazon could only do it within days.

To illustrate the potential for the Amazon model, Table 3 displays the differences in costs for products that have different weekly demand rates in two operating models, e-tailing and the traditional brick-and-mortar store. In both cases, the product has a 1-week lead time, the target in-stock probability is 0.99, and orders are placed weekly. The holding cost for traditional retailing is 60% of the product's value (as evaluated in Section 3.2.1), whereas e-tailing's holding cost is 40% to reflect the lower cost of warehouse space. However, e-tailing must incur 12% in fulfillment costs. (Amazon's fulfillment costs ranged from 10%–13% of the product's value from 2011–2016.) According to the table, the best model for a product depends considerably on its demand rate. If a product's total demand (across the area covered by 500 traditional stores) is fewer than 40 units per week, the e-tailing model dominates. For example, a product with total demand of 10 units per week would achieve inventory turns of only 1.1 when that demand is spread across 500 stores. The resulting cost is 56% of

Table 3. Operating Cost Differences (Measured as a Percentage of the Product's Cost) Between E-Tailing and Traditional Brick-and-Mortar Stores

	1 e-tailing warehouse		500 traditional stores		
Weekly demand rate	Annual turns	Total cost (%)	Annual turns	Total cost (%)	Best model
10	32.5	13.2	1.1	56.0	E-tailing
20	41.6	13.0	2.2	27.2	E-tailing
40	49.5	12.8	4.7	12.9	—
80	59.4	12.7	10.3	5.8	Traditional
160	68.2	12.6	24.2	2.5	Traditional

Notes. Assumptions: Annual holding cost rate for traditional store is 60%, and for e-tailing it is 40%. E-tailing incurs a fulfillment cost equal to 12% of the product's value. In both models: one-week lead time; weekly ordering; 0.99 in-stock probability target.

the product's value, which is more than four times the cost that would be incurred via the e-tailing model (13.2%). However, when the weekly demand surpasses about 40 units per week, the traditional model begins to do better because it avoids the expensive cost of fulfillment. In sum, if a product's turns are about 4.7 of greater when sold through a traditional store, then the traditional store is the better model. But if the traditional store is unable to achieve 4.7 turns, possibly because the demand is too low when spread across many stores, then the e-tailing model is better, potentially considerably better. This raises the question of whether there are enough slow-selling products for an e-tailer to have a decent business. In fact, Brynjolfsson et al. (2003) show that a considerable portion of Amazon's sales indeed come from slow-selling items—even though any one item has relatively small demand, the sum across millions of these items can be significant. (And, recall, a slow-selling item for Amazon with its limited locations would be an extremely slow-selling item for a retailer with many locations.)

For manufacturing firms, the scale-responsiveness trade-off continues to exist, but they also face a trade-off between production costs and responsiveness. Manufacturing in fewer, more distant locations gives the firm more options to find locations or suppliers with the lowest production costs. Nike's innovation was to move the manufacturing of its shoes from the United States to lower-cost Asian countries. Many other firms have followed a similar pattern as they seek to lower their costs. However, manufacturing in Asia while selling in North America generally requires shipping product by sea, which results in long lead times and lower responsiveness. (Air shipping is feasible only for products that have a high value-to-weight ratio.) Given past energy prices, the cost of transportation is of second-order importance relative to typical labor-cost differentials. This could change in the future, which would motivate a shift in strategies.

Zara's innovation was to emphasize responsiveness over costs. Zara manufactures in Europe and in nearby countries, which increases its manufacturing costs, but also gives it quick-response (QR) capabilities. With QR, a firm is able to make a small initial production commitment, which limits the chance the firm needs to mark down leftover inventory. At the same time, QR gives the firm the ability to respond later in the season with additional inventory if demand is strong—the firm can use QR to avoid the pitfall of excess inventory while also avoiding the opportunity cost of failing to respond to high demand. Even with exogenous demand, QR has been shown to be a very effective strategy (Iyer and Bergen 1997). But the true genius of the Zara model becomes apparent when one considers the impact on strategic consumer behavior.

Consider again the dynamic pricing model from Section 3.1.3. The one change is now that the firm has QR capability to place a second order, after observing period-1 demand. This second order becomes available at the start of period-2, but each unit in that order costs about 25% more. The advantage of the second order is that it is made with better information regarding demand. The disadvantage is that it costs substantially more.

Without QR, the firm's best strategy is to never discount, earning a profit of $1,534. With QR, the firm's best strategy is to adopt value pricing (which is never optimal without QR), with an initial price of $94 and a modest initial production quantity of 49.0 units. If demand turns out to be weak, units are discounted to $30, but given the small initial production, consumers know that a markdown is not likely (0.22 probability). However, if demand turns out to be strong, additional units can be purchased and sold at the profitable price of $94 without incurring lost sales. Combining these features yields an expected profit of $3,412, which is 222% greater than the profit with the best non-QR strategy. Cachon and Swinney (2009) find a similar result in a related, but broader, model. But the real confirmation of these ideas comes from Zara—their decision to locate manufacturing closer to their primary demand resulted in higher production costs, but gave the company a very powerful tool to avoid discounts, which led consumers to realize that there is little to gain from strategic behavior, which enabled Zara to maintain high margins, earn considerable profit, and grow to be the world's largest fashion apparel retailer.[12]

Although we have made some progress to understand the Amazon and Zara business models, the issue of location is by no means fully explored. For example, as Amazon has grown, and as customers have developed a taste for fast shipping, Amazon has begun to operate with more locations, which moves its inventory closer to consumers. Consequently, a customer might order items that are stored in different fulfillment centers, which requires Amazon to decide between sending all items in the order to one location so that they can be combined into one shipment to the customer, or should Amazon send multiple shipments to the customer (called split shipments) (Acimovic and Graves 2015)?

Although Amazon remains largely a pure e-tailer, many traditional retailers now operate both physical stores as well as an online store (e.g., Warby Parker for eyeglasses and Bonobos and Indochino for apparel). This creates a host of coordination issues between the channels. For example, some retailers now offer the customers the option to "buy-online, pickup-in-store" (BOPS). One might assume that this additional functionality would increase online sales. But in fact, Gallino and Moreno (2014) find that BOPS can decrease online sales while increasing sales overall because consumers use the online channel to do product research while using the store to do final product inspection. And Gallino et al. (2017) demonstrate that this capability increases sales dispersion across the product assortment, which can lead to higher inventory investments. Instead of using the online channel for product information and the store for physical fulfillment (BOPS), a retailer could reverse those roles: use physical showrooms to provide product information and the online channel to provide fulfillment (Bell et al. 2017).

Finally, there is growing interest on how location decisions influence environmental sustainability (e.g., Cachon 2014, Kabra et al. 2015) and pricing (Bimpikis et al. 2016, Besbes et al. 2018).

3.2.3. Control. Firms need a collection of resources to generate value. Each resource can be placed on a spectrum that describes the degree to which the firm has control over the resource. At one extreme, the firm "owns" a resource and therefore can dictate when, where, and how often it is used. At the other extreme, a firm "rents" a resource. In that case, the firm makes requests to the owner of the resource for its use, but use can only occur when an agreement is made between the two.

There are several benefits to ownership. To begin, it facilitates the coordination of resources. For example, when two resources are owned and need to be used concurrently or sequentially, then owning both allows the firm to guarantee that this can be done. For example, Zara owns its trucks used to deliver product from its fulfillment center to its stores. Consequently, clothing can be packaged in the truck that facilitates the store-employee process of unloading the clothes in the store to ready them for display and sale (e.g., the clothes arrive unwrinkled). Furthermore, truck deliveries can be timed when it is best for the store employees to receive (i.e., at times that are not crowded with customers), rather than when it is convenient for the third-party carrier to deliver. Second, ownership facilitates the standardization of a process. There are two benefits to standardization. First, customers may value standardized processes because the delivered quality has little variation over time or locations. Second, standardization allows a firm to purchase less-expensive resources. A machine that does a single, standardized task is less expensive than a machine that must be flexible. Wages for workers to complete routine/standardized tasks are lower than wages for workers who do varied/unstandardized tasks. This occurs in large part because unstandardized tasks require more judgment on the part of the employee, which requires an employee with greater skill and training.

The challenge with ownership is utilization. A firm's need for a resource can be volatile, and volatile needs lead to poor utilization, which increases the cost of the resource. Zara's outbound delivery to its stores are completed with a full truck, but then trucks return mostly empty. A third-party carrier's trucks are loaded a higher fraction of time, thereby lowering the cost of the asset. Employees' skills may not be needed all of the time, but the firm may be responsible for their salary throughout the year. The advantage of renting a resource is that the firm pays for the resource only when needed, assuming reasonable terms can be agreed upon with the renter. Hence, the key strategic decision for the firm is to get the correct balance with its resources between the virtues of control/ownership and the flexibility of renting.

Initial research on control focused on the theory of the firm boundary—that is, which of a firm's resources are included inside the firm and which are outside the firm. This can also be thought of as the "outsourcing decision," which is often referred to as the "make–buy" decision. The main insight is that there are frictions both within the firm (e.g., agency costs, i.e., workers/managers need to be compensated and may not have incentives that are fully aligned with the owners of the firm) and between firms (e.g., search and coordination costs due to incomplete contracts). The equilibrium structure is likely to make the best trade-off between these costs (e.g., Grossman and Hart 1986 and Grossman and Helpman 2002). And the balance in this trade-off tips toward more outsourcing when technology reduces search costs outside of a firm (Malone 1987, Brynjolfsson et al. 1994) and enables more complete contracting between firms (Cachon 2003, Cachon and Lariviere 2005).

Once a firm has decided to rent a resource, the firm must manage its acquisition of the resource. For instance, Van Mieghem (1999) demonstrates that it is not always in a firm's interest to try to specify and account for a priori all contingencies in the relationship between two firms—sometimes it is in their interest to decide their terms after observing information. Belavina and Girotra (2012) explain why even a large firm (e.g., Walmart) might choose to source through an intermediary (e.g., Li & Fung) rather than directly with suppliers. They demonstrate that an intermediary is better able to create long-run relationships with suppliers than the firm itself. Long-run relationships mitigate some of the frictions of disintermediation (e.g., supplier shirking), so a supply chain with an intermediary can generate more value than a supply chain without one, which justifies the existence of the intermediary. Ang et al. (2017) explore not only the boundary between two firms but, rather, the structure of the supply chain. In particular, they investigate how a firm should best mitigate the risk of production disruptions when managing its suppliers based on how the suppliers source one level higher in the chain from their own suppliers.

The previous examples generally consider the control boundary between a firm and a few other firms. The rise of the "sharing economy" creates situations in which the typical "one to few" relationship has become "one to many," illustrating an extreme penchant for renting over ownership. For example, Uber provides transportation without owning the vast majority of the vehicles actually used to provide the service. Nor are Uber drivers Uber employees. Instead, they are independent contractors. Consequently, Uber has limited control over when they work and how they work. Airbnb is another new firm that eschews owning assets. They provide a platform in which individuals (called hosts) can offer their property or a portion of their property (e.g., a single room) for rent to others. Airbnb is not able to specify the design of the property, nor when it is made available, nor the price the host charges. Hence, Airbnb needs to understand how its hosts operate and then how it can subtlety influence them. For instance, Cui et al. (2019) study the degree to which hosts exhibit discriminatory behavior. Li et al. (2016) study how hosts price and how they could price better if they had more experience.

A surprising feature of the success of Uber and Airbnb is that their strategies starkly contrast those of previously successful companies. For example, McDonald's became a dominant player in the (quick-serve) restaurant industry because it was able to standardize many of its processes—how its beef was raised, how its potatoes were grown, and how its hamburgers were cooked. Zara is able to provide timely fashions in large part because it has vertically integrated its supply chain to a larger extent than its competitors, thereby enabling greater coordination of its assets. In contrast, Uber and Airbnb have taken the approach to reduce control over their resources. This has enabled them to increase utilization of the resources (which lowers costs), but it sacrifices the benefits of standardization. Airbnb guests have more uncertainty as to the quality of their stay, and Uber customers are never sure exactly what type of vehicle they will ride in. What is remarkable about these companies is not that there are negative consequences to renting, but, rather, those negative consequences can either be managed or are small enough relative to the benefits of renting.

4. Conclusion

A firm's business model defines how it delivers (supply model) and profits from (revenue model) the customer value it creates through the effective satisfaction of needs. The revenue model involves high-level decisions regarding the mechanisms for determining prices (e.g., posted prices versus auctions), the structure of the transaction terms (e.g., per-use or fixed fees), and the dynamics of adjustments to transaction terms over time. The supply model requires decisions on the timing of resource activation, the locations of resources, and the level of control over those resources. The combination of these decisions determines the amount of customer value generated and, ultimately, the viability of the firm.

A business-model innovation generally involves one or more novelties in the revenue or supply models (or both). These novelties often challenge implicit assumptions, assumptions that are so taken for granted that they would be omitted from a list of key assumptions. For example, traditional fashion-apparel retailing involved outsourcing production to overseas suppliers and a hi–low pricing model that heavily relies on markdowns to clear excess inventory. Zara's fast-fashion business-model innovation begins with its supply-model decisions on location (local production) and control (vertical integration). These enable Zara to experience far fewer cases of overproduction because (a) design decisions are made closer to the selling season (an innovation in the timing component of the supply model) and (b) its quick response capabilities allow it to be conservative with initial production commitments. Consequently, Zara also innovated with its revenue model, adopting a value-pricing approach that uses markdowns sparingly. Like Zara, Amazon's e-tailing business model is grounded in the location decision of its supply model. But instead of moving closer to demand, Amazon moved further away, initially choosing to stock inventory in a single warehouse rather than in thousands of retail stores close to consumers. It turns out that many customers were willing to give up the ability to physically browse books (the traditional assumption) so that they could have access to a vastly greater breadth of books. More recently, the ride-sharing market has challenged our presumptions of urban transportation, which was dominated by taxis. Uber's entry to this market involved two key business-model innovations. First, it relinquished control of two key assets, the drivers and their vehicles. This reduced the standardization of their process in the sense that customers cannot know what type of car they will ride in. Second, they abandoned the rigid pricing of taxis and replaced it with a revenue model based on substantial and frequent adjustments to pricing. The combination of these two innovations allows Uber to provide far more vehicles on the road, especially when demand is high, which generates considerable value both for Uber and consumers.

The business-model examples given suggest that there have been many successful models and they have taken diverse forms. Nevertheless, in all cases, these firms made decisions regarding the elements of their revenue and supply models. New approaches have emerged when firms discovered new combinations of these decisions that in net yield better results than the industry status quo. Often (but not always, e.g., Zara), these new business models are enabled primarily by changes in technology. This pattern is likely to continue. For example, autonomous vehicles will alter the economics of transportation, which could not only influence how we move people and goods around, but could also influence where we choose to live. Technology could lead to radical changes in education. One possibility is that education becomes more standardized (i.e., the organization providing the education exerts more control over the process). Gone may be the days when professors across universities teach essentially customized classes. A more standardized approach to education may be delivered with less-expensive workers (standardized processes requires less skill and judgment), which would expand the potential scope of education and lower its cost. The extreme opposite on the standardization spectrum could also happen—technology could be used to customize education to each student. And these paths are also conceivable in medicine—more standardized and lower cost, or customized and higher quality. Beyond computer and communication technology, we are likely to experience substantial changes in our energy portfolio, which could influence where production occurs. In each of these cases, a firm may discover that moving its resource locations, changing its pricing mechanism, changing which resources it owns, or some mixture of these decisions leads to more consumer value and/or lower costs. When that happens, the next Zara, Amazon, or Uber will be well on its way to dominating its market.

Acknowledgments

The author thanks Robert Swinney, Jan van Mieghem, Marty Lariviere, the participants of the Kellogg Workshop, and the editorial team for helpful feedback on earlier versions of this manuscript.

Appendix

A.1. Probabilistic Selling

Probabilistic selling allows the firm to price-discriminate between customers that vary in the strength of their preferences across the product line. A simple example illustrates why this can improve revenue. Table A.1 displays the preferences of four types of customers for two products, A and B. The first and fourth types have strong preferences for only product A or B. The middle types are relatively indifferent between the two products. Normalize total demand in the market to equal 1, split evenly across the four types.

If the firm can only sell the two products, then the optimal pricing strategy is to charge 0.7 for each one, earning a total revenue of 0.7. (Charging 1 for each product only sells to the extreme types, earning 0.5 in revenue.)

To do better, the firm can create product C, which is a virtual/probabilistic product: If a customer selects product C, then they are given product A with a 0.5 probability and product B with a 0.5 probability. The optimal prices are 1 for products A and B and 0.6 for product C. The extreme types (1 and 4) prefer to purchase their preferred product, and the middle types prefer the probabilistic product C. Total revenue is 0.8, substantially higher than revenue without the virtual product.

For probabilistic selling to work, the expected value of the relatively indifferent types must be sufficiently high. Say the indifferent types' values are 0.5 and 0.3 for the two products. Now they are willing to pay at most 0.40 for the probabilistic product C. But if 0.4 is offered for product C,

Table A.1. Preferences for Four Customer Types over Two Products

Customer type	Value for product A	Value for product B	Fraction of market
1	1	0	0.25
2	0.7	0.5	0.25
3	0.5	0.7	0.25
4	0	1	0.25

the other types (1 and 4) also purchase product C, and total revenue drops to 0.4.

Selling virtual products is costly to the system: Total surplus is 0.85 when the firm sells only products A and B, but is 0.80 with the virtual product included. It works in the firm's favor because the firm is able to capture a greater share of that surplus: The firm's share is 0.7/0.85 = 82% with just the two products, and 0.8/0.8 = 100% with the virtual product. This undesirable feature of probabilistic selling could work against it the domain of public perception.

A.2. Subscription vs. Per-Use

This section describes a simplified version of the model in Cachon and Feldman (2011). There is one firm and two potential customers. Each customer values the service with probability ϕ. If they value the service, their value is uniformly distributed on the interval $[0,1]$; otherwise, they have no value for the service. To serve a customer requires one unit of capacity, and the firm can choose capacity $q = 1$ or $q = 2$. If $q = 2$, then both customers can be served immediately. If $q = 1$, then one customer can be served immediately and a second customer is served with delay cost w. The firm can charge a subscription to each customer or charge per use. With a subscription, each customer pays before observing their value and before observing the queue length. With per-use, customers decide whether to pay after observing their value for the good and the queue length. The cost for each unit of capacity is k. The marginal cost to serve a customer is $c < 1/2$.

There are four cases to consider: two types of contracts and two capacity levels. Table A.2 displays the profits and prices in each case.

To compare the strategies, first consider when congestion costs become small—that is, as $w \to 0$. In this case, subscription with $q = 1$ dominates because there is no need for a second unit of capacity if customers don't care about congestion ($\pi_s(q = 2) < \pi_s(q = 1)$), and subscription generates more revenue than per-use ($\pi_p(q = 1) < \pi_s(q = 1)$). However, when capacity is limited, $q = 1$, subscription profit is more sensitive to congestion costs than per-use:

$$\frac{\partial \pi_s(q = 1)}{\partial w} < \frac{\partial \pi_p(q = 1)}{\partial w} < 0.$$

This disadvantage of subscriptions is intuitive—subscriptions do not regulate usage and therefore lead to excessive consumption, which leads to additional congestion. But even though subscription is more sensitive to congestion, it starts with a higher profit, and therefore may be superior to per-use, even with high congestion costs (but not so high as to render either contract infeasible). To illustrate this point, the worst case for subscriptions relative to per-use pricing occurs when $w = k/\phi^2$: For smaller w subscription chooses $q = 1$ and its profit falls faster than per use profit, whereas for larger w subscription chooses $q = 2$, its profit is independent of w but per-use profit continues to decrease with w. Even in the worst case for subscription (i.e., $w = k/\phi^2$), subscription yields a higher profit than per-use for all $c \le \hat{c}$, where

$$\hat{c} = \sqrt{2} - 1 - \frac{k}{2\phi}.$$

So subscription is surely better than per-use when the marginal cost of service is low. But this does not tell the entire story. If

$$\frac{k}{\phi} < 3 - 2\sqrt{2} \approx 0.17.$$

then subscription, there exists a threshold c', $c' < \hat{c}$, such that for all $c \in [c', \hat{c}]$, subscription earns a positive profit but per-use does not—that is, $\pi_p < 0 < \pi_s$. In other words, it is possible that subscriptions yield higher profits than per-use for all feasible values of congestion cost. Counterintuitively, when it would seem that controlling congestion is most important (when consumers strongly dislike congestion), it can be optimal to use the pricing scheme that does not control congestion (subscriptions) because it is more important to generate revenue to enable more investment in capacity than it is to control consumer behavior with limited capacity.

A.3. Dynamic Pricing

This model extends the one studied in Su and Zhang (2008). One firm sells a single product over a selling season with two periods. The product is sold for the full price, p, in period 1 and potentially sold for the discount price, v, in period 2. The firm purchases q units before the selling season at a cost of c per unit. There is a random amount of demand (i.e., customers) in period 1, which creates the risk of leftover inventory. Let $F(\cdot)$ and $f(\cdot)$ be the distribution and density functions of demand, which satisfy the reasonable regularity property of an increasing generalized failure rate (IGFR) (Lariviere 2006). Let d be realized demand and μ be expected demand.

There are two types of consumers. A fixed fraction of demand, α, is "myopic": Myopic consumers receive utility $u = 1$ if they purchase in period 1 and they only consider purchasing in period 1 (e.g., the earn zero utility if they purchase in period 2). The other $1 - \alpha$ fraction of demand is "strategic": Strategic consumers choose which period to make a purchase and receive $u = 1$ utility for a purchase in either period. Let β be the fraction of demand that attempts to purchase in period 1. If the number of period-1 customers exceeds the firm's quantity, $q \le \beta d$, then units are randomly allocated. Otherwise, the unsold product from period 1 is

Table A.2. Profits and Prices with Subscriptions, "s," and Per-Use Pricing, "p," with Two Capacity Levels

Capacity	π_s	π_p	p_s	p_p
$q = 1$	$\phi(1 - 2c - \phi w) - k$	$\frac{1}{2}\phi\left(1 - c - \frac{1}{2}\phi w\right)^2 - k$	$\phi(1/2 - \phi w/2)$	$(2 + 2c - \phi w)/4$
$q = 2$	$\phi(1 - 2c) - 2k$	$\frac{1}{2}\phi(1 - c)^2 - 2k$	$\phi/2$	$(1 + c)/2$

sold in period 2 for v per unit, $v < c$. Strategic consumers who chose to wait until period 2 to purchase are the first to be served in period 2. If any inventory remains in period 2 after those consumers are served, the remaining inventory is sold for v to a large pool of discount shoppers. Thus, all remaining inventory at the start of period 2 is sold for v.

Strategic consumers choose in which period to purchase to maximize their net utility given their rational expectations. In particular, they believe there is a ϕ_t probability that they receive a unit if they choose to purchase in period t. In equilibrium, their beliefs are correct. For consumers, purchasing in period 1 has a higher price, $v < p$, but also a higher likelihood of product availability, $\phi_2 < \phi_1$.

Let $S_\beta(q) = E[\min\{\beta d, q\}]$ be the expected sales function when α is the fraction of demand that purchases in period 1:

$$S_\beta(q) = \int_0^{q/\beta} \beta x f(x) dx + (1 - F(q/\beta))q = \beta S(q/\beta),$$

where $S(q) = S_1(q)$ for notational convenience.

The firm's optimal period 1 price with the value strategy is the maximum price that induces consumers to purchase in period 1 given their beliefs:

$$p = 1 - (1 - v)\phi_2/\phi_1. \tag{A.1}$$

Assuming in equilibrium strategics purchase in period 1, then strategics believe they can purchase a unit in period 1 with probability $\phi_1 = S_1(q)/\mu$. (Note, from the consumer's perspective, $df(d)/\mu$ is the conditional probability of being in a market with d units of demand when all consumers participate.)[13] In period 2, the correct belief for strategics is

$$\phi_2 = \int_0^q \frac{xf(x)}{\mu} dx.$$

The firm's profit with price p, β fraction of consumers purchasing in period 1, and surely discounting in period 2 is

$$\pi(p, q, \beta) = (p - v)S_\beta(q) - (c - v)q. \tag{A.2}$$

With the hi–low price strategy $p = 1$ and $\beta = \alpha$. With value pricing, $p < 1$ and $\beta = 1$. With either strategy, the optimal quantity satisfies

$$F(q^*/\beta) = \tfrac{p-c}{p-v}. \tag{A.3}$$

(Note, q must be consistent with expectations, but changing q cannot change consumer expectations.) Given the optimal price, (A.1), the quantity equation for the value-price strategy can be written as

$$\frac{q\bar{F}(q^*)^2}{S(q^*)} = \frac{c - v}{1 - v}. \tag{A.4}$$

The left-hand side of (A.4) is decreasing for an IGFR distribution, which implies that there exists a unique optimal q^*.

With the never-discount strategy, the firm selects $p = 1$, strategics purchase in period 1, and to evaluate the firm's profit, it can be assumed that the firm's period-2 price is $v = 0$.

With the refund pricing strategy, the strategic consumers are indifferent between purchasing in period 1 or 2. To explain, their expected period-1 surplus is $(1 - v)\phi_2$ because they earn the $1 - v$ surplus if there is a period-2 discount. Their period-2 surplus is the same—they earn $1 - v$ in period 2 only if there is a discount, which occurs with probability ϕ_2. Thus, with this refund strategy, the firm's profit is

$$\pi(q) = (1 - v)q\bar{F}(q) - (c - v)q.$$

The firm's optimal quantity, q^*, uniquely satisfies (because demand is IGFR) the following equation:

$$\bar{F}(q^*)\left(1 - \frac{q^* f(q^*)}{\bar{F}(q^*)}\right) = \frac{c - v}{1 - v}.$$

With the quantity commitment contract, the price, given q, is given by (A.1). Substitution into the profit function yields the same profit function as with the refund strategy.

To compare the various strategies for the firm, construct a set of 108 scenarios from all combinations of the following parameters in which demand is modeled with a gamma distribution:

$$\alpha \in \{0.25, 0.5, 0.75\} \quad \mu = 100 \quad c =\in \{0.25, 0.5, 0.75\}$$
$$\sigma/\mu =\in \{0.25, 0.5, 1, 1.5\} \qquad v/c =\in \{0.3, 0.6, 0.9\}.$$

The best of the firm's four strategies always earns less than the firm's upper bound on profit, which occurs when all consumers are myopic, ($\alpha = 1$), and earns on average only 72%, 74%, and 82% of that bound when the share of myopic consumers is $\alpha = 0.25$, $\alpha = 0.50$, and $\alpha = 0.75$, respectively. Trying to woo strategic consumers with a low initial price (the value strategy) is generally not effective. It can be better than the hi–low strategy only when there are many strategics and few myopic consumers ($\alpha = 0.25$). However, the value strategy is never optimal among all of the strategies. The never-discount strategy is optimal in 80 of the 108 scenarios (74%). Because salvage revenues are of second-order importance relative to full-price sales, committing to never discount only performs poorly in extreme scenarios in which it is very costly to forgo salvage revenues (high demand uncertainty, high product cost, and high salvage prices). Finally, although committing to a full refund can be optimal, it rarely is: It is optimal in only 6 of 108 scenarios. (That said, there may be better designs for refund policies.)

A.4. Timing

See the main text for the description of the model. First, consider inventory. During the low phase, the firm can build inventory that will allow it to serve demand during the high period. Average inventory is

$$I = \frac{1}{4}\left(\frac{\lambda_h - \lambda_l}{\mu - \lambda_l}\right)(\lambda_h - \mu)\tau.$$

Let $\mu = \phi\delta + \lambda$, where $\phi \in [0, 1]$ and $\delta = (\lambda_h - \lambda_l)/2$. The firm chooses ϕ. The inventory equation can now be written as

$$I = \frac{1}{2}\tau\delta\left(\frac{1 - \phi}{1 + \phi}\right) = Q,$$

which is also the average queue length, Q, if the firm chooses to make customers wait rather than to build inventory.

Optimization. Let π be the firm's total cost. With the make-to-stock approach

$$\pi = \mu + hI = (\phi\delta + \lambda) + h\frac{1}{2}\tau\delta\left(\frac{1-\phi}{1+\phi}\right)$$
$$= \delta\left(\phi + \frac{1}{2}\tau h\left(\frac{1-\phi}{1+\phi}\right)\right) + \lambda,$$

π is convex in ϕ. The resulting cost is

$$\pi^* = \begin{cases} \lambda + \frac{1}{2}\delta\tau h & \tau h < 1 \\ \lambda + \delta\left(2\sqrt{\tau h} - 1 - \frac{1}{2}\tau h\right) & 1 \le \tau h \le 4 \\ \lambda + \delta & 4 < \tau h. \end{cases} \quad (A.5)$$

The firm's cost per unit of demand (π^*/λ) increases with both forms of seasonal variability, the amplitude(δ) and the duration (τ). If doing work in advance is relatively inexpensive ($h < 1/\tau$), then the firm chooses the minimum capacity and uses the entire low season to build inventory to satisfy high-season demand. On the other hand, if doing work in advance is sufficiently costly ($4/\tau < h$), then the firm chooses the maximum capacity and does not do any work in advance. If the firm just uses a queue, then the results are identical to (A.5), with the exception that h is replaced with the cost of waiting, w.

Now, consider the case in which the firm is able to do some of the work in advance and some that is done only after demand arrives. Say γ fraction of excess demand during the high phase is completed in advance. The amount of excess high-phase work done in advance is $\gamma(\lambda_h - \mu)\tau$. The average inventory of work is

$$I = \frac{1}{2}\gamma(\lambda_h - \mu)\tau\left(\frac{\tau_l + \tau}{2\tau}\right) = \frac{1}{4}\gamma\delta\tau(1 + \gamma + (1-\gamma)\phi)\left(\frac{1-\phi}{1+\phi}\right).$$

The remaining $(1-\gamma)(\lambda_h - \mu)\tau$ of work during the high phase creates a queue that is then drawn down during the low phase. The average queue length is then

$$Q = \frac{1}{4}\delta\tau(1-\gamma)(2 - \gamma(1-\phi))\left(\frac{1-\phi}{1+\phi}\right).$$

The firm's cost is

$$\pi = \mu + \tfrac{1}{4}h\gamma\tau\delta(1 + \gamma + (1-\gamma)\phi)\left(\tfrac{1-\phi}{1+\phi}\right)$$
$$+ \tfrac{1}{4}w(1-\gamma)\tau\delta(2 - \gamma(1-\phi))\left(\tfrac{1-\phi}{1+\phi}\right)$$
$$= \delta\Big(\phi + \tfrac{1}{4}\tau\left(\tfrac{1-\phi}{1+\phi}\right)(h\gamma(1 + \phi + \gamma(1-\phi))$$
$$+ w(1-\gamma)(2 - \gamma(1-\phi)))\Big) + \lambda.$$

Holding ϕ fixed, the optimal γ (unconstrained) is

$$\gamma^* = \frac{(3-\phi)(w/h) - (1+\phi)}{2(w/h+1)(1-\phi)}.$$

Note that

$$\frac{\partial\gamma^*}{\partial\phi} = \frac{w/h - 1}{(w/h+1)(1-\phi)^2},$$

which is positive for $0 < w/h$. It is possible to show that the upper bound on $\gamma < 1$ is achieved with $\phi = 1$ and $w/h = 3$. Thus, if the cost of waiting is more than three times the cost of holding inventory, the optimal solution for all ϕ is $\gamma = 1$, which means to build inventory. Analogously, if $w/h < 1/3$, then $\gamma = 0$ is optimal for all values of ϕ : If waiting costs are low, then it is best to build a queue.

Endnotes

[1] There are alternative frameworks for business models, such as the Business Model Canvas (https://en.wikipedia.org/wiki/Business_Model_Canvas). The Business Model Canvas creates nine dimensions that managers can use to describe their business. The emphasis is more on the details of each of these dimensions rather than on high-level strategic decisions and their interactions. Nevertheless, all frameworks for business models include some element of how firms earn revenue and how they produce value.

[2] Haggling is yet another option, with similar strengths and limitations as auctions (Desai and Purohit 2004).

[3] The break-even marginal cost solves $3.45 - 10 \times 0.885 \times c = 10(1-c)^2/4$.

[4] Learning about demand can be another motivation for dynamic pricing: Besbes and Zeevi (2015) and Yu et al. (2015).

[5] Chevalier and Goolsbee (2009) and Li et al. (2014) provide empirical evidence of strategic consumer behavior.

[6] This model does not consider the possibility that the firm can take actions to manipulate/distort consumer perceptions of availability, and therefore the likelihood of future discounts (Yin et al. 2009, Özer and Zheng 2016). Nor do these consumers care whether others are able to purchase (Tereyagoglu and Veeraraghavan 2012).

[7] Never discount and value pricing are actually extreme versions of a single strategy in which the firm randomly marks down to \$30 in period 2 with probability θ if there is inventory left over: Never discount is a random markdown with $\theta = 0$, and value pricing is a random markdown with $\theta = 1$. In this case, the firm's optimal strategy is the extreme to never discount. See Moon et al. (2018) for a broader analysis of random markdowns.

[8] There is one additional strategy that initially seems different but in fact is equivalent to refunds—the firm announces before the season starts the number of units purchased, q. Assuming this is feasible, the goal is to signal to consumers that product availability is limited, so they should purchase in period 1, even though the price is high.

[9] Research on ride sharing is rapidly expanding: for example, Bimpikis et al. (2016), Besbes et al. (2018), and Hu and Zhou (2017).

[10] There is a 0.5 probability of having a value greater than $p = 1/2$, and $3/4$ is her expected value conditional on having a value greater than p. So expected value is $(3/4 - p)1/2 = 1/8$.

[11] One might argue that space costs shouldn't be included in h because space is not adjustable in the short term. True, but when comparing two operating systems, long-run costs should be compared, and space is surely adjustable in the long term.

[12] Although Zara does not rely heavily on discounting, they nevertheless must discount some merchandise. See Gallien and Caro (2012) for a discussion of how Zara optimizes markdowns.

[13] To explain, suppose demand can either be $d = 1$ or $d = 99$, each equally likely. A consumer should not assume that there is a 0.5

probability of being the only consumer in the market ($d = 1$). Instead, the conditional probability of being in a market with d consumers is $df(d)/\mu$: The probability a consumer is the only one in the market is $1f(1)/\mu = 0.01$ and the probability the consumer is in a market with many others is $99f(99)/\mu = 0.99$.

References

Acimovic J, Graves S (2015) Making better fulfillment decisions on the fly in an online retail environment. *Manufacturing Service Oper. Management* 17(1):34–51.

Alexandrov A, Lariviere MA (2012) Are reservations recommended? *Manufacturing Service Oper. Management* 14(2):218–230.

Allon G, Gurvich I (2009) Pricing and dimensioning competing large-scale service providers. *Manufacturing Service Oper. Management* 12(3):449–469.

Amaldoss W, Jain S (2008) Joint bidding in the name-your-own-price channel: A strategic analysis. *Management Sci.* 54(10):1685–1699.

Ang E, Iancu DA, Swinney R (2017) Disruption risk and optimal sourcing in multitier supply networks. *Management Sci.* 63(8): 2397–2419.

Bakos Y, Brynjolfsson E (1999) Bundling information goods: Pricing, profits, and efficiency. *Management Sci.* 45(12):1613–1630.

Belavina E, Girotra K (2012) The relational advantages of intermediation. *Management Sci.* 58(9):1614–1631.

Bell D, Gallino S, Moreno A (2017) Offline showrooms in omnichannel retail: Demand and operational benefits. *Management Sci.* 64(4):1629–1651.

Besanko D, Winston WL (1990) Optimal price skimming by a monopolist facing rational consumers. *Management Sci.* 36(5): 555–567.

Besbes O, Zeevi A (2015) On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Sci.* 61(4):723–739.

Besbes O, Castro F, Lobel I (2018) Surge pricing and its spatial supply response. Columbia Business School Research Paper 18-25, Columbia Business School, New York.

Bimpikis K, Candogan O, Saban D (2016) Spatial pricing in ride-sharing networks. Working paper, Stanford University, Stanford, CA.

Blattberg RC, Eppen GD, Lieberman J (1981) A theoretical and empirical evaluation of price deals for consumer nondurables. *J. Marketing* 45(1):116–129.

Brynjolfsson E, Hu Y, Smith MD (2003) Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Sci.* 49(11):1580–1596.

Brynjolfsson E, Malone TW, Gurbaxani V, Kambil A (1994) Does information technology lead to smaller firms? *Management Sci.* 40(12):1628–1644.

Bulow JI (1982) Durable-goods monopolist. *J. Political Econom.* 90(2): 314–332.

Cachon GP (2003) Supply chain coordination with contracts. Graves SC, de Kok AG, eds. *Supply Chain Management: Design, Coordination and Operation*, Handbooks in Operations Research and Management Science, vol. 11 (Elsevier, Amsterdam), 227–339.

Cachon GP (2014) Retail store density and the cost of greenhouse gas emissions. *Management Sci.* 60(8):1907–1925.

Cachon GP, Feldman P (2011) Pricing services subject to congestion: Charge per-use fees or sell subscriptions? *Manufacturing Service Oper. Management* 13(2):244–260.

Cachon GP, Feldman P (2018) Pricing capacity over time and recourse strategies: Facilitate reselling, offer refunds/options, or overbook? Working paper, Wharton School, University of Pennsylvania, Philadelphia.

Cachon GP, Lariviere MA (2005) Supply chain coordination with revenue-sharing contracts: Strengths and limitations. *Management Sci.* 51(1):30–44.

Cachon GP, Swinney R (2009) Purchasing, pricing, and quick response in the presence of strategic consumers. *Management Sci.* 55(3):497–511.

Cachon GP, Swinney R (2016) The value of fast fashion: Quick response, enhanced design, and strategic consumer behavior. *Management Sci.* 62(2):410–435.

Cachon GP, Zhang F (2006) Procuring fast delivery: Sole sourcing with information asymmetry. *Management Sci.* 52(6):881–896.

Cachon GP, Daniels KM, Lobel R (2017) The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing Service Oper. Management*. 19(3):368–384.

Casadesus-Masanell R, Llanes G (2011) Mixed source. *Management Sci.* 57(7):1212–1230.

Chen N, Gallego G (2018) Welfare analysis of dynamic pricing. *Management Sci*. 65(1):139–151.

Cheng HK, Liu Y (2012) Optimal software free trial strategy: The impact of network externalities and consumer uncertainty. *Inform. Systems Res.* 23(2):488–504.

Chevalier J, Goolsbee A (2009) Are durable goods consumers forward-looking? Evidence from college textbooks. *Quart. J. Econom.* 124(4): 1853–1884.

Chu LY, Zhang H (2011) Optimal preorder strategy with endogenous information control. *Management Sci.* 57(6):1055–1077.

Clifford S, Rampell C (2013) Sometimes, we want prices to fool us. *New York Times* (April 13), https://www.nytimes.com/2013/04/14/business/for-penney-a-tough-lesson-in-shopper-psychology.html.

Cui R, Li J, Zhang DJ (2019) Discrimination with incomplete-information in the sharing economy: Evidence from field experiments on Airbnb. *Management Sci*. Forthcoming.

Cui Y, Duenyas I, Şahin Ö (2014) Should event organizers prevent resale of tickets? *Management Sci.* 60(9):2160–2179.

Desai P, Purohit D (1998) Leasing and selling: Optimal marketing strategies for a durable goods firm. *Management Sci.* 44(11-part-2):S19–S34.

Desai P, Purohit D (2004) "Let me talk to my manager": Haggling in a competitive environment. *Marketing Sci.* 23(2):219–233.

Etzion H, Pinker E, Seidmann A (2006) Analyzing the simultaneous use of auctions and posted prices for online selling. *Manufacturing Service Oper. Management*. 8(1):68–91.

Fay S (2004) Partial-repeat bidding in the name-your-own-price channel. *Marketing Sci.* 23(3):407–418.

Fay S, Xie J (2008) Probabilistic goods: A creative way of selling products and services. *Marketing Sci.* 27(4):674–690.

Gale IL, Holmes TJ (1993) Advance-purchase discounts and monopoly allocation of capacity. *Amer. Econom. Rev.* 83(1):135–146.

Gallego G, Sahin O (2010) Revenue management with partially refundable fares. *Oper. Res.* 58(4):817–833.

Gallien J, Caro F 2012. Clearance pricing optimization for a fast-fashion retailer. *Oper. Res.* 60(6):1404–1422.

Gallino S, Moreno A (2014) Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management Sci.* 60(6):1434–1451.

Gallino S, Moreno A, Stamatopoulis I (2017) Channel integration, sales dispersion, and inventory management. *Management Sci.* 63(9):2813–2831.

Gneezy A, Imas A, Brown A, Nelson LD, Norton M (2012) Paying to be nice: Consistency and costly prosocial behavior. *Management Sci.* 58(6):179–187.

Grossman GM, Helpman E (2002) Integration versus outsourcing in industry equilibrium. *Quart. J. Econom.* 117(1):85–120.

Grossman S, Hart O (1986) The costs and benefits of ownership: A theory of vertical and lateral integration. *J. Political Econom.* 94(4):691–719.

Guo L (2009) Service cancellation and competitive refund policy. *Marketing Sci.* 28(5):901–917.

Hann I-H, Terwiesch C (2003) Measuring the frictional costs of online transactions: The case of a name-your-own-price channel. *Management Sci.* 49(11):1563–1579.

Hansen S (2012) How Zara grew into the world's largest fashion retailer. *New York Times* (November 9), https://www.nytimes.com/2012/11/11/magazine/how-zara-grew-into-the-worlds-largest-fashion-retailer.html.

Ho T-H, Tang CS, Bell DR (1998) Rational shopping behavior and the option value of variable pricing. *Management Sci.* 44(12-part-2): S145–S160.

Hoch S, Drèze X, Purk ME (1994) EDLP, hi-lo, and margin arithmetic. *J. Marketing* 58(4):16–27.

Hosanagar K, Abhishek V (2013) Optimal bidding in multi-item multi-slot sponsored search auctions. *Oper. Res.* 61(4):855–873.

Hu M, Zhou Y (2017) Price, wage and fixed commission in on-demand matching. Working paper, University of Toronto, Toronto.

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

Jerath K, Netessine S, Veeraraghavan SK (2010) Revenue management with strategic customers: Last-minute selling and opaque selling. *Management Sci.* 56(3):430–448.

Kabra A, Belavina E, Girotra K (2015) Bike-share systems: Accessibility and availability. Working paper, University of Chicago, Chicago.

Kumar V (2014) Making "freemium" work. *Harvard Business Rev.* (May), https://hbr.org/2014/05/making-freemium-work

Lai G, Debo LG, Sycara K (2010) Buy now and match later: Impact of posterior price matching on profit with strategic consumers. *Manufacturing Service Oper. Management* 12(1):33–55.

Lariviere M (2006) A note on probability distributions with increasing generalized failure rates. *Oper. Res.* 54(3):602–604.

Lazear E (1986) Retail pricing and clearance sales. *Amer. Econom. Rev.* 76(1):14–32.

Lee HL, Tang CS (1997) Modelling the costs and benefits of delayed product differentiation. *Management Sci.* 43(1):40–53.

Lee HL, Padmanabhan V, Whang S (1997) Information distortion in a supply chain: The bullwhip effect. *Management Sci.* 43(4): 546–558.

Li J, Granados N, Netessine S (2014) Are consumers strategic? structural estimation from the air-travel industry. *Management Sci.* 60(9):2114–2137.

Li J, Moreno A, Zhang DJ (2016) Pros vs. joes: Agent pricing behavior in the sharing economy. Working paper, University of Michigan, Ann Arbor.

Malone TW (1987) Modeling coordination in organizations and markets. *Management Sci.* 33(10):1317–1332.

McWilliams G (2004) Analyzing customers, Best Buy decides not all are welcome. *Wall Street Journal* (November 8), https://www.wsj.com/articles/SB109986994931767086

Mendelson H, Parlaktūrk AK (2008) Product-line competition: Customization vs. proliferation. *Management Sci.* 54(2):2039–2053.

Moon K, Bimpikis K, Mendelson H (2018) Randomized markdowns and online monitoring. *Management Sci.* 64(3):1271–1290.

Mukhopadhyay T, Kekre S (2002) Strategic and operational benefits of electronic integration in B2B procurement processes. *Management Sci.* 48(10):1301–1313.

Nicolae M, Arikan M, Deshpande V, Ferguson M (2017) Do bags fly free? An empirical analysis of the operational implications of airline baggage fees. *Management Sci.* 63(10):3187–3206.

Özer Ö, Zheng Y (2016) Markdown or everyday low price? The role of behavioral motives. *Management Sci.* 62(2):326–346.

Parker GG, Van Alstyne MW (2005) Two-sided network effects: A theory of information product design. *Management Sci.* 51(10): 1494–1504.

Rob R, Waldfogel J (2006) Piracy on the high C's: Music downloading, sales displacement, and social welfare in a sample of college students. *J. Law Econom.* 49(1):29–62.

Rochet J-C, Tirole J (2006) Two-sided markets: A progress report. *RAND J. Econom.* 37(3):645–667.

Schmidt KM, Spann M, Zeithammer R (2015) Pay what you want as a marketing strategy in monopolistic and competitive markets. *Management Sci.* 61(6):1217–1236.

Simonsohn U, Ariely D (2008) When rational sellers face nonrational buyers: Evidence of herding on eBay. *Management Sci.* 54(9): 1624–1637.

Spann M, Tellis GJ (2006) Does the Internet promote better consumer decisions? The case of name-your-own-price auctions. *J. Marketing* 70(1):65–78.

Stokey NL (1981) Rational expectations and durable goods pricing. *Bell J. Econom.* 12(1):112–128.

Su X (2007) Intertemporal pricing with strategic consumer behavior. *Management Sci.* 53:726–741.

Su X (2010) Optimal pricing with speculators and strategic consumers. *Management Sci.* 56(1):25–40.

Su X, Zhang F (2008) Strategic customer behavior, commitment, and supply chain performance. *Management Sci.* 54(10):1759–1773.

Tereyagoglu N, Veeraraghavan S (2012) Selling to conspicuous consumers: pricing, production, and sourcing decisions. *Management Sci.* 58(12):2168–2189.

Terwiesch C, Hann I-H, Savin S (2005) Online haggling at a name-your-own-price retailer: Theory and application. *Management Sci.* 51(3):339–351.

Wang T, Chatterjee R, Gal-Or E (2009) The name-your-own-price channel in the travel industry: An analytical exploration. *Management Sci.* 55(6):968–979.

Van Mieghem JA (1999) Coordinating investment, production and subcontracting. *Management Sci.* 45(7):954–971.

Wang R (1993) Auctions versus posted-price selling. *Amer. Econom. Rev.* 83(4):838–851.

Xie J, Gerstner E (2007) Service escape: Profiting from customer cancellations. *Marketing Sci.* 26(1):18–30.

Xie J, Shugan S (2001) Electronic tickets, smart cards, and online prepayments: When and how to advance sell. *Marketing Sci.* 20(3):219–243.

Yin R, Aviv Y, Pazgal A, Tang CS (2009) Optimal markdown pricing: Implications of inventory display formats in the presence of strategic customers. *Management Sci.* 55(8):1391–1408.

Yu M, Debo L, Kapuscinski R (2015) Strategic waiting for consumer-generated quality information: Dynamic pricing of new experience goods. *Management Sci.* 61:723–739.

Zhang L (2018) Intellectual property strategy and the long tail: Evidence from the recorded music industry. *Management Sci.* 64(1): 24–42.