The impact of committing to customer orders in online retail

Citation for published version (APA):

Document license:
CC BY

DOI:
10.1287/msom.2022.1124

Document status and date:
Published: 01/01/2023

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Manufacturing & Service Operations Management

The Impact of Committing to Customer Orders in Online Retail

Gonçalo Figueira, Willem van Jaarsveld, Pedro Amorim, Jan C. Fransoo

To cite this article:

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article’s accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2022 The Author(s)

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org
The Impact of Committing to Customer Orders in Online Retail

Gonçalo Figueira, Willem van Jaarsveld, Pedro Amorim, Jan C. Fransoo

Abstract. Problem definition: Online retailers are on a consistent drive to increase on-time delivery and reduce customer lead time. However, in reality, an increasing share of consumers places orders early. Academic/practical relevance: Such advance demand information can be deployed strategically to reduce costs and improve the customer service experience. This requires inventory and allocation policies that make optimal use of this information and that induce consumers to place their orders early. An increasing number of online retailers not only offer customers a choice of lead time but also, actively back-order missing items from a consumer basket. Methodology: We develop new allocation policies that commit to a customer order upon arrival of the order rather than at the moment the order is due. We provide analytical results for the performance of these allocation policies and evaluate their behavior with real data from a large food retailer. Results: Our policy leads to a higher fill rate at the expense of a slight increase in average delay. The analysis based on real-life data suggests a sizeable impact that should impact current best practices in online retail. Managerial implications: With the changing landscape in online retail, customers increasingly place baskets of orders that they would like to receive at a planned and confirmed moment in time. Especially in grocery, this has grown fast. This fundamentally changes the strategic management of inventory. We demonstrate that online retailers should commit early to customer orders to enhance the customer service experience and eventually, to also create opportunities for reducing the cost of operations.

1. Introduction

Online retail is one of the fastest-growing industries, despite the thin and often negative margins. At the end of 2019, the e-commerce share of total U.S. retail sales reached 11.2%, with a substantial spike up to 16.1% during the coronavirus disease 2019 pandemic (FRED 2020). The preference for online retail over in-store purchasing depends greatly on the product category, being highest for books and entertainment and lowest for groceries, with the latter currently having the strongest growth; electronics and apparel are in between (Hamory and Rampoldt 2015). In terms of margins, even large online retailers struggle. Amazon has averaged just 1.8% in operating margins over the past three years (Amazon.com 2020), whereas online grocery retailers have netted as low as 0.5% (Foley 2017). These thin margins result from high fulfillment costs, which are triggered by the more complex and fragmented operations in the online retail environment. Fulfillment in online retail includes picking, packing, and shipping to individual customers, the latter being particularly costly (e.g., logistics costs in 2019 represented 27.9% of net sales for Amazon.com (Amazon.com 2020)).

In an attempt to counter these costs, online retailers strategically offer alternative delivery options—such as different lead times—to allow for consolidation in picking and transport. In grocery retail, different time slots with different lengths and different horizons are offered for attended home delivery (Agatz et al. 2011). The latter not only increases options for consolidation but also, helps consumers plan in more detail when...
they plan to be home such that they are able to receive the groceries they ordered.

However, service levels in online grocery retail are low and stockouts are common (Jing and Lewis 2011). It is increasingly common for retailers to back-order missing items. For instance, Target, one of the largest online retailers in the United States, back-orders items that are missing in a basket, explaining to their customers that “‘Backordered’ means there’s a delay on your order because an item is temporarily out of stock. The item will ship as soon as it’s in stock” (Target 2020). Electronics retailer Best Buy offers a similar service. Such back orders are costly because they involve additional shipments to consumers. Offering back orders as an option also complicates the service-level experience by the customer; apart from the fill rate at the moment of ordering, the delay experienced in getting the back order delivered also becomes an important metric. Hence, online retailers now have a larger portfolio of managerial decisions to make in their service strategy; inducing customers to place orders early provides advance information and allows for order consolidation, whereas the option to have products back-ordered adds an additional dimension to the service-level experience. Although we focus our narrative on online retail, similar trends can be observed in business-to-business (B2B) commerce.

Online retailers can leverage this additional information and improve their operations. A key piece of information results from the fact that customers place orders for a future due date.

Having the ability to select a convenient delivery window has been of great importance to online consumers for a while now (Buhler 2016). In online grocery retail, attended home delivery—where consumers need to indicate a preferred time slot in advance—is the current norm in the industry. This results in a lead time that is larger than the shipping time; it provides “advance demand information” (ADI) that the retailer can use to better manage its inventory (Harigaran and Zipkin 1995) and hence, increase operating margins.

In current inventory models with ADI, the commitment to the consumer is made upon delivery (commit upon delivery (CUD)). This is at the moment when most information is available and the best decision can be made in terms of minimizing delay. However, the growing reality in online retail is that customers pick their preferred lead time and expect a higher service level when they place their order in advance. This implies that in addition to using ADI for inventory control, this information should also be used for inventory allocation. This allocation decision may need to be made when the customer places the order rather than when the order is due for delivery. This fundamentally changes the inventory logic and the associated strategic choices. Current ADI inventory models lack this option to allocate early and hence, do not take advantage of this new reality in online retail.

The service level of an online retailer can be measured in terms of the fill rate of the orders placed and the delay of the orders that are backlogged. Any policy needs to be evaluated against these performance measures. Strategically, however, another property of any policy is important, and this is whether a policy is advancement inducing. An advancement-inducing policy is a policy that rewards consumers who place orders early by committing early to such consumers and by providing them with better service than consumers ordering late. Such commitment needs to be firm and thus, prevent the item from being reallocated to a later order. Note that inducing of advancement may go at the expense of the overall delay.

Committing to an order upon delivery (CUD) hence may not be the best choice for the allocation decision in online commerce because it is not advancement inducing. In fact, consumers who would place orders early would not be able to get a firm commitment at the moment that they place the order. For such consumers, it would be rational to wait to place their order until they are close to the moment of delivery. Such delay will lead to considerably fewer opportunities for the retailer to realize efficiency gains in the operation.

Instead, we propose that it is better to commit upon arrival of the order, which is advancement inducing. In this paper, we define such an allocation policy and evaluate it both analytically and numerically. Our numerical evaluation is based on actual data of a European grocery retailer.

Our analysis demonstrates that when committing upon arrival of an order (commit upon arrival (CUA)), it is not a good idea to allocate units of the supply pipeline following first in, first out (FIFO). This is too naïve and does not make effective use of the time between the arrival of an order and its due date. Its service level is worse than CUD. Therefore, we propose a second rule, which allocates the “least-critical unit” (LCU) (i.e., the unit in the supply pipeline that will arrive latest but still in time to fulfill the order). The commit upon arrival with least-critical unit allocation (CUA-LCU) policy dominates commit upon arrival with first in, first out allocation (CUA-FIFO) in every indicator. When compared with CUD, it results in longer delays but higher fill rate. This is a relevant trade-off in the online setting. To study this trade-off analytically, we define and study sensible policies (i.e., policies that satisfy an intuitive constraint on back-order clearance that is reasonable in environments where customers are all equally important). The corresponding broad class of policies contains CUD, CUA-LCU, and CUA-FIFO as well as a wide range of other policies. A particularly interesting analytical finding of our work is then
that the sensible policy that maximizes the fill rate is precisely the CUA-LCU policy, whereas the classical CUD policy is the sensible policy that minimizes the average waiting time. Further, we specifically analyze the system under study using a composite performance measure that weighs the relative importance of customers who place their orders early versus those who place their orders late.

With the changing landscape in online retail, customers increasingly place baskets of orders that they would like to receive at a planned and confirmed moment in time. Especially in grocery, this has grown fast. This fundamentally changes the strategic management of inventory not only because the customers may have different expectations but also, because the willingness of customers to order early provides an opportunity for cost reduction and service-level improvement for thin-margin online retailers. We demonstrate in this paper that online retailers have the tools at hand to do so; they should commit early to customer orders to enhance the customer service experience and eventually, to also create opportunities for reducing the cost of operations. Our application in a grocery retail setting demonstrates that the effect size of our novel strategy is substantial and hence, would need to be considered by many retailers.

The next section frames our work in the literature, shedding light on its novelty and the similarities with related research. Section 3 describes the problem formulation and highlights the underlying assumptions. Section 4 provides analytical results that pave the ground for the comparison of the different allocation policies. In Section 5, the numerical study is presented in order quantify the differences of the policies and derive relevant insights. Finally, Section 6 summarizes the main results and discusses future research directions.

2. Literature Review

Our work is at the intersection of two main streams of inventory research: ADI and allocation policies. The ADI literature has usually assumed a given allocation policy, such as Harirhan and Zipkin (1995) for continuous review and Gallego and Özer (2001) for periodic review policies. These seminal studies have been extended in multiple ways, such as production-inventory systems (Özer and Wei 2004) and systems with imperfect ADI (Tan et al. 2007).

In our work we specifically consider the allocation problem in the context of ADI. Allocation may be required in several contexts, such as divergent/convergent flows, multiple demand classes, early delivery, or when attaining multiple service-level measures or strategic directives.

There is substantial work on models that consider inventory allocation in multiechelon distribution systems (Özer 2003). In these divergent one-warehouse multiple-retailer supply chains, inventory can be reserved upstream at the warehouse and allocated to different downstream retailers. The reverse configuration, with multiple warehouses satisfying each customer, also requires inventory allocation (Mahar and Wright 2009). Acimovic and Graves (2014) approach one such problem in an online retail setting, where fulfillment centers (FCs) and shipping methods are selected for each customer order.

When dealing with multiple demand classes, allocation policies are usually referred to as “rationing policies,” and they reserve inventory for customers who are more sensitive to stockouts (Gayon et al. 2009). Similarly, in the presence of flexible delivery, they may either reserve inventory for future demand or prioritize current orders, considering the trade-off between back-order and holding costs, respectively (Wang and Toktay 2008). Van Foreest et al. (2018) consider a different trade-off because their inventory model includes both fixed and time-dependent backorder costs. When clearing backlog, they reserve some maximum number of items in stock for future demand so that new back orders are prevented. This allows for finding a balance between back-order level and fill rate.

Because customers expect firm commitments from the seller in the online retail setting, the actual time of allocation is critical. Acimovic and Graves (2014) study a problem in online retail where they model the decision of how to fulfill orders that had already been committed prior to the allocation decision. Wang et al. (2014) consider actual commitment to orders. Their commitment is, however, loose because customers are given an expected rather than a firm delivery date upon arrival. The delivery date is not guaranteed, and they can be further backlogged. Kang et al. (2016) discuss this separation between commitment/acceptance and allocation and extend previous work by considering acceptance postponement and order abandonment. However, their allocation problem refers to capacity in manufacturing and service systems, not inventory. Their formulation is thus appropriate for applications such as manufacturing facilities, healthcare service providers, and law companies, where capacity has to be carefully managed and the company is frequently negotiating with customers, but not for online retail.

Marklund (2006) was the first to recognize the relevance of including timed allocation decisions in inventory models with advanced demand information. In a setting with one warehouse, multiple retailers, and perfect ADI, he provides exact and approximate cost evaluation techniques for a general reservation policy along
with two specific policies: complete reservation (comparable with our “commit upon arrival”) and last-minute allocation (comparable with our “commit upon delivery”). Our work differs from Marklund’s pioneering work in several dimensions. (1) Although Marklund (2006) focuses on allocation timing, we also consider different ways of selecting units in the pipeline other than FIFO (specifically, selecting the “least-critical unit,” which provides better performance in multiple metrics). (2) Marklund (2006) considers an overall service level, whereas we analyze the interplay between allocation policies and customers’ expectations in terms of both fill rate and waiting time but also, fairness, which may depend on the amount of ADI they provide.

Our problem also has slight similarities with that of Van Foreest et al. (2018) because we consider multiple service-level measures. Other than that, it is substantially different, as in their case, there is no ADI, the underlying decision is based on reservation levels, and the policy makes a nonsensible prioritization of orders.

In addition to commitment and sensibleness, we look at advance inducement. Some papers have discussed the estimation of the value of ADI in order to support decisions on how much sellers should give (for instance, with promotions) to obtain more ADI. This appears to be particularly important because promotions or other types of compensations could extract some (if not all) of the value ADI brings (Gayon et al. 2009).

In short, although there are papers that have somehow analyzed some of the issues we identify, none have attempted to study them together. Instead of assuming a given strategy, we evaluate different policies, their service-level performance, and their alignment with multiple strategic directives. Moreover, we look at two decision dimensions: time of allocation/commitment and pipeline unit selection. The former was previously studied but in different problem settings. The latter is completely new.

3. Problem Description
3.1. Model and Preliminaries
We consider a company that operates a number of single-item inventory systems, and we focus on one such single-item system. Customers arrive in continuous time to place single-unit orders for the product for some future date. Shipments to customers are only initiated at certain discrete and equidistant shipment times. We define time such that one time unit corresponds to the distance between two consecutive shipment times (e.g., if we ship every day, then the time unit corresponds to one day.) Thus, shipment times occur at \( t \in \mathbb{N} \), where \( \mathbb{N} = \{1, 2, \ldots \} \). To be in line with common inventory terminology, we will mostly refer to each shipment time \( t \in \mathbb{N} \) as a period. In addition to initiating shipments, in each period we also review inventory and place replenishment orders. Inventory ordered in period \( t \in \mathbb{N} \) is available for satisfying customer demands in period \( t + L \), where \( L \in \mathbb{N} \) is the supplier lead time.

We number the customer orders in the order in which they arrive to the system. The first customer order placed is order 1, and the second order is order 2. In general, the \( i \)th order is order \( i \). For each order \( i \), let \( t_i \in \mathbb{R} \) be the (random) time at which it is placed. Each order arrives with a customer order lead time (also demand lead time), which may be translated into a requested shipment time (RST). We denote by \( r_i \in \mathbb{N} \) the RST for order \( i \). If order \( i \) is shipped at time \( r_i \), then we say that the order/request is met. We require \( t_i \geq r_i \), which is reasonable because a request \( i \) for which \( r_i < t_i \) can never be met. Orders that are not met are back-ordered.

The results derived in this paper hold under fairly general conditions on the demand and RST process. For the sake of concreteness, we put down some specific assumptions. Throughout, we use \( \delta \) as a dummy variable for continuous time points to distinguish them from periods, which are denoted by \( t \). Suppose that customer orders are placed following a nonhomogeneous Poisson process with rate \( \lambda(\delta) \), and let \( \Lambda(\delta) = \int_{\delta'=0}^{\delta} \lambda(\delta')d\delta' \) be the expected number of customers between zero and \( \delta \). Then, \( N(\delta) - \text{Pois}(\Lambda(\delta)) \) represents the total number of customers who arrive between zero and \( \delta \) for \( \delta \in \mathbb{R} \). To capture, for example, seasonality, \( \lambda(\delta) \) could be any bounded, continuous, nonnegative function of \( \delta \). Furthermore, the requested shipment dates \( r_i \) for \( i \in \mathbb{N} \) could, for example, satisfy \( r_i = [t_i + \ell_i] \), where \( \ell_i \geq 0 \) for \( i \in \mathbb{N} \) are some independent and identically distributed (i.i.d.) random variables. \([x]\) denotes the smallest integer larger than \( x \).

To avoid technicalities, we assume a finite demand horizon in the sense that there exists some \( T \in \mathbb{N} \) such that \( \Lambda(\delta) = 0 \) for \( \delta > T \). Then, \( N(T) \) equals the total number of orders over the horizon; we have \( \text{P}(N(T) \in \mathbb{N}) = 1 \), and orders are numbered \( i = 1, 2, \ldots, N(T) \). Note that although no new demands come in after \( T \), some demands may have request dates \( r_i > T \).

We assume that inventory is controlled using a periodic review \((s, nQ)\) policy, with reviews for every \( t \in \mathbb{N} \). In particular, at each \( t \in \mathbb{N} \), the quantity to order \( q_t \) is determined as the smallest multiple of \( Q \in \mathbb{N} \) that can raise the inventory position (IP) above the reorder point \( s \in \mathbb{N} \) (Silver et al. 1998). Such policies are commonly used because they can buffer against demand variability and because they can control fixed
ordering costs. In our setting, the inventory position is defined as units on hand plus units in outstanding resupply orders minus unsatisfied customer orders (cf. Hariharan and Zipkin 1995). Our assumption of an \((s, nQ)\) policy is for the sake of concreteness. In fact, the results that we will derive on allocation policies are independent of the key assumptions regarding the replenishment policy (cf. Lemma A.1). The on-hand inventory in period \(t\) after receiving the replenishment \(q_{t-1}\) is denoted by \(I(t)\), and the on-hand inventory at the end of that period is denoted by \(J(t)\). Initial inventory is denoted by \(I_0\), and for ease of exposition, we assume that there are no customer or replenishment orders in the system at time \(0\); thus, \(I(0) = J(0) = I_0\), and \(q_1 = 0\) for \(t \leq 1\).

At each period \(t \in \mathbb{N}\), we go through a specific sequence of events. The first two steps of this sequence deal with placing and receiving replenishment orders.

1. The replenishment quantity \(q_{t-1}\) ordered in period \(t - 1\) arrives: \(I(t) = I(t - 1) + q_{t-1}\).
2. The quantity \(q_t\) to order for this period is determined based on the replenishment policy.

Before discussing the allocation step, we first introduce some notation. An allocation policy decides for each period which customer orders to satisfy. Clearly, in each period there is a specific set of orders that are available to be satisfied. To define this set formally, we let \(R(t)\) be the set of all orders with RST at or before \(t \in \mathbb{N}\), including orders that are already shipped. That is, \(i \in R(t)\) if and only if \(r_i \leq t\). We let \(\sigma_i \in \mathbb{N}\) be the time at which order \(i \in \mathbb{N}\) is satisfied, and we let \(S(t)\) be the set of orders satisfied in period \(t\) (i.e., \(i \in S(t)\)) if and only if \(\sigma_i = t\). In line with discussions in Section 1, customer orders can be satisfied at or after their RST; thus, we require \(\sigma_i \geq r_i\), which implies \(S(t) \subseteq R(t)\). The set \(A(t) := R(t) \setminus \bigcup_{i < t}S(i')\) contains precisely the orders that are due at or before \(t\) but that were not yet satisfied in the periods before \(t\) (i.e., the orders that are available to be satisfied at \(t\)). For two sets \(X\) and \(Y\), the set \(X \setminus Y\) contains the elements in \(X\) that are not in \(Y\). Note that \(A(t)\) contains all orders that were available to be met last period minus orders satisfied last period plus orders that are due in this period. Thus, \(A(t)\) is the union of \(A(t - 1) \setminus S(t - 1)\) and the orders \(i\) with \(r_i = t\). This leads to the last step of the sequence of events that occurs every period \(t \in \mathbb{N}\).

3. The allocation policy determines the (set of) orders \(S(t)\) to be satisfied in this period, such that \(S(t) \subseteq A(t)\) and \(|S(t)| \leq I(t)\). (For \(i \in S(t)\), we have \(\sigma_i = t\). The inventory remaining at the end of the period equals \(J(t) = I(t) - |S(t)|\).

Here, \(|X|\) denotes the number of elements in the set \(X\).
Same Position those demands at the on-hand and the on-order units will be allocated to Figure 2.

3.2.2. Commit upon Arrival with FIFO Allocation. We next discuss an easy to implement policy that commits to a specific delivery time upon arrival of a customer.

\[ \sigma^{\pi(CUD)}_{\tau_i} \] at time \( \tau_i \) because it may depend on demand after \( \tau_j \).

In the example provided in Figure 1, the starting inventory is \( I_0 = 1 \) unit and \( L = 4 \). Order 1 arrives at \( \tau_1 = 0.3 \) with RST \( r_1 = 5 \) (the time when the first replenishment unit \( u_2 \) arrives at the on-hand inventory). Under \( \pi(CUD) \), the delivery time \( \sigma^{\pi(CUD)}_{\tau_i} \) depends on the demand trajectory after \( \delta = 0.3 \). If no more than one customer order arrives between 0.3 and 4, then \( \sigma^{\pi(CUD)}_{\tau_1} = r_1 = 5 \). However, if two or more customer orders with \( r_i \leq 4 \) arrive in the interval \((1, 4)\), then both the on-hand and the on-order units will be allocated to those demands at \( t \leq 4 \), implying that \( \sigma^{\pi(CUD)}_{\tau_1} > r_1 = 5 \) because further replenishments arrive at the earliest in period 6. This is what happens with order 3, which is delayed because orders 4 and 5 are prioritized.

To better serve their customers, companies may want to adopt allocation policies \( \pi \) that are able to determine \( \sigma^{\pi}_{\tau_i} \) based on information available at time \( \tau_i \). As a consequence, such policies can effectively commit to a certain delivery time at the moment that the customer arrives; hence, we will say that such policies commit upon arrival. Note that the example implies that the CUD policy \( \pi(CUD) \) does not commit upon arrival.

3.2.2. Commit upon Arrival with FIFO Allocation. We next discuss an easy to implement policy that commits to a specific delivery time upon arrival of a customer.

This simple policy physically or virtually sets inventory apart when customers place orders for some future RST. Inventory that is set apart is preallocated to specific customer orders, and the policy must keep track of which units have been preallocated and which units remain unallocated.

More precisely, the simple policy sets an unallocated on-hand unit apart each time a customer order (say, order \( i \)) arrives, and this unit is preallocated to that order. At \( t = \tau_i \), the unit is shipped to customer \( i \) if no unallocated on-hand inventory is available when order \( i \) arrives at time \( \tau_i \). If no unallocated on-hand inventory is available when order \( i \) arrives at time \( \tau_i \), then the company preallocates the unallocated unit that is part of a pipeline replenishment order that is due to arrive earliest.

When there is neither unallocated on-hand inventory available at \( \tau_i \) nor unallocated units that are part of pipeline replenishment orders at \( \tau_i \), then the company preallocates an unallocated unit that is part of one of the replenishment orders that will be placed during the first review period \([\tau_i] \) after \( \tau_i \). These replenishments will arrive at \([\tau_i] + L \); note that some units in those replenishment orders are guaranteed to be unallocated because \( s > 0 \).

The CUA-FIFO policy has the considerable advantage that it is relatively easy to implement in practice. Setting inventory apart physically may be a very convenient approach for ensuring that promised deliveries can be made (see also Marklund 2006).

We illustrate this allocation policy in Figure 2, which describes the same example as before \((I_0 = 1, L = 4, and
3.2.3. CUA with LCU Allocation. So, we can improve customer service compared with the CUA-FIFO policy by instead preallocating the LCU. For a customer order $i$ with given $r_i$, the LCU is the unallocated unit that is part of a replenishment order and that is planned to arrive latest but before $r_i$. If no such order is available, then the LCU is an unallocated on-hand unit. If all on-hand inventory is also already preallocated, then no unallocated inventory is available that can meet the deadline $r_i$. In this case, an unallocated unit in a replenishment order is allocated to order $i$ that minimizes the lateness $o_i - r_i$ of the order; this unit may already be in replenishment at $r_i$ or it may be part of a replenishment order that will be placed at $r_i$.

For customer orders $i$ with a demand lead time that exceeds the supply lead time $L$, some care is needed to appropriately allocate the least-critical unit. Indeed, when $r_i > \tau_i + L$, then the LCU for order $i$ may not have been ordered yet at $\tau_i$. Accordingly, when such an order $i$ arrives at $\tau_i$, the company commits to delivering the order at $r_i$ and in period $r_i - L$, the least-critical unit can be allocated in line with the discussion. By this time, all units that can meet order $i$ are guaranteed to be on hand or in the pipeline. (Also, note that there will be at least one unallocated unit in the pipeline at $r_i - L$ because $s > 0$.)

The example shows that the CUA-FIFO policy has the disadvantage that preallocating on-hand inventory may needlessly reduce the service experience of some customers. In particular, if we had preallocated to order 1 the unit ordered at $t = 1$ (instead of an on-hand item), then we could still have guaranteed $o_1 = r_1$. Moreover, we could have used the on-hand unit to meet customer order $i = 2$ to guarantee $o_2 = r_2$.

3.2.3. CUA with LCU Allocation. So, we can improve customer service compared with the CUA-FIFO policy by instead preallocating the LCU. For a customer order $i$ with given $r_i$, the LCU is the unallocated unit that is part of a replenishment order and that is planned to arrive latest but before $r_i$. If no such order is available, then the LCU is an unallocated on-hand unit. If all on-hand inventory is also already preallocated, then no unallocated inventory is available that can meet the deadline $r_i$. In this case, an unallocated unit in a replenishment order is allocated to order $i$ that minimizes the lateness $o_i - r_i$ of the order; this unit may already be in replenishment at $r_i$ or it may be part of a replenishment order that will be placed at $r_i$.

For customer orders $i$ with a demand lead time that exceeds the supply lead time $L$, some care is needed to appropriately allocate the least-critical unit. Indeed, when $r_i > \tau_i + L$, then the LCU for order $i$ may not have been ordered yet at $\tau_i$. Accordingly, when such an order $i$ arrives at $\tau_i$, the company commits to delivering the order at $r_i$ and in period $r_i - L$, the least-critical unit can be allocated in line with the discussion. By this time, all units that can meet order $i$ are guaranteed to be on hand or in the pipeline. (Also, note that there will be at least one unallocated unit in the pipeline at $r_i - L$ because $s > 0$.)

The example shows that the CUA-FIFO policy has the disadvantage that preallocating on-hand inventory may needlessly reduce the service experience of some customers. In particular, if we had preallocated to order 1 the unit ordered at $t = 1$ (instead of an on-hand item), then we could still have guaranteed $o_1 = r_1$. Moreover, we could have used the on-hand unit to meet customer order $i = 2$ to guarantee $o_2 = r_2$.

4. Classical Performance Measures

We consider the replenishment policy and initial conditions (i.e., $s, Q,$ and $l_0$) to be given, and we look into the following measures of performance for allocation policies $\pi \in \Pi$.

- The expected average delay for customer orders is
  \[
  EW(\pi) := E\left(\frac{1}{N(T)} \sum_{i=1}^{N(T)} (\sigma_i - r_i)\right).
  \] (1)

- The expected fill rate on customer orders (i.e., the fraction of RSTs that are met) is
  \[
  EF(\pi) := 1 - E\left(\frac{1}{N(T)} \sum_{i=1}^{N(T)} 1[\sigma_i > r_i]\right),
  \] (2)
  where $1[\sigma_i > r_i] = 1$ if $\sigma_i > r_i$ and $1[\sigma_i > r_i] = 0$ otherwise.

In case $N(t) = 0$, then we define the quantity inside the expectation on the right hand side (RHS) as zero in both cases. As mentioned in the previous section, we expect the $\pi(CUD)$ policy to perform well in general. It has the advantage that after each allocation round $t \in \mathbb{N}$, there is either zero inventory or zero unmet orders (i.e., $S(t) = A(t)$). It should come as no surprise that this leads to good performance in terms.
of the average waiting time $EW$ for customers. In fact, for given $s$, $Q$, and $I_0$, the policy minimizes the average customer delay $EW$ over the entire class of policies $\Pi$.

**Theorem 1.** For the CUD policy $\pi(CUD)$, we have $\pi(CUD) \in \arg\min_{\pi \in \Pi} EW(\pi)$.

The proof of Theorem 1 is included in the appendix; our formal proof of Theorem 1 rigorously underpins assumptions that underlie previous literature (see, e.g., Hariri and Zipkin 1995).

Theorem 1 implies that in terms of expected waiting time, the $\pi(CUD)$ policy is superior to the $\pi(CUA-FIFO)$ and $\pi(CUA-LCU)$ policies because $\pi(CUA-FIFO)$, $\pi(CUA-LCU) \in \Pi$. We may wonder about the fill-rate performance of the three policies. We now come to the main result of this section.

**Theorem 2.** $EF(\pi(CUA-LCU)) \geq \max[EF(\pi(CUA-FIFO)), EF(\pi(CUD))]$.

The proof of Theorem 2 and all other proofs are included in the online appendix. One of the challenges in proving Theorem 2 is getting a grip on the fill rate of specific policies. To overcome this challenge, we instead prove a stronger result that yields Theorem 2 by specialization. The stronger result is presented next because it is of interest in its own right.

**4.1.1. Sensible Policies.** We use the insight that although the three investigated policies ($\pi(CUD)$, $\pi(CUA-FIFO)$, $\pi(CUD)$) differ on important points, they all exhibit a sensible approach for clearing back orders in the following sense.

**Definition 2.** An allocation policy $\pi \in \Pi$ is sensible if for all $t \in \mathbb{N}$ and for any late order $i \in A^l(t)$ (i.e., with $r_i < t$), one of the following two conditions holds.

1. Order $i$ is satisfied now ($i \in S^l(t)$).
2. Any other orders $i' \in S^l(t)$ satisfied now are either delayed by more than $i$ ($r_{i'} < r_i$) or they arrived earlier than $i$ ($r_{i'} < r_i$).

We denote the family of sensible allocation policies $\pi$ by $\Pi_s$. Thus, $\Pi_s \subseteq \Pi$.

It is not difficult to show that $\pi(CUD)$, $\pi(CUA-FIFO)$, $\pi(CUA-LCU) \in \Pi_s$. Clearly, Theorem 1 then implies $\pi(CUD) \in \arg\min_{\pi \in \Pi_s} W(\pi)$. Analogously, we may ask ourselves which sensible policy then maximizes the fill rate. The following result answers that question.

**Theorem 3.** For the CUA-LCU policy $\pi(CUA-LCU)$, we have $\pi(CUA-LCU) \in \arg\max_{\pi \in \Pi_s} EF(\pi)$.

Note that this directly implies the claim in Theorem 2. So, restricting attention to sensible policies can be seen as a proof idea toward proving the claim stated in Theorem 2.

However, Theorem 3 is interesting beyond implying Theorem 2. Indeed, the conditions imposed in Definition 2 are intuitive. If an order is less entitled to receive a unit of inventory than another order (because it arrived later and it has a later due date), then it should not be satisfied before the other order. Companies may in fact wish to rule out any nonsensible allocation policies for two reasons: (1) Nonsensible policies may be perceived as very unfair by customers, and (2) under nonsensible policies, a customer who cancels a delayed order and later places that order again may receive better service than a customer who waits patiently (i.e., nonsensible policies may promote cancellations). Thus, finding the sensible policy that maximizes the fill rate is interesting from a practical point of view.

Another matter of interest is that although there are many policies in $\Pi$ that do not commit on arrival, the policy $\pi(CUA-LCU) \in \Pi$ that maximizes the fill rate does commit on arrival. So, a key insight here is that although committing upon arrival bears costs in terms of increasing the average customer delays (cf. Theorem 1), it does not have an adverse impact on the fill rate in many practical settings where policies need to be sensible. Indeed, we can increase the fill rate when moving from the noncommitting CUD policy to the committing CUA-LCU policy.

Theorems 1 and 3 together imply that $\pi(CUA-LCU)$ delays fewer customers, but the delay incurred by customers who are delayed is higher on average. A relevant question is then which customers are delayed more under $\pi(CUA-LCU)$, and this question is answered in the next section.

**4.2. Stratified Performance Measures**

In the previous subsection, we investigated the ability of policies that commit upon arrival to achieve good overall performance. We found that, perhaps surprisingly, $\pi(CUA-LCU)$ improves the fill-rate performance compared with the benchmark $\pi(CUD)$. However, we also found that the average delay increases in $\pi(CUA-LCU)$ compared with $\pi(CUD)$. These results imply that deciding between $\pi(CUD)$ and $\pi(CUA-LCU)$ involves a subtle trade-off. When we think closely about this trade-off, we observe that both allocation policies treat customers differently depending on their customer order lead time. Because customers with long customer order lead times may have different characteristics from customers with short customer order lead times, companies need to be aware of these subtleties when selecting an allocation policy.

Remember that customer $i$ arrives at time $\tau_i$ and has associated requested shipment date $r_i = [\tau_i + \ell_i]$. In order to obtain crisp insights to illustrate the subtleties mentioned, in this section we will restrict attention to a situation with two types of customers. Type 1 customers have $\ell_i = l_1$, and type 2 customers have $\ell_i = l_2$, where $l_1, l_2 \in \{0, 1, \ldots\}$ and $l_1 < l_2$. More precisely, we let $P(\ell = l_1) = \theta$ and $P(\ell = l_2) = 1 - \theta$, where $\theta$ is the
probability that a customer is of type 1 and $\ell_i - \ell$ for all $i$. For any customer demand (and customer demand lead time) realization, let $I_k = \{ i \in \{1, \ldots, N(T)\} \mid \ell_i = \ell \}$ be the set of type $k = 1, 2$ customers.

Then, for this special case, we can define the following stratified performance measures for $k = 1, 2$.

- The expected average delay for type $k$ orders is
  \[
  EW_k(\pi) := \mathbb{E}\left(\frac{1}{|I_k|} \sum_{i \in I_k} (\sigma_i^n - r_i) \right).
  \] (3)

- The expected fill rate on type $k$ orders is
  \[
  EF_k(\pi) := 1 - \mathbb{E}\left(\frac{1}{|I_k|} \sum_{i \in I_k} 1[\sigma_i^n > r_i] \right).
  \] (4)

If $|I_k| = 0$, we define the quantity inside the expectation on the RHS as zero in both cases.

We have the following results.

**Theorem 4.** The relative performance of the two policies for the two customer classes is as follows.

1. For the type 1 customer class, $\pi(CUD)$ gives superior fill rates and waiting times compared with $\pi(CUA-LCU)$ ($EF_1(\pi(CUA-LCU)) \leq EF_1(\pi(CUD))$ and $EW_1(\pi(CUA-LCU)) \geq EW_1(\pi(CUD))$).

2. For the type 2 customer class, $\pi(CUA-LCU)$ gives superior fill rates and waiting times compared with $\pi(CUD)$ ($EF_2(\pi(CUA-LCU)) \geq EF_2(\pi(CUD))$ and $EW_2(\pi(CUA-LCU)) \leq EW_2(\pi(CUD))$).

Note that the type 2 customers place their orders far in advance, whereas the type 1 orders have a shorter lead time. Hence, when comparing the two policies, $\pi(CUD)$ is favorable to the short demand lead time customers, whereas $\pi(CUA-LCU)$ is favorable to the long demand lead time customers. In that sense, companies may wish to adopt $\pi(CUA-LCU)$ allocation. It rewards customers who place their orders long in advance with higher fill rates and lower waiting times. In this way, companies can induce their customers to place orders with a longer demand lead time.

Companies may aggregate performance over multiple customer segments via composite key performance indicators (KPIs). We next analyze a common example, viz. the weighted waiting time:

\[
EW_{1,2}(\pi) := \mathbb{E}\left[ \frac{\alpha_1 (\sum_{i \in I_1} (\sigma_i^n - r_i)) + \alpha_2 (\sum_{i \in I_2} (\sigma_i^n - r_i))}{\alpha_1 |I_1| + \alpha_2 |I_2|} \right].
\] (5)

Here, $\alpha_1, \alpha_2 > 0$ are weights that represent the relative importance of type 1 and type 2 customers, respectively. For example, if $\alpha_2$ is larger than $\alpha_1$, then the waiting time for a customer who places an order far in advance carries more weight in the composite KPI.

We investigate the performance of $\pi(CUA-LCU)$ and $\pi(CUD)$ when this is the case.

**Theorem 5.** For $\alpha_2 \geq (l_2 - l_1 + 1)\alpha_1$, we have $EW_{1,2}(\pi(CUA-LCU)) \leq EW_{1,2}(\pi(CUD))$.

Hence, it turns out that even though $\pi(CUA-LCU)$ is outperformed by $\pi(CUD)$ when considering the average waiting time criterion $EW(\cdot)$, it outperforms $\pi(CUD)$ on the composite KPI $EW_{1,2}(\cdot)$ when customers who place orders far in advance carry sufficient weight. For example, if type 1 business-to-consumer (B2C) customers demand next-day delivery (meaning that the items must be ready for shipment the same evening; i.e., $l_1 = 0$) and type 2 customers place their orders a week in advance (i.e., $l_2 = 6$), then $\pi(CUA-LCU)$ is guaranteed to outperform $\pi(CUD)$ on the composite KPI when $\alpha_2 > 7\alpha_1$.

Finally, we investigate what happens when customers with a shorter order lead time are more important to the company. As expected, in those cases, companies should generally prefer $\pi(CUD)$.

**Theorem 6.** For $\alpha_1 \geq \alpha_2$, we have $EW_{1,2}(\pi(CUD)) \leq EW_{1,2}(\pi(CUA-LCU))$.

Taken together, Theorems 5 and 6 demonstrate that $\pi(CUA-LCU)$ tends to be preferable when the customers who place their orders far in advance are more important in terms of avoiding waiting time. From the perspective of customers, arguably it makes sense for a customer who places an order a week in advance to expect to receive all goods on time and be inconvenienced to a larger extent when the order is delayed. Hence, such a customer should typically carry more weight in a composite KPI.

Remember that $\pi(CUA-LCU)$ has an additional characteristic that may be considered; it increases the number of customer orders that are delivered on time by Theorem 3. Because a delay, regardless of its length, may involve rescheduling activities, this may in some cases be an additional reason to adopt $\pi(CUA-LCU)$.

5. Committing Now or Later: Numerical Study

In this section, we provide further insights into the different allocation policies by simulating them and quantifying their performance in terms of service level and advance inducement. This is conducted for a variety of scenarios based on real data from a large food retailer. Also, some sensitivity analysis is performed to extrapolate the obtained results.

The data we have collected from the food retailer concern two fulfillment centers and all online orders in the period between October 2016 and September 2017. Regarding the former, we have collected information of supply lead time (in days) and the policy...
parameters (reorder point and order quantity) for every pair \( \langle \text{item}, \text{FC} \rangle \). A total of 39,240 unique items exist, amounting to 74,556 records across the two FCs (note that a few items exist in only one FC). The online orders consisted of 455,663 records and included information on the time the order was placed, the requested shipment time, and other information (such as shopping basket composition, allocated FC, delivery fee, and delivery location).

5.1. Simulation Design

From the examples explored in the previous section, we can observe that CUA-LCU appears to be more effective than CUD when two conditions are met: (i) orders crossover, and (ii) there are multiple outstanding replenishment orders. The frequency of order crossovers depends on the distribution of the demand lead time \( \ell \). The number of outstanding orders is determined by the ratio \( \lambda L/nQ \). Figure 4 shows the distribution of these two factors in the case of our food retailer. The distribution of the demand lead time exhibits a strongly asymmetric shape, with a mean of 0.844 and a coefficient of variation (CV) of 0.703. The ratio \( \lambda L/nQ \) can be as low as 0.1 and reach values greater than 5. However, more than 99% of the items are in the range of 0.5–4. Therefore, we consider the following geometric sequence \( (0.25, 0.5, 1, 2, 4, \text{and } 8) \), which extrapolates that range one step on each side.

We also perform sensitivity analysis on the distribution of \( \ell \) by adding and subtracting 20% to the mean and CV. This gives a total of nine distributions (D1–D9) whose parameters are provided in Table 1 (the real case corresponds to distribution D5). By performing a full factorial experiment combining the nine distributions and the six levels of \( \lambda L/nQ \), we generate a total of 54 scenarios, of which 6 concern the base case (with the mean and CV of the real case).

Because we do not specify a cost structure, we do not optimize the replenishment policy parameters \( s \) and \( nQ \). Instead, we simulate the different allocation policies with the same replenishment policy. For each scenario, \( s \) is determined such that a fill rate of around 95% is achieved (the values of \( s \) for the 54 scenarios are provided in Table 2). The value of \( nQ \) results from the ratio \( \lambda L/nQ \), where \( L \) is one and \( \lambda \) is eight for all scenarios (this is similar to the real average lead time demand and provides a value of one for \( nQ \) when the ratio reaches eight, thus avoiding rounding approximations).

### Table 1. Parameters of the \( \ell \) Distributions

| \( p(x) \) | 0.089 | 0.583 | 0.179 | 0.079 | 0.038 | 0.020 | 0.012 | — | — |
| \( x \)  | 0.255 | 0.511 | 0.772 | 1.159 | 1.545 | 1.931 | 2.317 | 0.676 | 0.562 |
| \( x^2 \) | 0.218 | 0.437 | 0.874 | 1.310 | 1.747 | 2.184 | 2.621 | 0.676 | 0.703 |
| \( x^3 \) | 0.183 | 0.365 | 0.971 | 1.457 | 1.943 | 2.429 | 2.914 | 0.676 | 0.843 |
| \( x^4 \) | 0.319 | 0.639 | 0.965 | 1.448 | 1.931 | 2.414 | 2.896 | 0.844 | 0.562 |
| \( x^5 \) | 0.273 | 0.546 | 1.092 | 1.638 | 2.184 | 2.730 | 3.276 | 0.844 | 0.703 |
| \( x^6 \) | 0.228 | 0.456 | 1.214 | 1.821 | 2.429 | 3.036 | 3.643 | 0.844 | 0.843 |
| \( x^7 \) | 0.383 | 0.766 | 1.159 | 1.738 | 2.317 | 2.896 | 3.476 | 1.013 | 0.562 |
| \( x^8 \) | 0.328 | 0.655 | 1.310 | 1.966 | 2.621 | 3.276 | 3.931 | 1.013 | 0.703 |
| \( x^9 \) | 0.274 | 0.548 | 1.457 | 2.186 | 2.914 | 3.643 | 4.371 | 1.013 | 0.843 |

### Table 2. Reorder Points

| \( \lambda L/nQ \) | 0.25 | 0.5 | 1 | 2 | 4 | 8 |
| \( \ell \) | D1 | 1 | 3 | 4 | 5 | 6 |
| D2 | 2 | 3 | 4 | 5 | 6 | 6 |
| D3 | 0 | 2 | 3 | 4 | 4 | 4 |
| D4 | 1 | 2 | 3 | 4 | 4 | 5 |
| D5 | 1 | 3 | 4 | 4 | 5 | 5 |
| D6 | 0 | 1 | 1 | 2 | 3 | 3 |
| D7 | 0 | 1 | 2 | 3 | 3 | 4 |
| D8 | 1 | 2 | 3 | 4 | 4 | 5 |
The customers arrival follows a homogeneous Poisson process, and each order requests a single unit.

Our simulation focuses on estimating the performance measures introduced in Section 4.1. Let $\bar{W}$ and $\bar{F}$ denote the Monte Carlo estimates of the expected waiting time $E[W]$ and expected fill rate $E[F]$, respectively. To ensure a consistent interpretation where low values are good, we report results for the nonfill-rate fraction $1 - \bar{F}$. We also occasionally report the estimated on-hand inventory $\bar{I}$. The simulation length is determined for each scenario using a confidence interval of 0.1% on the nonfill-rate fraction, with a 95% confidence level.

5.2. Service-Level Performance

We start by looking at average results across the six scenarios of the base case. Table 3 summarizes the performance of the different allocation policies. Note that in all these cases, the underlying replenishment policy uses all the available ADI. We also tested not using ADI, and for the same reorder points, it resulted in 10 times more stockouts, which demonstrates the importance of getting as much ADI as possible. Regarding the allocation policies, CUA-FIFO is dominated by the other policies for all the scenarios in all the metrics, especially on the number of stockouts. For the other two, as expected, there is a trade-off between stockouts and average delay. CUD has considerably shorter delays, whereas CUA-LCU fails with fewer customers.

As previously mentioned, these results were generated for values of $s$ such that fill rate would be around 95%. We have run the same tests targeting 90% and 97.5%, but the obtained relative differences were similar.

The average results provide a general idea, but the actual differences of performance, namely with respect to fill rate, highly depend on the scenario. We plot in Figure 5(a) the performance for different levels of outstanding orders (given by $\lambda L/nQ$). In the y-axis, we represent the advantage of CUA-LCU over CUD, with respect to the number of stockouts ($1 - \bar{F}$). As expected, when $\lambda L/nQ \leq 1$, CUA-LCU has virtually no advantage over CUD in terms of the number of stockouts. Then, it grows logarithmically with that ratio, reaching around 20% for a ratio of eight.

By using the distribution of ratios displayed in Figure 4(b), we find that in the real case, the average improvement of CUA-LCU is around 4%. In Figure 5(b), we show this distribution (i.e., for each level of $\lambda L/nQ$, the orders that CUA-LCU can prevent from being delayed when compared with CUD). From the chart, we can see that 70% of that advantage is obtained by one-third of the items (those with $2 \leq \lambda L/nQ \leq 3.5$).

In order to understand whether these results will generalize to other scenarios, we perform sensitivity analysis on the distribution parameters of $\ell$.

In Figure 6, we can see that the advantage of CUA-LCU tends to increase with the mean of $\ell$ but apparently not with $CV$. This seems to indicate that, even when $CV$ reduces 20%, there is still enough variability to produce a good amount of order crossovers.

Contrary to fill rate, the advantage of CUD regarding the average delay ($\bar{W}$) is not substantially impacted by $\lambda L/nQ$. Therefore, at least for cases where $\lambda L/nQ \geq 1$, the selection of the policy should depend on the values that are specified for both fixed and variable back-order costs. In the other cases, CUD might be

<table>
<thead>
<tr>
<th>Policy</th>
<th>$1 - \bar{F}, %$</th>
<th>$\bar{W}$</th>
<th>$\bar{I}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUD</td>
<td>5.0</td>
<td>0.04</td>
<td>6.3</td>
</tr>
<tr>
<td>CUA-FIFO</td>
<td>11.8</td>
<td>0.59</td>
<td>7.8</td>
</tr>
<tr>
<td>CUA-LCU</td>
<td>4.6</td>
<td>0.23</td>
<td>6.3</td>
</tr>
</tbody>
</table>
better in terms of performance. However, this policy does not meet important strategic directives, such as committing early. Another critical directive is inducing advance in customer orders, which is following analyzed.

5.3. Advance Inducement

Theorem 4 already states that customers with longer lead times will get better service in terms of both fill rate and waiting time under CUA-LCU compared with CUD. This property should induce customers to provide their orders more in advance in order to obtain a better service. If customers do so, more ADI will be available, which then greatly improves the overall performance at every dimension. In this subsection, we illustrate numerically the difference in service that customers with different lead times will experience under each of the allocation policies.

Figure 7 plots, for each policy, the service provided to customer orders (either \(1 - \bar{F}\) or \(\bar{W}\)) depending on their lead time. Although both policies provide better service to customers with longer lead times, this positive discrimination is much more evident with CUA-LCU. The latter reduces continuously the probability of failing to customers as their lead time increases. This happens so that at some point, the service level of those customers is 100%. This is not guaranteed by the CUD policy, which always maintains some probability of failing to customers and hence, some delay.

6. Conclusion

With the changing landscape in grocery retail, customers are increasingly placing baskets of orders that they would like to receive at a planned and confirmed moment in time. This model is also increasingly adopted by other leading players in the industry and requires these retailers to also deliver missing items as back orders, like for instance, Target and Best Buy are doing. This new reality fundamentally changes the strategic management of inventory because the willingness of customers to order early provides an opportunity for cost reduction and service-level improvement for thin-margin online retailers. We demonstrate in this paper that online retailers have the tools at hand to do so; they should commit early to customer orders to enhance the customer service experience and eventually, to also create opportunities for reducing the cost of operations.

We explored three allocation policies both analytically and numerically. Two of them have already been studied in the literature; the third one was devised here. We compare these policies looking not only at their operational performance but also, at multiple strategic directives to which they may contribute. Moreover, we do not use a specified objective function. Instead, we show the relative performance of each policy in each indicator so that trade-offs can be

Figure 6. Advantage of CUA-LCU over CUD on \(1 - \bar{F}\) for Different \(\ell\) Distributions

![Figure 6](image)

Notes: (a) Probability of failing customer orders as a function of their lead time. (b) Average delay to customer orders as a function of their lead time.
made a posteriori. The indicators include fill rate, average delay, and a weighted variant of the latter. Table 4 summarizes the policies comparison according to the identified strategic directives and service-level performance.

We conclude that CUA-FIFO and CUA-LCU have all the desirable properties to meet the strategic directives. However, the latter dominates the former in all performance metrics. CUA-LCU is indeed shown to be the optimal “sensible” policy in terms of fill rate, whereas the traditional CUD minimizes the average waiting time. Note, however, that CUD does not meet the strategic directives. With the changes in online ordering and delivery policies taking place as we observe, it seems appropriate for online retailers to commit upon arrival, hence improving customer experience and delivering on promises made. This also enables the retailers to get a better handle on reducing the very high operations cost, namely by inducing customers to provide more advance demand information.

Our application in a grocery retail setting demonstrates that the effect size of our novel strategy is substantial and hence, would need to be considered by many retailers.

Acknowledgments
The authors thank the editorial team for their assistance in developing and improving the manuscript. Further, they thank the (anonymous) food retailer for making available their transaction data for this study.

Appendix. Proof of Theorem 1
To prove the results in the main text, we first prove some auxiliary results that are required for a rigorous analysis.

Throughout, let \( \omega \) (with \( \omega \in \Omega \)) denote a possible customer demand (and customer demand lead time) realization. With slight abuse of notation, let \( N(T) \) denote the total number of customers arriving for this demand realization. As a consequence of our assumptions, note that we have \( N(T) \in \mathbb{N} \) with probability 1; hence, for each realization, \( N(T) \) is a natural number. For \( i \in \{1, \ldots, N(T)\} \), let \( \tau_i \in \mathbb{R} \), \( r_i \in \mathbb{R} \) denote the corresponding realization of the random variables \( \tau_i \) and \( r_i \).

We first note that the replenishment policy acts independently of the allocation policy because replenishment decisions are taken based on the inventory position, which does not depend on allocation. Thus, we have Lemma A.1.

**Lemma A.1.** Consider any fixed reorder point \( s \), order quantity \( Q \), and initial inventory level \( I_0 \), and let \( \pi \in \Pi \) be any allocation policy. Denote by \( q_t^\pi \) the (random) replenishment quantity in period \( t \). (More precisely, let \( q_t^\pi(\omega) \in \mathbb{N} \) denote the replenishment quantity in period \( t \) for demand realization \( \omega \).) Then, \( \forall t : q_t^\pi(\omega) = q_t^\pi(\omega) \) for any two allocation policies \( \pi, \pi' \in \Pi \).

(In the lemma, the possible dependence of \( q_t^\pi \) on \( \pi \) is explicit through the superscript \( \pi \). The lemma verifies that \( q_t^\pi \) is in fact independent of \( \pi \); so, we revert back to the notation \( q_t \).)

**Proof of Lemma A.1.** Because orders are decided based on the inventory position, they are not influenced by allocation decisions as can be shown rigorously by a straightforward induction argument. We omit details for brevity. \( \square \)

Lemma A.1 implies that an allocation policy has some given amount of inventory available until time \( t \) that can be allocated to customer orders. To formalize this, recall that \( J(t) = I(t) - |S(t)| = I(t-1) + q_{t-1-L} - |S^\pi(t)| \). Because \( I(0) = I_0 \) by assumption, we find

\[
J(i) = I_0 + \sum_{i=1}^{\hat{i}} (q_{i-L} - |S^\pi(i)|),
\]

where \( \sum_{i=a}^{b} = 0 \) if \( b < a \). Rewriting yields

\[
\sum_{i=1}^{\hat{i}} |S^\pi(i)| = I_0 + \sum_{i=1}^{\hat{i}} q_{i-L} - J(i) \leq I^\pi(\hat{i}), \tag{A.1}
\]

where \( I^\pi(\hat{i}) := I_0 + \sum_{i=1}^{\hat{i}} q_{i-L} \) is independent of \( \pi \) and where we used that \( J(i) \geq 0 \) almost surely. Thus, the total amount \( \sum_{i=1}^{\hat{i}} |S^\pi(i)| \) satisfied until \( \hat{i} \) cannot exceed the total inventory \( I^\pi(\hat{i}) \) available until \( \hat{i} \).

Now, note that the first \( I_0 \) units of inventory are available to satisfy demand from period 0 onward; the subsequent \( q_1 \) units are available from period \( 1 + L \) onward (because \( q_t = 0 \) for \( t \leq 0 \)), and the subsequent \( q_2 \) units are available from period \( 2 + L \) onward, etc. In what follows, we will number the units in the order of arrival (i.e., the units \( 1, \ldots, I_0 \) correspond to initial on-hand inventory, and the units \( I_0 + 1, \ldots, I_0 + q_1 \) correspond to the units in the

<table>
<thead>
<tr>
<th>Policy</th>
<th>Committing</th>
<th>Advancement inducing</th>
<th>Fill rate</th>
<th>Average delay$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUD</td>
<td>−</td>
<td>−</td>
<td>**</td>
<td>***</td>
</tr>
<tr>
<td>CUA-FIFO</td>
<td>+</td>
<td>+</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>CUA-LCU</td>
<td>+</td>
<td>+</td>
<td>***</td>
<td>**</td>
</tr>
</tbody>
</table>

$^a$More asterisks mean lower average delays.

### Table 4. Comparison of Allocation Policies According to Strategic Properties and Service-Level Performance
first replenishment order, etc.). Then, the jth unit is available at
\[
a(j) := \min_{t \in \{0, 1, 2, \ldots \}} \{t | I^j(t) \geq j \}. \tag{A.2}
\]
Note that a(j) is random in the sense that it depends on the demand realization and that a(j) is increasing in j almost surely (we use increasing and decreasing in the nonstrict sense in this paper). By Lemma A.1, a(j) is independent of \( \pi \) for j \in N.

We next briefly discuss a technicality. For \( \pi \in \Pi \), remember that each order i \( \in \{1, \ldots, N(T)\} \) is satisfied eventually. Thus, by definition of \( A^j(t) \) and by \( S^j(t) \subseteq A^j(t) \), there is for every order i, a unique \( t \in \mathbb{N} \) such that \( i \in S^j(t) \). Thus, \( \alpha^j_i = t \in \mathbb{N} \) is uniquely defined, as claimed in the main text.

Recall that customer orders i \( \in \{1, \ldots, N(T)\} \) are numbered in the order of arrival and that the units of inventory \( j = 1, 2, 3, \ldots \) are numbered in the order in which they are available to satisfy demand. The allocation policy \( \pi \) specifies the orders \( S^j(t) \) to be satisfied in period t. Now, for writing our proofs, it will be convenient to introduce a rule that specifies which specific unit j is used to satisfy each order i. For this, we will use the FIFO rule. Item j = 1 is used to satisfy the first order that is satisfied; item j = 2 is used to satisfy the second order that is satisfied, etc. (We note that FIFO here only pertains to bookkeeping the match between orders and units; the allocation policy \( \pi \) remains in place.) Remember that the orders \( S^j(t) \) delivered in period t are satisfied simultaneously, but for purposes of bookkeeping, we will assume that they are satisfied in the order in which they originally arrived to the inventory system. We introduce the function \( g_n(\cdot) \): it assigns to the ith customer order the unit numbered \( j = g_n(i) \) used to satisfy it.

**Example A.1.** Suppose allocation policy \( \pi \) satisfies the orders \( S^1 = \{2, 5, 6\} \) in period 1, the orders \( S^2 = \{1, 4\} \) in period 2, and the orders \( S^3 = \{3\} \) in period 3. Then, using FIFO allocation of the units to the orders, units 1, 2, and 3 are assigned to orders 2, 5, and 6; units 4 and 5 are assigned to orders 1 and 4; and unit 6 is assigned to order 3. This implies that \( g_n(1) = 4, g_n(2) = 1, g_n(3) = 6, g_n(4) = 5, g_n(5) = 2, \) and \( g_n(6) = 3 \).

In general according to these rules, the unit \( j = g_n(i) \) used to satisfy order i is given by
\[
g_n(i) := \sum_{t=1}^{\alpha^i - 1} |S^t(i)| + |\{i' \in S(\alpha^i) | i' \leq i \}|.
\]
(\note{\( |\{i' \in S(\alpha^i) | i' \leq i \}| \) is the number of orders that are satisfied in the same period as order i but that arrived to the system at or before the moment at which order i arrived.}

In line with the intuitive interpretation of \( g_n(\cdot) \), we would expect the following.

**Lemma A.2.** For \( \pi \in \Pi \), \( g_n(\cdot) \) satisfies the following equations almost surely
\[
1. \text{ For } i, \tilde{i} \in \{1, \ldots, N(T)\} \text{ subject to } i \neq \tilde{i}, \text{ we have } g_n(i) \in \{1, \ldots, N(T)\} \text{ and } g_n(\tilde{i}) \neq g_n(i).
2. \text{ We have } \alpha_n^i \geq \max\{r_n(i), a(g_n(i))\).
\]
The interpretation of the first property is that the units numbered \( 1, \ldots, N(T) \) are used to satisfy demand (which

consists of orders \( 1, \ldots, N(T) \) and that each of those units is used to satisfy at most one order. Recalling the definition of \( a(\cdot) \), the second property states that we cannot satisfy order \( i \) before the unit \( j = g_n(i) \) that is used to satisfy that order is available.

**Proof of Lemma A.2.** Property (1) holds by construction of \( g_n(\cdot) \). Now, property (2). Consider some arbitrary \( i \in \mathbb{N} \), which is satisfied at time \( \alpha^i_n \) such that \( i \in S(\alpha^i_n) \). It follows from the definition of \( g_n(i) \) that \( g_n(i) \leq \sum_{t=1}^{\alpha^i_n} |S^t(i)| \). Substituting \( \alpha^i_n \) for \( \tilde{i} \) in (A.1), we find \( \sum_{t=1}^{\alpha^i_n} |S^t(i)| \leq I^\tilde{i}(\alpha^i_n) \). Hence, \( g_n(i) \leq I^\tilde{i}(\alpha^i_n) \). However, together with (A.2), this implies that
\[
a(g_n(i)) = \min_{t \in \{0, 1, 2, \ldots \}} \{t | I^\tilde{i}(t) \geq g_n(i) \} \leq \alpha^i_n.
\]
Combining this with \( \alpha^i_n \geq \alpha^i \) proves the claim.

Now, note that for any \( \pi \in \Pi \), the function \( g_n(\cdot) \) is a permutation of \( \{1, \ldots, N(T)\} \) (i.e., a bijection of the set \( \{1, \ldots, N(T)\} \) onto itself). (Equivalently, \( g_n(\cdot) \) is a mapping from the orders \( \{1, \ldots, N(T)\} \) to the units used to satisfy those orders \( \{1, \ldots, N(T)\} \).) In particular, let \( \Xi_{N(T)} \) denote the set of permutations of \( \{1, \ldots, N(T)\} \). Then, for any \( \omega \in \Omega \) and any \( \pi \in \Pi \), we have \( g_n(\cdot) \in \Xi_{N(T)} \). With this insight, we obtain a lower bound on the minimal waiting time that can be obtained for some given demand realization \( \omega \in \Omega \). To be more specific, let \( W(\pi) = \frac{1}{N(T)} \sum_{i=1}^{N(T)} (\alpha^i_n - r_i) \) denote the average ("realized") waiting time for the demand realization \( \omega \) when applying policy \( \pi \). Then, we find for any demand realization \( \omega \in \Omega \) and \( \pi \in \Pi \) that
\[
N(T)W(\pi) \geq \hat{W}(g_n(\cdot)) \geq \min_{\sigma \in \Xi_{N(T)}} \hat{W}(\sigma(\cdot)). \tag{A.3}
\]
Here, \((x)^+\) denotes \max(0, x), and \( \hat{W}(\sigma(\cdot)) = \sum_{i=1}^{N(T)} (\alpha^i_n - r_i)^+ \), with \( \sigma(\cdot) \) defined in (A.2). The first inequality is by (1) and Lemma A.2(2), and the second inequality holds because \( g_n(\cdot) \in \Xi_{N(T)} \) by Lemma A.2(1).

We are now ready to prove Theorem 1. We note that high-level proofs of the theorem may very well exist; we instead present a constructive and detailed proof, elements of which will be reused in Theorem 3.

**Proof of Theorem 1.** Note that in period \( a(g_n(CUD)(i)) \), the unit used to satisfy order i is available. Thus and because \( \pi(CUD) \) never has inventory and back orders at the end of any period, we find that order i is satisfied at \( \max(r_n(i), \sigma(g_n(CUD)(i))) \) under policy \( \pi(CUD) \). This establishes that, for any \( \omega \in \Omega \), we have
\[
N(T)W(\pi(CUD)) = \hat{W}(g_n(CUD)(\cdot)). \tag{A.4}
\]
We next show that \( \pi(CUD) \in \arg\min_{\pi \in \Xi_{N(T)}} \hat{W}(\sigma(\cdot)) \). To this end, we will construct, for any \( \sigma \in \Xi_{N(T)} \), a sequence of permutations \( \sigma_k \in \Xi_{N(T)} \) for \( k = 1, \ldots, n \) such that \( \sigma_1 = \sigma \) and \( \hat{W}(\sigma_n) = \hat{W}(g_n(CUD)) \) and such that \( \hat{W}(\sigma_{k+1}) = \hat{W}(\sigma_k) \) for \( k = 1, \ldots, n - 1 \). We construct the sequence recursively. For any \( \sigma \in \Xi_{N(T)} \), set \( \sigma_1 = \sigma \). After defining \( \sigma_1(\cdot) \), let
\[
\hat{i}_k := \arg\min_{i \in \{1, \ldots, N(T)\}} \{g_n(CUD)(i) | \sigma_k(i) \neq g_n(CUD)(i)\}.
\]
That is, $i_k$ is an order that is mapped to different units under $\sigma_i(\cdot)$ and $g_{n,CUD}(\cdot)$; of these orders, it is the order that, under the $\pi(CUD)$ policy, is satisfied using the unit that arrives earliest. Now, there are two options. (1) If $i_k = \infty$ (if and only if $\sigma_i(\tilde{i}_k) = g_{n,CUD}(\tilde{i}_k)$ for $i \in \{1, \ldots, N(T)\}$), then set $n = k$; the complete sequence becomes $\sigma_1, \ldots, \sigma_n$. (2) If $i_k < \infty$, then let $\tilde{i}_k := \sigma_i(\tilde{i}_k), \tilde{j}_k := g_{n,CUD}(\tilde{i}_k)$, and $\tilde{i}_k := \sigma_i^{-1}(\tilde{i}_k)$ such that $\tilde{i}_k = g_{n,CUD}(\tilde{i}_k)$. Now, define $\sigma_{k+1}(\cdot)$ as follows: $\sigma_{k+1}(\tilde{i}_k) := \tilde{j}_k$. For $i \in \{1, \ldots, N(T)\}$, $\sigma_{k+1}(i) := \sigma_i(i)$.

To show that this sequence indeed has the desired properties, we first show that $\tilde{j}_k > \tilde{i}_k$ and that $r_{i_k} \leq r_{i_k}$. For the first inequality, note that $\tilde{j}_k \neq \tilde{i}_k$ by the definition of $i_k$ and because $\tilde{i}_k = g_{n,CUD}(\tilde{i}_k)$. Now, let $i' \neq g_{n,CUD}(\tilde{i}_k)$. Then, by construction of $\tilde{i}_k$, we must have $\sigma(i') = g_{n,CUD}(i')(= \tilde{i}_k)$. However, this implies $\sigma(i') = \sigma(\tilde{i}_k)$, which contradicts $i' \neq \tilde{i}_k$. Thus, assuming $\tilde{j}_k < \tilde{i}_k$ also leads to a contradiction, and the only possibility that is left is thus that $\tilde{j}_k > \tilde{i}_k$.

Now, for the second inequality, we note that $g_{n,CUD}(\tilde{i}_k) \geq \tilde{j}_k \geq g_{n,CUD}(\tilde{i}_k)$. To see this, note that the reverse inequality $g_{n,CUD}(\tilde{i}_k) < \tilde{j}_k = g_{n,CUD}(\tilde{i}_k)$ would imply (by definition of $i_k$) that $g_{n,CUD}(\tilde{i}_k) = \sigma_i(\tilde{i}_k)$. However, $g_{n,CUD}(\tilde{i}_k) = \tilde{i}_k$ and $\tilde{j}_k \neq \tilde{i}_k$, which together imply $g_{n,CUD}(\tilde{i}_k) \neq \sigma_i(\tilde{i}_k)$. Thus, $g_{n,CUD}(\tilde{i}_k) < \tilde{j}_k$ leads to a contradiction. Moreover, $g_{n,CUD}(\tilde{i}_k) = \tilde{j}_k = g_{n,CUD}(\tilde{i}_k)$ contradicts that $\tilde{i}_k \neq \tilde{j}_k$. Thus, $g_{n,CUD}(\tilde{i}_k) > g_{n,CUD}(\tilde{i}_k)$. However, this last inequality implies that the CUD policy satisfies order $i_k$ before order $\tilde{i}_k$ and by definition, this policy satisfies orders with earlier RST before orders with later RST. Thus, this implies that $r_{i_k} \leq r_{i_k}$.

Now, consider any order $i$ for which $g_{n,CUD}(i) < \tilde{j}_k$. By definition of $i_k$, we must have $g_{n,CUD}(i) = \sigma_i(i)$, and because $g_{n,CUD}(\tilde{i}_k) = \tilde{j}_k \geq \tilde{i}_k$ and $g_{n,CUD}(\tilde{i}_k) = \tilde{i}_k > \tilde{j}_k$, this implies $\sigma_{k+1}(i) = \sigma_i(i) = g_{n,CUD}(i)$. Now, remember that $g_{n,CUD}(\tilde{i}_k) = \tilde{j}_k$ and that $\sigma_{k+1}(\tilde{i}_k) = \tilde{j}_k$; thus, $g_{n,CUD}(\tilde{i}_k) = \sigma_{k+1}(\tilde{i}_k)$. However, this implies that for all $i$ such that $g_{n,CUD}(i) \leq \tilde{j}_k$, we find $g_{n,CUD}(\tilde{i}_k) = \sigma_{k+1}(i)$. This implies that $g_{n,CUD}(\tilde{i}_k) > g_{n,CUD}(\tilde{i}_k)$. However, because the number of finite values that $g_{n,CUD}(\cdot)$ can have is finite, this implies that eventually the sequence terminates.

Now, by definition of $\hat{W}(\cdot)$, we find (with $r(\cdot)$ denoting $r$ for legibility)

$$
\hat{W}(\sigma_1) - \hat{W}(\sigma_{k+1}) = [a(\sigma_i(\tilde{i}_k)) - r(\tilde{i}_k)]^+ + [a(\sigma_i(\tilde{j}_k)) - r(\tilde{j}_k)]^+

= a(\sigma_{k+1}(\tilde{i}_k)) - r(\tilde{i}_k)]^+ + [a(\sigma_{k+1}(\tilde{j}_k)) - r(\tilde{j}_k)]^+

= [a(\tilde{j}_k) - r(\tilde{i}_k)]^+ + [a(\tilde{j}_k) - r(\tilde{j}_k)]^+

= [a(\tilde{j}_k) - r(\tilde{i}_k)]^+ + [a(\tilde{j}_k) - r(\tilde{j}_k)]^+ \geq 0.
$$

Here, the equality follows by substituting in the introduced notation, whereas the inequality follows because.


