

Inventory Management of a Fast-Fashion Retail Network

Felipe Caro * J eremie Gallien †

November 5, 2008

Abstract

Working in collaboration with Spain-based retailer Zara, we address the problem of distributing, over time, a limited amount of inventory across all the stores in a fast-fashion retail network. Challenges specific to that environment include very short product life-cycles, and store policies whereby an article is removed from display whenever one of its *key sizes* stocks out. To solve this problem we first formulate and analyze a stochastic model predicting the sales of an article in a single store during a replenishment period as a function of demand forecasts, the inventory of each size initially available and the store inventory management policy just stated. We then formulate a mixed-integer program embedding a piece-wise linear approximation of the first model applied to every store in the network, allowing us to compute store shipment quantities maximizing overall predicted sales, subject to inventory availability and other constraints. We report the implementation of this optimization model by Zara to support its inventory distribution process, and the ensuing controlled pilot experiment performed to assess the model’s impact relative to the prior procedure used to determine weekly shipment quantities. The results of that experiment suggest that the new allocation process increases sales by 3 to 4%, which is equivalent to \$275M in additional revenues for 2007, reduces transshipments, and increases the proportion of time that Zara’s products spend on display within their life-cycle. Zara is currently using this process for all of its products worldwide.

1. Introduction

The recent impressive financial performance of the Spanish group Inditex (its 2007 income-to-sales ratio of 13.3% was among the highest in the retail industry) shows the promise of the *fast-fashion* model adopted by its flagship brand Zara; other fast-fashion retailers include Sweden-based H&M, Japan-based World Co., and Spain-based Mango. The key defining feature of this new retail model lies in novel product development processes and supply chain architectures relying more heavily on local cutting, dyeing and/or sewing, in contrast with the traditional outsourcing of these activities to developing countries. While such local production obviously increases labor costs, it also provides greater supply flexibility and market responsiveness. Indeed, fast-fashion retailers offer in each season a larger number of articles produced in smaller series, continuously changing the assortment of products displayed in their stores: Ghemawat and Nueno 2003 report that Zara offers on average

*UCLA Anderson School of Management, Los Angeles, CA 90095, fcaro@anderson.ucla.edu

†MIT Sloan School of Management, Cambridge, MA 02142, jgallien@mit.edu

11,000 articles in a given season, compared to only 2,000 – 4,000 items for key competitors. This increases Zara’s appeal to customers: A top Zara executive quoted in Fraiman et al. 2002 states that Zara customers visit the store 17 times on average per year, compared to 3 to 4 visits per year for competing (non fast-fashion) chains. In addition, products offered by fast-fashion retailers may result from design changes decided upon as a response to actual sales information during the season, which considerably eases the matching of supply with demand: Ghemawat and Nueno 2003 report that only 15 – 20% of Zara’s sales are typically generated at marked-down prices compared with 30 – 40% for most of its European peers, with an average percentage discount estimated at roughly half of the 30% average for competing European apparel retailers.

The fast-fashion retail model just described gives rise to several important and novel operational challenges. The work described here, which has been conducted in collaboration with Zara, addresses the particular problem of distributing, over time, a limited amount of merchandise inventory between all the stores in a retail network. Note that while the general problem just stated is not specific to fast-fashion retailing, we believe that several features which are specific to this retail paradigm (short product life cycles, unique store inventory display policies) justify new approaches. Indeed, Zara’s interest in this area of collaboration was motivated by its desire to improve the inventory distribution process it was using at the beginning of our interaction for deciding the quantity of each article to be included in the weekly shipment from the warehouse to each store (see Figure 1 (a) for an illustration).

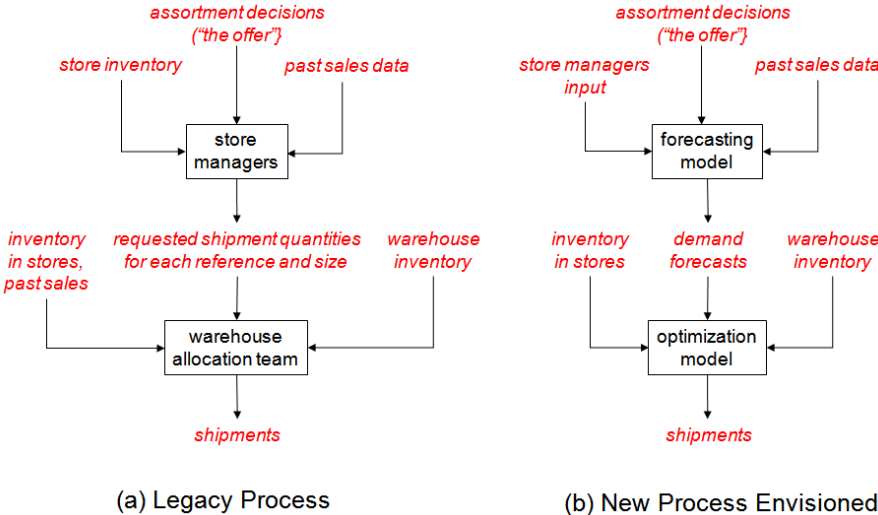


Figure 1: Legacy Process and New Process Envisioned to Determine Weekly Shipments to Stores.

According to that process, which we call the Legacy Process, each store manager would receive a weekly statement of the subset of articles available in the central warehouse for which he/she may request a shipment to his/her store. Note that this weekly statement (dubbed “the offer”)

would thus effectively implement any high-level assortment decision made by Zara’s headquarters for that particular store. However, it would not mention the total quantity of inventory available in the warehouse for each article listed. After considering the inventory remaining in their respective stores, store managers would then transmit back requested shipment quantities (possibly zero) for every size of every one of those articles. A team of employees at the warehouse would then reconcile all those requests by modifying (typically lowering) these requested shipment quantities so that the overall quantity shipped for each article and size was feasible in light of the remaining warehouse inventory.

At the beginning of our interaction, Zara expressed some concerns about the process just described, stating that while it had worked well for the distribution network for which it had been originally designed, the growth of its network to more than a thousand stores (and recent expansion at a pace of more than a hundred stores per year) might justify a more scalable process. A first issue centered on the incentives of store managers, whose compensation and career promotion prospects are driven to a significant degree by the total sales achieved in their stores. We believe that as a consequence store managers frequently requested quantities exceeding their true needs, particularly when suspecting that the warehouse may not hold enough inventory of a top-selling article to satisfy all stores (among others, Cachon and Lariviere 1999 study a stock rationing model capturing this behavior). Another issue is that store managers are responsible for a large set of tasks beyond determining shipment quantities, including building, sustaining and managing a team of several dozen sales associates in environments with high employee turnover, and are thus subject to important time pressures. Finally, we also believe that the very large amount of data that the warehouse allocation team was responsible for reviewing (shipments of several hundred articles offered in several sizes to more than a thousand stores) also created significant time pressures which made it challenging to balance inventory allocations across stores and articles in a way that would globally maximize sales. Motivated by these observations, we started discussing with Zara the alternative New Process for determining these weekly shipment quantities that is illustrated in Figure 1 (b). The New Process consists of using the shipment requests from store managers along with past historical sales to build demand forecasts. It then uses these forecasts, the inventory of each article and size remaining both in the warehouse and each store, and the assortment decisions as inputs to an optimization model having shipment quantities as its main decision variables.

The forecasting model considered takes as input from store managers their shipment requests, which is the very input they provide in the Legacy Process. This approach was believed to constitute the easiest implementation path, because it does not require any changes in the communication infrastructure with the stores or the store managers’ incentives. Note that Zara’s inventory distri-

bution process could be further improved in the future by introducing explicit incentives for the stores to contribute accurate forecasts.¹ However, the implementation reported here shows that substantial benefits can be obtained without any such change in the incentive structure.

While the forecasting component of the New Process provides a critical input, we also observe that the associated forecasting problem is a relatively classical one. In addition, the forecasting and optimization models supporting this new distribution process are relatively independent from each other, in that both may be developed and subsequently improved in a modular fashion. For these reasons, and for the sake of brevity and focus, the remainder of this paper is centered on the optimization component, and we refer the reader to Correa (2007) for more details and discussion on the forecasting model developed as part of this collaboration.

We proceed as follows: After a discussion of the relevant literature in §2, we discuss in §3 the successive steps we followed to develop the optimization model, specifically the analysis of a single-store stochastic model (§3.1) and then the extension to the entire network (§3.2). Section 4 discusses a pilot implementation study we conducted with Zara in order to assess the impact of our proposed inventory allocation process. Finally, we offer concluding remarks in §5. The (online) Appendix contains a technical proof, a validation of the store inventory display policy, a detailed computation of the financial impact, a model extension that considers articles offered in multiple colors, and some additional material related to the software implementation of this work.

2. Literature Review

The fast-fashion retail paradigm described in the previous section gives rise to many novel and interesting operational challenges, as highlighted in the case studies on Zara by Ghemawat and Nueno (2003), and Fraiman et al. (2002). However, we are aware of only one paper besides the present one describing an analytical model formulated to address an operational problem that is specific to fast-fashion companies. Namely, Caro and Gallien (2007) study the problem of dynamically optimizing the assortment of a store (i.e. which products it carries) as more information becomes available during the selling season. The present paper constitutes a logical continuation to that previous work since Zara’s inventory allocation problem takes the product assortment as an exogenous input (see Figure 1).

The generic problem of allocating inventory from a central warehouse to several locations satisfying separate demand streams has received much attention in the literature. Nevertheless, the optimal allotment is still an open question for most distribution system. When demand is assumed to be deterministic however, there are very effective heuristics with data-independent worst case

¹See Chen (2005) for a study on compensation packages that induce salespeople to forecast accurately and to work hard.

performance bounds for setting reorder intervals (see Muckstadt and Roundy 1993 for a survey). For the arguably more realistic case of stochastic demand that we consider here, available performance bounds depend on problem data. Focusing on stochastic periodic-review models (Zara replenishes stores on a fixed weekly schedule), Table 1 summarizes the main features of representative existing studies along with that of the present one. A first feature is the scope of inventory decisions considered: *ordering* refers to the replenishment of the warehouse from an upstream retailer; *withdrawal* to the quantity (and sometimes timing) of inventory transfers between the warehouse and the store network; and *allocation* to the split of any inventory withdrawn from the warehouse between individual stores. For a more exhaustive description of this body of literature, see Axsäter, Marklund and Silver (2002) or the earlier survey by Federgruen (1993).

	decision scope			time horizon		shortage model		retailers		
	<i>ordering</i>	<i>withdrawal</i>	<i>allocation</i>	<i>finite</i>	<i>infinite</i>	<i>backorder</i>	<i>lost sales</i>	<i>identical</i>	<i>non-identical</i>	<i>pull back display policy</i>
Eppen and Schrage (1981)	•		•		•	•		•		
Federgruen and Zipkin (1984)	•		•	•		•			•	
Jackson (1988)		•	•	•		•			•	
McGavin, Schwarz and Ward (1993)		•	•	•		•			•	
Graves (1996)	•	•	•		•		•		•	
Axsäter, Marklund and Silver (2002)	•	•	•		•	•		•		
this paper		•	•	•			•		•	•

Table 1: Main Features of Representative Periodic Review, Stochastic Demand Models for Inventory Management in Distribution Networks.

We observe that the operational strategy of fast-fashion retailers consists of offering through the selling season a large number of different articles, each having a relatively short life-cycle of only a few weeks. As a first consequence, the infinite horizon timeline assumed in some of the papers mentioned above does not seem appropriate here. Furthermore, typically at Zara a single manufacturing order is placed for each article, and that order tends to be fulfilled as a single delivery to the warehouse without subsequent replenishment. Ordering on one hand and withdrawal/allocation on the other thus occur at different times, and in fact, Zara uses separate organizational processes for these tasks. Consequently, we have chosen to not consider the ordering decisions and assume instead that the inventory available at the warehouse is an exogenous input (see Figure 1). While we do consider the withdrawal decisions, it should be noted that these critically depend in our model on an exogenous input by the user of a valuation associated with warehouse inventory, and any development of a rigorous methodology for determining the value of that parameter is beyond the scope of this work (see §3.2 for more details and discussion). We also point out that Zara stores do not take orders from their customers for merchandise not held in inventory, which seems to be part of a deliberate strategy (Fraiman et al. 2002). This justifies

the lost-sales model we consider.

The most salient difference between our analysis and the existing literature on inventory allocation in distribution networks is arguably that our model, which is tailored to the apparel retail industry, explicitly captures some dependencies across different sizes and colors of the same article. Specifically, in Zara stores (and we believe many other fast-fashion retail stores) a stockout of some selected *key* sizes or colors of a given article triggers the removal (or *pull back*) from display of the entire set of sizes or colors. While we refer the reader to §3.1.1 and §E for a more complete description and discussion of the associated rationale, that policy effectively strikes a balance between generating sales on one hand, and on the other hand mitigating the shelf space opportunity costs and negative customer experience associated with incomplete sets of sizes or colors. The literature we have found on these phenomena is scarce, but consistently supports the rationale just described: Zhang and Fitzsimons (1999) provide evidence showing that customers are less satisfied with the choice-process when, after learning about a product, they realize that one of the options is actually not available (as when a size in the middle of the range is not available and cannot be tried on). They emphasize that such negative perceptions affect the store’s image and might deter future visits. Even more to the point, the empirical study by Kalyanam et al. (2005) explores the implications of having key items within a product category, and confirm that they deserve special attention. Their work also suggests that stockouts of key items have a higher impact in the case of apparel products compared to grocery stores. We also observe that the inventory removal policy described above guarantees that when a given article is being displayed in store a minimum quantity of it is exposed, which is desirable for adequate presentation. In that sense, the existing studies on the broken assortment effect are also relevant (see Smith and Achabal 1998 and references therein).²

Finally, we point out that our goal was to develop an operational system for computing actual store shipment quantities for a global retailer, as opposed to deriving insights from a stylized model. Consequently, our model formulation sacrifices analytical tractability for realism, and our theoretical contribution is small relative to that of the seminal papers by Eppen and Schrage (1981) or Federgruen and Zipkin (1984) for example. In fact, the key approximation that our optimization model formulation implements was derived in essence by Federgruen and Zipkin (1984), whose analysis suggests that such approximation leads to good distribution heuristics (see §3.2). On the other hand, the present paper is the only one we are aware of which presents a controlled pilot implementation study for an inventory allocation model accounting for operational details in a large

²The broken assortment effect refers to the empirical observation that the sales rate for an article decreases as the inventory available diminishes even though that inventory may still be positive. This is explained by the reduced visibility of that article to customers in the store then, and the fact that popular sizes and colors of that article may become progressively harder to find.

distribution network (see §4). We also believe that the simple performance evaluation framework we developed when designing that study may be novel and potentially useful to practitioners.

3. Model Development

In this section we successively describe the two hierarchical models that we formulated to develop the optimization software supporting the New Process for inventory distribution discussed in §1. The first (§3.1) is descriptive and focuses on the modeling of the relationship between the inventory of a specific article available at the beginning of a replenishment period in a single store and the resulting sales during that period. The second model (§3.2) is an optimization formulation that embeds a linear approximation of the first model applied to all the stores in the network, in order to compute a globally optimal allocation of inventory between them.

3.1 Single Store Inventory-to-Sales Model.

3.1.1 Store Inventory Display Policy at Zara. In many clothing retail stores, an important source of negative customer experience stems from customers who have identified (perhaps after spending much time searching a crowded store) a specific article they would like to buy, but then cannot find their size on the shelf/rack (Zhang and Fitzsimons 1999). These customers are more likely to solicit sales associates and ask them to go search the backroom inventory for the missing size (increasing labor requirements), leave the store in frustration (impacting brand perception), or both. Proper management of size inventory seems even more critical to a fast-fashion retailer such as Zara, which offers a large number of articles produced in small series throughout the season. The presence of many articles with missing sizes would thus be particularly detrimental to the customers' store experience.

We learned through store visits and personal communications that Zara store managers tend to address this challenge by differentiating between *major* sizes (e.g. S,M,L), and *minor* sizes (e.g. XXS, XXL) when managing in-store inventory. Specifically, upon realizing that the store has run out of one of the major sizes for a specific article, store associates move all of the remaining inventory of that article from the display shelf/rack to the backroom and replace it with a new article, thus effectively removing the incomplete article from customers' sight.³ In contrast, no such action is taken when the store runs out of one of the minor sizes. The incomplete article that was removed might later return to the floor if missing sizes can be shipped again from the warehouse. Otherwise, it is either transferred to another store where the sizes are consolidated, or remains in the backroom until clearance sales.

³Weekly shipments to a Zara store include 10 – 20% of new articles. If a new article is not available to replace an incomplete one, the latter is still removed, but the store manager rearranges the articles remaining on the floor to avoid large empty spaces.

Zara does not have a product catalogue, and in fact strives to maintain among its customers a sense of scarcity and continuous assortment freshness (see §2). Consequently, customers do not typically enter a Zara store looking for a specific article, and do not expect articles not displayed on shelves/racks to still be available in the backroom. The store inventory removal policy just described can thus be seen as a balancing act between keeping inventory displayed to generate sales and mitigating the negative impact of missing sizes on brand perception.

Interestingly, the definition of major and minor sizes may reflect that some sizes (e.g. M) account for considerably more demand than others (e.g. XXL), but also more subtle psychological effects: when sizes XS, M and L of a given article are available but size S is not for example, S customers will tend to attribute that stockout to Zara’s mismanagement of its inventory. However, it appears that size XS customers will place less blame on Zara when a continuous set of sizes S, M and L is available but XS is not. This is because customers may not realize then that some units of that article were made in size XS in the first place (not all articles are offered in extreme sizes), and also because these customers may be blaming themselves instead for their own seemingly unusual dimensions. As a result, Zara managers seem to define as major sizes either a single size (e.g. M) or a continuous set of sizes (e.g. S,M,L) in the middle of the size range, even in (common) cases where an extreme size such as XS or XL accounts for more demand than S or M.

We also learned that the inventory removal rule just described was not prescribed by any formal policy imposed upon store managers, rather it constituted empirical observation of common store behavior. Because this seemed a key modeling issue, we decided to verify through analysis how prevalent this policy was. Specifically, in absence of available data for whether in-store inventory is located in the display area or in the backroom, we measured the adherence to the inventory display policy using the ratio DPF_j/DPA_j for each store j , where DPA_j is the number of days, summed over all articles, when there was a stockout of a major size but there was still some inventory available in another size, and DPF_j corresponds to the subset of those days characterized by the additional requirement that no sales were observed for any size.⁴ The details of this analysis are given in §B of the Appendix, and results are summarized in Figure 2, which shows the distribution of those ratios found across Zara’s entire network of approximately 900 stores (at the time when the data was collected). According to our metric, less than 2% of the stores showed an adherence lower than 80% with an average and median across stores both equal to 89%. We find these results to be quite compelling. In particular, they justify that the inventory display policy based on major sizes can be used as a representation of store execution behavior.

We next describe a stochastic model developed to answer the following question: Given the

⁴DPA and DPF stand for “Days when the Policy was Applicable” and “Days when the Policy was Followed”, respectively.

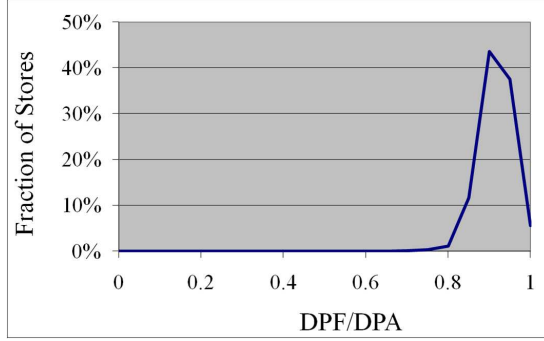


Figure 2: Histogram of the Adherence Ratios DPF_j/DPA_j Across All Stores.

dependency between inventory and sales of different sizes introduced by the store inventory management policy based on major sizes, how many sales of each article should be expected between successive replenishments when starting from a given initial profile of inventory across sizes? As part of this first modeling effort, we initially assume away the dependencies between inventory levels of different articles. That assumption is clearly not tenable in all cases, as there may be in practice significant demand substitution (e.g. garments available in different colors but otherwise identical) and demand complementarity (e.g. assorted vest and trousers sold separately) across articles. In section §E of the Appendix however, we discuss how our model may be extended to the case of products offered in multiple colors.

3.1.2 Model Description. Consider an article offered in a set of sizes $\mathcal{S} = \mathcal{S}^+ \cup \mathcal{S}^-$, where \mathcal{S}^+ denotes the major sizes (e.g. $\{S, M, L\}$) and \mathcal{S}^- the minor sizes (e.g. $\{XS, XL\}$). Sale opportunities for each size $s \in \mathcal{S}$ are assumed to be independent across sizes and follow a Poisson process with rate λ_s and cumulative counting measure $\{N_s(t), t \geq 0\}$, where t denotes the time elapsed since the last replenishment (i.e., $N_s(t)$ is the random number of sale opportunities for size s that occurred between 0 and t). While this Poisson assumption is primarily made for tractability reasons (the derivation to follow may be considerably harder, or even impossible, with other demand process assumptions), we also believe that it is a priori realistic, even though we were not able to obtain data enabling a validation study. In addition, as mentioned above, although there may be in practice some demand dependencies across sizes (e.g. a customer preferring size XS may still go for size S if XS is not available, customers may choose the wrong size), we ignore these effects here.

Let $q_s \in \mathbb{N}$ represents the inventory level of size s immediately after replenishment at time 0 (that is the sum of any leftover inventory unsold in the previous period and the quantity contained in the new shipment), we can now define a *virtual stockout time* $\tau_s(q_s)$ for every size $s \in \mathcal{S}$ as:

$$\tau_s(q_s) \triangleq \inf\{t \geq 0 : N_s(t) = q_s\}.$$

In words, $\tau_s(q_s)$ is the time at which, starting from an initial inventory of q_s units, the store would

run out of size s , assuming that all inventory of that size remains always exposed to customers and that no subsequent replenishment ever occurs (these provisions justify the adjective “virtual”). The earliest time at which one of the major sizes runs out, assuming no replenishment occurs, can then be expressed naturally as

$$\tau_{\mathcal{S}^+}(\mathbf{q}) \triangleq \min_{s \in \mathcal{S}^+} \tau_s(q_s).$$

In the following we will omit the dependence on the variables $\mathbf{q} = (q_s)_{s \in \mathcal{S}}$ when no ambiguity arises, and use the notation $a \wedge b \triangleq \min(a, b)$.

As described in §3.1.1, all inventory is removed from customer view as soon as one of the major sizes runs out at any point between successive replenishments. Under that policy, the (random) total number of sales in a replenishment period resulting from an initial profile \mathbf{q} of inventory across sizes can be expressed as

$$G(\mathbf{q}) \triangleq \sum_{s \in \mathcal{S}^+} N_s(\tau_{\mathcal{S}^+} \wedge T) + \sum_{s \in \mathcal{S}^-} N_s(\tau_{\mathcal{S}^+ \cup \{s\}} \wedge T), \quad (1)$$

where $T > 0$ denotes the time between consecutive replenishments (this would be one week for Zara). Given our purposes, we are particularly interested in characterizing the expected value of the random sales function just defined, or $g_{\boldsymbol{\lambda}}(\mathbf{q}) \triangleq \mathbb{E}[G(\mathbf{q})]$, where the subscript $\boldsymbol{\lambda} \triangleq (\lambda_s)_{s \in \mathcal{S}}$ highlights the dependence of that function on the demand rate parameters characterizing the cumulative sales process $(N_s)_{s \in \mathcal{S}}$ (we will omit that subscript when it is obvious from context however). In the following, we establish some of its properties and develop an approximation for $g_{\boldsymbol{\lambda}}$ that may easily be embedded in a mixed integer program (MIP).

3.1.3 Model Analysis and Approximation. Intuitively, the descriptive model just defined captures how sales should increase when more inventory is available for display in a store. Our expected sales function g should thus obviously be non-decreasing with the inventory vector \mathbf{q} . A slightly less straightforward requirement is that function g should also reflect the decreasing marginal returns associated with shipping more inventory to a store, which follow from demand saturation. This feature is particularly important given our ultimate goal, as it will effectively dictate the relative values of marginal returns associated with sending a unit of inventory to different stores, depending on how much inventory is already present in these stores. Finally, the expected sales function should also capture the complementarity effects across sizes following from the display inventory removal policy described in §3.1.1. Specifically, the marginal returns associated with shipping inventory of one size should be non-decreasing with the inventory level of the other sizes (the major sizes in particular), since all sales processes terminate as soon as a major size runs

out.⁵ The expected sales function $g(\mathbf{q})$ associated with our model indeed exhibits those desirable qualitative features, as formally established by the following proposition (where the notation \mathbf{e}_s denotes a vector with all components equal to zero except the s -th, which is equal to one).

Proposition 1 *The expected sales function g is non-decreasing and discretely concave in each variable, and supermodular. That is, $g(\mathbf{q})$ is non-decreasing in x_s and its marginal differences $\Delta_s g(\mathbf{q}) \triangleq g(\mathbf{q} + \mathbf{e}_s) - g(\mathbf{q})$ are non-increasing in q_s and non-decreasing in $q_{s'}$ for all $\mathbf{q} \in \mathbb{N}^{\mathcal{S}}$ and $s, s' \in \mathcal{S}$ with $s \neq s'$.*

We turn next to the approximation. The first step is to note that each compensated Poisson process $\tilde{N}_s(t) \triangleq N_s(t) - \lambda_s t$ defines a martingale with $\tilde{N}_s(0) = 0$, and that the random variables $\tau_{\mathcal{S}^+} \wedge T$ and $\tau_{\mathcal{S}^+ \cup \{s\}} \wedge T$ appearing in (1) are bounded stopping times. Doob's optional sampling theorem thus applies, and

$$g\lambda(\mathbf{q}) = \lambda_{\mathcal{S}^+} \mathbb{E}[\tau_{\mathcal{S}^+} \wedge T] + \sum_{s \in \mathcal{S}^-} \lambda_s \mathbb{E}[\tau_{\mathcal{S}^+ \cup \{s\}} \wedge T], \quad \text{where } \lambda_{\mathcal{S}^+} \triangleq \sum_{s \in \mathcal{S}^+} \lambda_s. \quad (2)$$

Next, it follows from Jensen's inequality that for any subset of sizes $\mathcal{D} \subset \mathcal{S}$,

$$\mathbb{E}[\tau_{\mathcal{D}} \wedge T] \leq \min_{s \in \mathcal{D}} \mathbb{E}[\tau_s \wedge T]. \quad (3)$$

In turn, the minimum operand in (3) can be calculated as

$$\mathbb{E}[\tau_s \wedge T] = \frac{1}{\lambda_s} \sum_{k=1}^{q_s} \mathbb{P}(N_s(T) \geq k) = \sum_{k=1}^{q_s} \frac{\gamma(k, \lambda_s T)}{\lambda_s \Gamma(k)}, \quad (4)$$

where $\Gamma(a) \triangleq (a-1)!$ is the Gamma function and $\gamma(a, b) \triangleq \int_0^b v^{a-1} e^{-v} dv$ is the lower incomplete Gamma function – this follows from the optional sampling theorem and properties of Poisson processes (see e.g. Johnson, Kotz and Kemp 1993). As is clear from (4), the expected stopping time $\mathbb{E}[\tau_s \wedge T]$ can be expressed as a sum of q_s decreasing positive terms, so that it is a discretely concave function of q_s . That function is thus equal to the lower envelope of its (discrete) tangents at every point, or

$$\mathbb{E}[\tau_s \wedge T] = \min_{i \in \mathbb{N}} \{a_i(\lambda_s)(q_s - i) + b_i(\lambda_s)\} \quad (5)$$

with $a_k(\lambda_s) \triangleq \frac{\gamma(k, \lambda_s T)}{\lambda_s \Gamma(k)}$, $b_i(\lambda_s) \triangleq \sum_{k=0}^{i-1} a_k(\lambda_s)$ for $i \geq 1$ and $b_0(\lambda_s) \triangleq 0$ (we define by extension $a_\infty(\lambda_s) \triangleq 0$ and $b_\infty(\lambda_s) \triangleq T$). Note that the parameter $a_k(\lambda_s)$ is equal to the average inter-arrival time weighted by the probability that the $(k+1)$ -th unit of size s will sell before the next replenishment. Our proposed approximation consists of only computing the minimum in equation (5) over a (small) finite subset $\mathcal{N} \subset \mathbb{N}$ instead of the entire set of non-negative integers \mathbb{N} . That

⁵This prediction is consistent with our assumption that any substitution effects across sizes driven by customer behavior are negligible relative to the complementary effects driven by the inventory display policy described. While relaxing that assumption seems an interesting research challenge, we do not explore this further here.

is, we approximate function $\mathbb{E}[\tau_s \wedge T]$ by the lower envelope of only a few of its discrete tangents, thus obtaining an upper bound for its exact value. While that approximation can conceptually be made arbitrarily close (by considering a very large number $|\mathcal{N}|$ of discrete tangents), in practice we have used small sets $\mathcal{N}(\lambda_s)$ defined as

$$\mathcal{N}(\lambda_s) \triangleq \{i \in \mathbb{N} \cup \{\infty\} : b_i(\lambda_s) \approx 0, 0.3T, 0.6T, 0.8T, 0.9T, T\}, \quad (6)$$

which are straightforward to compute numerically. That is, our approximation of $\mathbb{E}[\tau_s \wedge T]$ as a function of q_s with values in $[0, T]$ consists of the lower envelope of six tangents chosen so that their intercepts cover the range of fractions of T shown in (6). Notice that this range of values is slightly denser as it gets close to T , that is when market saturation effects start to appear. Finally, substituting the approximate expression thus obtained for $\mathbb{E}[\tau_s \wedge T]$ in (3) for the sets $\mathcal{D} = \mathcal{S}^+$ and $\mathcal{D} = \mathcal{S}^+ \cup \{s\}$, then substituting in turn the resulting expressions in (2) yields the following approximation \tilde{g}_λ for our original expected sales function g_λ :

$$\tilde{g}_\lambda(\mathbf{q}) = \lambda_{\mathcal{S}^+} \min_{\substack{s \in \mathcal{S}^+, \\ i \in \mathcal{N}(\lambda_s)}} \{a_i(\lambda_s)(q_s - i) + b_i(\lambda_s)\} + \sum_{s \in \mathcal{S}^-} \lambda_s \min_{\substack{s' \in \mathcal{S}^+ \cup \{s\}, \\ i \in \mathcal{N}(\lambda_{s'})}} \{a_i(\lambda_{s'})(q_{s'} - i) + b_i(\lambda_{s'})\}. \quad (7)$$

Note that \tilde{g}_λ can be expressed as a linear combination of minimums of linear functions of \mathbf{q} , and may thus easily be embedded in an MIP formulation having \mathbf{q} as its primary decision variables (as we proceed to do in the following section). In addition, each of the two successive approximation steps (3) and (6) results in an upper bound of the original value, so that $\tilde{g}_\lambda(\mathbf{q})$ is also an upper bound for $g_\lambda(\mathbf{q})$. Finally, it is easy to see that the approximating function $\tilde{g}_\lambda(\mathbf{q})$ still exhibits the desirable qualitative properties of the original function $g_\lambda(\mathbf{q})$ stated in Proposition 1.

3.2 Network Sales Optimization Model. As stated in §1, our main objective is to implement an optimization model for distributing a limited amount of warehouse inventory between all stores in our industrial partner's retail network over time, with the goal of maximizing total expected revenue. As shown in Figure 1(b), the primary input data of that model includes a demand forecast for every store as well as the inventory available in stores and in the warehouse. Note that, in practice, this problem has a dynamic component, because the shipment decisions in any given week impact future warehouse and store inventory, and therefore both the feasible set of shipments and the achievable sales in subsequent weeks. In addition, this problem may involve connections between different articles because of possible substitution or complementarity effects at the store level. For computational and other practical reasons such as data availability however, we have implemented a mixed integer programming (MIP) formulation which considers each article independently (we return to this issue in section §E of the Appendix), and only captures dynamic

effects in a heuristic manner:

$$(MIP) \quad \max \quad \sum_{j \in J} P_j z_j + K \left(\sum_{s \in \mathcal{S}} (W_s - \sum_{j \in J} x_{sj}) \right) \quad (8)$$

$$s.t. \quad \sum_{j \in J} x_{sj} \leq W_s \quad \forall s \in \mathcal{S} \quad (9)$$

$$z_j \leq \left(\sum_{s \in \mathcal{S}^+} \lambda_{sj} \right) y_j + \sum_{s \in \mathcal{S}^-} \lambda_{sj} v_{sj} \quad \forall j \in J \quad (10)$$

$$y_j \leq a_i(\lambda_{sj})(I_{sj} + x_{sj} - i) + b_i(\lambda_{sj}) \quad \forall j \in J, s \in \mathcal{S}^+, i \in \mathcal{N}(\lambda_{sj}) \quad (11)$$

$$v_{sj} \leq a_i(\lambda_{sj})(I_{sj} + x_{sj} - i) + b_i(\lambda_{sj}) \quad \forall j \in J, s \in \mathcal{S}^-, i \in \mathcal{N}(\lambda_{sj}) \quad (12)$$

$$v_{sj} \leq y_j \quad \forall j \in J, s \in \mathcal{S}^- \quad (13)$$

$$z_j, y_j \geq 0 \quad \forall j \in J; v_{sj} \geq 0 \quad \forall (s, j) \in \mathcal{S}^- \times J; x_{sj} \in \mathbb{N} \quad \forall (s, j) \in \mathcal{S} \times J \quad (14)$$

In the formulation just stated, the primary decision variables x_{sj} represent the shipment quantities of each size $s \in \mathcal{S}$ to each store $j \in J$ for the current replenishment period, which are constrained to be integer. These variables are subjected to the warehouse inventory constraint (9), which insures that the total shipment of a given size across all stores never exceeds the inventory W_s available in the warehouse for that size. The secondary decision variables z_j correspond to the approximate expected sales across all sizes in each store j for the current period under consideration. The relationship assumed between these expected sales and the shipment decisions $\mathbf{x}_j \triangleq (x_{sj})_{s \in \mathcal{S}}$, existing store inventory $\mathbf{I}_j \triangleq (I_{sj})_{s \in \mathcal{S}}$ and demand forecast $\boldsymbol{\lambda}_j \triangleq (\lambda_{sj})_{s \in \mathcal{S}}$ for each size s in each store j follows the approximate inventory-to-sales function derived in §3.1.3. Specifically, constraints (10)-(13) along with the maximization objective (8) ensure that in any optimal solution to (MIP) the variable z_j is equal to $\tilde{g}_{\boldsymbol{\lambda}_j}(\mathbf{x}_j + \mathbf{I}_j)$, where \tilde{g} is the approximate expected sales function defined in (7). The expression for \tilde{g} is given by the right-hand-side of constraint (10) and makes use of two auxiliary variables, namely y_j and v_{sj} , which are equal to $\min_{\substack{s \in \mathcal{S}^+, \\ i \in \mathcal{N}(\lambda_{sj})}} \{a_i(\lambda_{sj})(q_s - i) + b_i(\lambda_{sj})\}$ and $\min_{\substack{s' \in \mathcal{S}^+ \cup \{s\}, \\ i \in \mathcal{N}(\lambda_{s'j})}} \{a_i(\lambda_{s'j})(q_{s'} - i) + b_i(\lambda_{s'j})\}$ respectively in any optimal solution. The latter follows from constraints (11)-(13), and again from the fact that this is a maximization problem.

The objective (8) is the sum of the expected revenue in the current period and a heuristic approximation of expected future revenue in subsequent periods. Its first term thus features the unit selling price P_j of the article considered in each store j , which constitutes input data. Note that (as is typical in the retail industry) the selling price may vary across stores but is identical for all sizes sold in the same store. The second objective term provides a monetary evaluation of the total inventory remaining in the warehouse after the shipment decisions considered are executed, using an exogenous unit value K for that inventory. That value K can thus be interpreted as the unit opportunity cost of shipping inventory to a particular store and is meant as a control lever allowing the model user to affect its output: A high value of the warehouse inventory value

K relative to the store selling prices P_j results in “conservative” shipments, possibly appropriate shortly after a product introduction (when forecast uncertainty is high), or when the returns and transshipment costs associated with excessive inventory sent to low-selling stores may be particularly high. In contrast, a low relative value of K results in “aggressive” shipments, perhaps suitable when forecasts are deemed more reliable, and/or toward the end of the planned shelf life of an article.

Note that any value of K above the selling price P_j of a store will effectively prevent any shipments to that store in any solution to (MIP) , since any unit of store inventory has a sales probability that is no larger than one under our inventory-to-sales model (10)-(13). More generally, K serves as a cutoff related to the expected revenue associated with every marginal unit of inventory shipped to a store j , itself equal to the sales price P_j times the sales probability of that unit. For example, a value of $K = 0.8P_j$ will prevent the incremental shipment to store j of any additional unit with an estimated sales probability lower than 80%.

The warehouse inventory value K appearing in (8) thus effectively allows the recommended shipments to reflect some of the dynamic considerations discussed earlier, even though the model is otherwise myopic. Note that we do not provide here any systematic method for deciding what the appropriate value of K should be, leaving the determination of that control to the users’ judgement and experience with the model. In addition, it is clear that the second term in (8) is only a very crude approximation of the expected revenue-to-go function that would appear in the objective of any dynamic programming formulation of the same problem, as it does not reflect the existing distribution of inventory in stores, does not account for a possible unbalance of the warehouse inventory across sizes, etc. However, our warehouse inventory value approximation does constitute a simple implementation of an idea consistently described as fruitful in the literature when applied to comparable stochastic inventory distribution models. In particular, Federgruen and Zipkin (1984) found that decoupling the overall inventory distribution problem into a *withdrawal* decision (how much inventory in total should be shipped to the stores) and *allocation* decision (how should that inventory be assigned to individual stores), then solving the allocation problem in a myopic manner, led to a good approximation (see also Chapter 8 of Zipkin 2000). Even though the policies proposed by these authors for the withdrawal problem are obviously more explicit and elaborated than our proposed withdrawal solution, as described above our model formulation otherwise implements the approximation scheme just described fairly closely: In the optimal solution to (MIP) , the value of the overall quantity shipped $\sum_{s \in \mathcal{S}} \sum_{j \in \mathcal{J}} x_{sj}$ is determined by the choice of K , and the individual shipments x_{sj} also solve the myopic allocation problem obtained when the total withdrawal amount $\sum_{s \in \mathcal{S}} \sum_{j \in \mathcal{J}} x_{sj}$ is constrained to be equal to that value.

4. Pilot Implementation Study

We were fortunate to help a team at Zara implement the new inventory allocation process described in earlier sections, and test it as part of a small scale but real pilot experiment conducted during the 2006-2007 Spring-Summer season (see the Appendix and Correa 2007 for more details on that implementation and the software developed to support it). This pilot test had three objectives: (i) prove the feasibility of the New Process envisioned through an actual implementation; (ii) collect feedback from real users to refine model features and the interface of supporting software; and (iii) estimate the specific impact of the New Process on some key operational performance metrics. In this section, we focus on the latter. Specifically, we describe our methodology in §4.1, discussing in turn the experimental set-up (in §4.1.1), the operational performance metrics used (in §4.1.2) and our impact assessment method (in §4.1.3). Results are then discussed in §4.2, and the financial impact is summarized in §4.3.

4.1 Methodology.

4.1.1 Experimental Design. Because Zara is comprised of three sub-brands or sections (Women, Men, and Children) that are organizationally distinct, it was felt that the pilot study would be best organized within a single one of them. Fifteen articles of the Women’s section were thus selected by a senior Zara manager as a test bed for the New Process, with the goal of constructing as representative a sample as possible. Because articles at Zara are split between a *basic* group (standard garments produced in large quantities and sold during the whole season) and a *fashion* group (trendier items with short life-cycles produced in small batches), the relative proportions of these two groups in the sample selected were in particular representative of the entire article population.

Our experimental set-up leveraged the fact that Zara currently has only two major warehouses worldwide, the first in Arteixo (next to A Coruña, northwest of Spain) shipping directly to about 500 stores in Western Europe, and the second in Zaragoza (about halfway between Barcelona and Madrid) shipping directly to about the same number of stores located in the rest of the world.⁶ Specifically, the new inventory allocation process was only implemented in Arteixo at some point during the life-cycle of the fifteen articles mentioned above, while the Legacy Process was still used to distribute all articles (including these) in Zaragoza. As further discussed in §4.1.3, our estimation of impact associated with the New Process is based on a comparison between that sample group of articles and a carefully selected control group of paired articles, conducted with data from stores assigned to Arteixo. Because both the sample and the control groups of articles

⁶Because stores in Western Europe tend to be more established and sell more merchandise, the Arteixo warehouse currently ships roughly 75% more volume than the one in Zaragoza.

were distributed using the Legacy Process in Zaragoza, by leveraging in turn the data from stores assigned to that second warehouse we were also able to validate our estimation methodology. That is, we could quantify the likely experimental error associated with our estimation of the impact specifically attributable to the New Process, since any non-negligible impact estimate in Zaragoza was obviously only attributable to such error as opposed to the New Process.

For each article in the sample group, the warehouse team in Arteixo thus switched from the Legacy to the New Process at some point after July 2006, possibly after that article had already been offered in stores for one or several weeks, and kept using the New Process until the end of the pilot, which was set in November 2006 (December is not a representative month due to Christmas and the end of the season). An important feature of that implementation is that the recommended shipments computed by the optimization model were only presented as a suggestion to the warehouse team, which retained the ability to freely modify them. We were initially concerned that, because of this discretion, any positive results would not be easily attributed to the New Process. However, it turned out that very few modifications of the model output were actually performed after the first couple of model runs, which proved our apprehension to be unfounded. Unfortunately, we were not able to choose the exact week when the model would be used for the first time for each article. For that reason, in the end we were only able to collect more than three weeks of data associated with the New Process for ten articles out of the original fifteen, and thus decided to remove the other five from our analysis. However, among the remaining ten were four basic and six fashion articles, which corresponds to approximately the same proportions as in the overall assortment.

4.1.2 Operational Performance Metrics. We now present the framework we developed to measure the operational performance of Zara’s inventory distribution over time, and applied in particular to evaluate the impact of our proposed process change. That framework is in essence the same one that underlies the classical newsvendor problem, in that it captures the goal of neither shipping too much nor too little inventory with respect to actual demand. Specifically, the two primary metrics we used, to be described shortly, respectively measure any overstock (i.e. any amount of excess inventory shipped to the stores) and understock (i.e. missed sales). In contrast with the newsvendor model however, these metrics have been designed to assess the distribution of a large number of articles across a network of many selling locations.

The primary data available to us included sales (V_{rsj}^d), shipments to store (X_{rsj}^d) and returns to warehouse/transhipments from store (R_{rsj}^d), all expressed in number of units, on each day d of the entire study period, for each available size s of each article r in a group of 118 (including the pilot articles described in §4.1.1), for every Zara store j in the world. Using basic inventory balance

equations we derived the corresponding daily inventory positions (I_{rsj}^d), and computed the corresponding weekly sales $Sales_{rsj}^w$, shipments $Shipments_{rsj}^w$ and returns/transhipments $Returns_{rsj}^w$ series, by summing up the daily data series over each day d in each (calendar) week w .⁷ Finally, we computed the corresponding network-wide cumulative weekly series $Sales_r^t$, $Shipments_r^t$ and $Returns_r^t$ for each article r , by summing up the previous series over all stores j in the network, sizes s of article r , and weeks w in the selling season up to and including week t .

In order to quantify missed sales, we constructed data series $Demand_{rsj}^w$ and $Demand_r^t$, defined over the same index set and providing estimates of uncensored customer demand, that is the sales that would have been observed had all merchandise been displayed without any stockout. Specifically, we first computed

$$DND_{rsj}^w \triangleq \sum_{d \in w} 1_{\{I_{rsj}^d=0\} \text{ or } \{ \min_{\tilde{s} \in \mathcal{S}_r^+} I_{r\tilde{s}j}^d=0 \text{ and } \max_{\tilde{s} \in \mathcal{S}_r^-} V_{r\tilde{s}j}^d=0 \}},$$

or number of Days in week w when size s of article r was Not on Display at store j , either because that size was out of stock, or because the article was removed due to the inventory display policy described in §3.1.1. Secondly, we estimated $Demand_{rsj}^w$ by increasing sales according to the number of days $7 - DND_{rsj}^w$ during which in the item was actually on display and when those sales were observed, according to the following procedure:

```

if  $Sales_{rsj}^w > 0$  and  $DND_{rsj}^w < 7$  then
  |  $Demand_{rsj}^w = Sales_{rsj}^w \left( \frac{7}{7 - DND_{rsj}^w} \right)$ 
else
  |  $Demand_{rsj}^w =$  most recent non-negative demand, otherwise zero.
end

```

Note that our estimation of demand implicitly assumes that average daily sales are identical throughout the week, whereas many Zara stores do experience some predictable variability within each week (e.g. surge of customer visits on Saturday). We also ignore the time of the day in which a stockout occurs, i.e, if an item sells out by noon, these (reduced) sales still count as a whole day. Although the resulting demand estimate could thus be biased, we do not believe that this bias is likely to affect the new inventory process and the old one in different ways (shipments occur on a weekly basis according to a constant schedule), and therefore feel that this simple approach is appropriate given our purposes.

⁷Point of sale data tends to be very accurate at Zara, but data inconsistencies concerning the inventory position could be more frequent. While we were not able to precisely estimate these inconsistencies during the pilot, we measured impact based on a differential relative to control articles (see §4.1.3 below) and there is no reason why any data inaccuracy concerning the inventory position would selectively affect more the pilot articles than the control ones.

We used the ratio of cumulative sales to cumulative shipments $S/S_r^t \triangleq Sales_r^t/Shipments_r^t$, or *shipment success ratio*, as our primary metric for quantifying any excess inventory (i.e. overstock) in Zara’s network. This represents the fraction of all units of a given article shipped to stores since the beginning of the season that have actually been sold to date.⁸ At Zara, the shipment success ratio S/S_r^t was actually used and closely monitored to evaluate operational performance long before our collaboration. For that reason, and also due to its convenience to estimate the impact on sales (see §4.1.3), we chose this metric to measure overstock instead of inventory turns (or its reciprocal, the average flow time), which might be more natural for other retailers.

The primary metric we used for quantifying missed sales due to lack of inventory (i.e. understock) is the ratio of cumulative sales to cumulative demand $S/D_r^t \triangleq Sales_r^t/Demand_r^t$, or *demand cover ratio*, where the cumulative weekly demand series $Demand_r^t$ is calculated analogously to $Sales_r^t$ and $Shipments_r^t$. This metric is to be interpreted as the proportion of demand that Zara was able to convert into sales through its display of inventory. In contrast with the first metric however, that second one was new to Zara. We argued when introducing it that both were required to form a comprehensive framework for evaluating distribution performance, as illustrated by Figure 3. Observe that while overstock and understock may not occur simultaneously in any inventory model describing the sales of a given product in a single location (as the newsvendor), in the network setting considered here both demand cover and shipment success ratios may be low at the same time, as explained in the lower left quadrant of that figure.

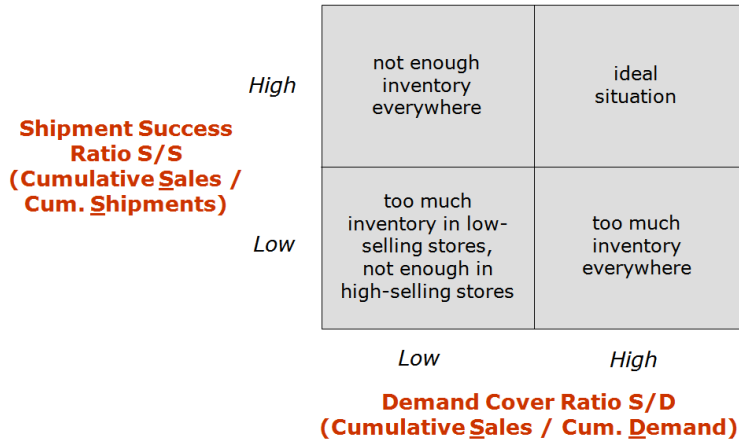


Figure 3: Proposed Evaluation Framework for Zara’s Distribution Performance.

Besides the two primary metrics just discussed, we used three additional secondary metrics. The first is the *Stock Retention ratio*, defined as $SR_r^t \triangleq 1 - Returns_r^t/Shipments_r^t$. This represents the fraction of units shipped to date that were not sent to another store or sent back to the warehouse

⁸The value of the S/S_r^t metric at the end of the season thus corresponds to the widely used metric of “sell-thru” or “retail turnover,” which is defined as the fraction of sales to total inventory received by suppliers.

by the store manager who received them originally, and therefore provides an alternative measure for overstock (although one that depends on the store manager’s actions).⁹ Our last two metrics are the *Store Cover ratio* (SC_r^t) and the *Display Cover ratio* (DC_r^t), formally defined as

$$SC_r^t \triangleq 1 - \frac{\sum_{w \leq t} \sum_{s \in \mathcal{S}_r} \sum_{j \in J} \sum_{d \in w} 1_{\{I_{rsj}^d = 0\}}}{7 \times t \times |\mathcal{S}_r| \times |J|} \quad \text{and} \quad DC_r^t \triangleq 1 - \frac{\sum_{w \leq t} \sum_{s \in \mathcal{S}_r} \sum_{j \in J} DND_{rsj}^w}{7 \times t \times |\mathcal{S}_r| \times |J|}.$$

The store cover ratio is the fraction of cumulative days \times sizes \times stores with stock at the store (possibly in the backroom), while the display cover ratio is the (smaller) fraction of these same days \times sizes \times stores with stock at the store *and* in sufficient quantity to be displayed to customers, according to the store inventory policy described in §3.1.1. They therefore both provide alternative measures for understock, although they are arguably coarser than the demand cover ratio. This is because SC_r^t and DC_r^t are inversely related to the number of days without stock, as opposed to their economic consequence (i.e. the number of units that could have been sold during those days). These last two metrics therefore do not differentiate between stockouts for the same period of time in high and low selling stores, in contrast with the demand cover ratio, but they give a sense of the service level.

Finally, note that values closer to one are more desirable for all the performance metrics just defined. In addition, the shipment success ratio S/S of a given article tends to improve over time in Zara’s environment, as weekly sales progressively deplete the inventory already shipped to stores and new shipments abate due to increasing warehouse inventory scarcity.¹⁰ The increasing inventory scarcity over a typical life cycle of a Zara article also explains the natural tendency for the demand cover ratio S/D (along with the other metrics SR , SC and DC) to decrease over time, as stockouts progressively occur earlier in the week following each replenishment, and become more widespread.

Several senior distribution managers at Zara interviewed independently emphasized to us another important aspect of the metric S/S , which is that, in their experience, improving the shipment success ratio of a given article becomes increasingly difficult for higher initial values of that metric.¹¹ While we did not conduct a specific study in order to investigate such observation, we notice that it is plausibly explained by the broken assortment effect described in Smith and Achabal (1998) and others (c.f. §2). Specifically, a higher initial value of the shipment success ratio is typically associated with lower inventory levels in stores, so that the broken assortment effect could make

⁹At Zara, store transshipments and returns to the warehouse also require the approval of the regional manager.

¹⁰From now on we omit the indices t and r when no ambiguity arises.

¹¹For example, increasing the shipment success ratio of an article from 0.8 to 0.9 is regarded within Zara as a considerably superior achievement by a distribution manager than an increase of that ratio for another article from 0.4 to 0.5 over the same period of time.

it harder to maintain or increase the S/S metric compared to a situation with higher levels of available inventory.

The prevalence within Zara of that nonlinear notion of performance relative to the metric S/S seemed potentially problematic given our purposes. This is because our impact assessment method (to be described in the next section) involved the comparison of differences in the value of the metrics defined over a given time period across distinct articles. As a result, we decided to also consider as an alternative metric the logarithmic transformation $-\ln(1 - S/S)$, which seemingly captured the nonlinear behavior more faithfully.¹² Likewise, limiting the decrease of the demand cover ratio and all other metrics from a given point in time was perceived by Zara managers to be considerably more challenging when the starting value of these ratios was closer to one. As a result, we also considered the transformations $\ln(M)$, where M is any of the other metrics S/D , SR , SC and DC (this alternative transformation was chosen because these metrics tend to decrease over time).

4.1.3 Impact Assessment Method. Estimating the impact of the new inventory process tested during the pilot experiment on the metrics defined in §4.1.2 presents a significant but classical methodological challenge. Specifically, the most relevant basis for comparison, that is the values that these metrics would have taken for the pilot articles over the same period of time had the new inventory process not been employed (i.e. the counterfactual), cannot be directly observed. Our solution is known as the difference-in-differences method, and is also used in many other empirical event studies found in the literature (e.g. Barber and Lyon 1996, Hendricks and Singhal 2005). It involves using instead a control group as a basis for comparison, where that group is designed by carefully matching individuals in the population receiving the treatment considered (in our case the articles included in the pilot experiment) with others in the population at large (the articles still distributed using the Legacy Process).

For each one of the ten pilot articles, we thus identify a control article among the 118 articles included in our dataset that were distributed using the Legacy Process over the same period of time. Our matching procedure is the following: (1) a basic (resp. fashion) pilot article may only be matched with a basic (resp. fashion) control article (see §4.1.1); (2) dates when a pilot article and its matched control were introduced cannot differ by more than one week; and (3) subject to those restrictions, the matched control article minimizes the initial difference in performance with the pilot article, as measured by the aggregate relative difference across shipment success and demand

¹²Continuing the previous example, changes of S/S from 0.4 to 0.5 and from 0.8 to 0.9 respectively correspond to increases of 0.2 and 0.7 for $-\ln(1 - S/S)$.

cover ratios

$$\frac{|S/S_r^t - S/S_{\tilde{r}}^t|}{\max\{S/S_r^t, S/S_{\tilde{r}}^t\}} + \frac{|S/D_r^t - S/D_{\tilde{r}}^t|}{\max\{S/D_r^t, S/D_{\tilde{r}}^t\}}, \quad (15)$$

where r is the pilot article considered, \tilde{r} its matched control article, and t is the week before the New Process was used for the first time to distribute article r . That is, the notion of proximity across articles that we use is driven by the values of our primary performance metrics immediately before the treatment begins (Barber and Lyon 1996 find that matching on such criteria leads to well-specified test statistics).

We carried out this matching procedure independently in the Arteixo and Zaragoza warehouses for the ten pilot articles (although the New Process was only used in Arteixo, see §4.1.1). For one article in Arteixo the control article initially selected was a clear outlier with an unusually bad performance in the second half of the season (this was confirmed by a standard box plot and Grubb’s test), resulting in an overly optimistic assessment of the New Process impact. We thus discarded that control article and repeated the matching procedure. The outcome is summarized in Table 2.

	Arteixo	Zaragoza
Number of pilot articles matched	10	10
Mean (median) shipment success ratio S/S of pilot articles	52.3%(46.8%)	50.0%(46.7%)
Mean (median) shipment success ratio S/S of control articles	51.8%(52.1%)	46.5%(39.6%)
Pearson (Spearman nonparametric) correlation coefficient	0.94*** (0.89***)	0.98*** (0.96***)
t-statistic (Wilcoxon signed-rank W-statistic) on the paired differences	0.19(9)	2.00 ^o (33)
Mean (median) demand cover ratio S/D of pilot articles	62.0%(63.4%)	61.1%(55.4%)
Mean (median) demand cover ratio S/D of control articles	53.4%(58.1%)	55.7%(53.8%)
Pearson (Spearman nonparametric) correlation coefficient	0.85*** (0.84***)	0.24(0.08)
t-statistic (Wilcoxon signed-rank W-statistic) on the paired differences	2.13*(39*)	0.95(21)

Note: the p values are two-tailed, except for the correlation coefficient, and the level of statistical significance from zero is noted by ^o $p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.025$, *** $p \leq 0.01$.

Table 2: Outcome of the Matching Process at Arteixo and Zaragoza.

Note that both performance ratios show significant correlation among pilot and control articles in Arteixo. For that warehouse the mean and median of the S/S ratios across articles in the pilot and control groups are not statistically different (p value > 0.1), but there is statistical evidence showing that the S/D ratios are larger for the pilot articles. In the case of Zaragoza, only the S/S ratios are significantly correlated. The means and medians of the S/D ratio and medians of the S/S ratios are not statistically different across pilot and control groups. While the mean of the S/S ratio is somewhat larger for the pilot articles, this is not quite significant (p value ≤ 0.1). Since the S/D ratios are uncorrelated, we also performed the unpaired tests and found that the mean and median were still not different. While such matching can never be perfect, we believe

ours to be suitable for our purposes and we are in particular unaware of any systematic bias that could make the final results be unduly optimistic. On the contrary, the fact that the initial values of the S/D ratios are larger in Arteixo is actually disadvantageous for the New Process relative to the Legacy Process, because this leaves less room for improvement to the pilot articles. During our collaboration with Zara we were able to confirm that making additional improvements to the ratios defined in §4.1.2 (or equivalently, impeding their deterioration) is much more challenging when the ratios are closer to one, as discussed in §4.1.2.

Our next step is to compute the difference-in-differences for each metric M defined in §4.1.2 and each matched pair of articles (r, \tilde{r}) in each warehouse as

$$\Delta(M) \triangleq (M_r^{End} - M_r^{Start}) - (M_{\tilde{r}}^{End} - M_{\tilde{r}}^{Start}), \quad (16)$$

where M_r^{Start} (resp. $M_{\tilde{r}}^{Start}$) is the value of the metric considered for the pilot (resp. control) article the week before the New Process was used for the first time, and M_r^{End} (resp. $M_{\tilde{r}}^{End}$) is the corresponding value at the end of the pilot experiment in November 2006. For data relative to the Arteixo warehouse, the expression in (16) thus provides an estimate for the specific impact of the New Process on the metric considered: the differences within parentheses excludes any time period other than that when the New Process was used, while the difference between the two pairs of parentheses is meant to exclude the effects of factors other than the New Process (such as seasonality, weather, exceptional events), based on the rationale that these external factors similarly affect pilot and control articles. Because the Legacy Process was used for both pilot and control articles in Zaragoza, expression (16), when calculated with data relative to that warehouse, provides an estimate of the error associated with our impact assessment methodology (see §4.1.1). Expression (16), when computed for the S/S and S/D metrics, can also be interpreted as a control-adjusted estimation of the increase in sales attributable to the New Process, relative to either shipments (S/S) or demand (S/D). Rearranging the terms defining $\Delta(S/S)$ for example yields

$$\Delta(S/S) = \left(\frac{Sales_r^{End} - \frac{Sales_r^{Start}}{Shipments_r^{Start}} \cdot Shipments_r^{End}}{Shipments_r^{End}} \right) - \left(\frac{Sales_{\tilde{r}}^{End} - \frac{Sales_{\tilde{r}}^{Start}}{Shipments_{\tilde{r}}^{Start}} \cdot Shipments_{\tilde{r}}^{End}}{Shipments_{\tilde{r}}^{End}} \right); \quad (17)$$

while the two terms in parenthesis in (17) respectively correspond to the pilot article and the control article as before, the numerator of each term represents the difference between the actual final cumulative sales and a proportional prediction of what these sales would have been with the Legacy Process, based on conditions immediately preceding the implementation of the new one. Because $Sales_r^t \leq \min \{ Shipments_r^t, Demand_r^t \}$, note that $\Delta(S/S)$ and $\Delta(S/D)$ can thus also be interpreted as a somewhat conservative estimate of the relative impact of the New Process on sales.

4.2 Results. The results of the live pilot test are summarized in Tables 3 and 4. Our observations are based on averages across articles of the values obtained for each metric using equation (16), which as discussed in §4.1.3 provides a control-adjusted estimation for the impact of the New Process on that metric. Note that considering averages across articles is justified by the need to factor out the randomness (noise) that we cannot control (indeed, the focus on such a statistic is prevalent in studies involving a pairwise matching procedure to construct a control group, e.g. Hendricks and Singhal 2005). In addition, we report associated t-statistics indicating whether these means are significantly different from zero, as well as the corresponding median for each metric and the respective nonparametric Wilcoxon signed-rank W-statistic (which likewise indicates whether the median is significantly different from zero). The significance of the statistics is reported conservatively by considering the two-tailed versions of the tests. Notice that, since our sample size is very small (only ten articles), a difference from zero has to be fairly large for it to be statistically significant.

The results shown in Table 3 should be qualified in light of the relatively low statistical significance of the impact on the primary metrics defined in §4.1.2, which is largely driven by the limitation of sample size imposed upon us. Nevertheless, Table 3 still suggests a positive impact on the primary metrics. These results are not driven by outliers since the mean and median changes all have the same sign. Considering all pilot articles, the changes in the value of the S/S and S/D ratios in Arteixo are 3.0% and 5.2% respectively, while the corresponding estimation errors given by measuring the same metrics in Zaragoza are -2.4% and 3.8% .¹³ The impact measured by the logarithmic transforms of these two metrics is even larger and different from zero with a high level of statistical significance, while the corresponding estimation errors obtained from the Zaragoza warehouse are not. This latter set of results is particularly noteworthy, as the logarithmic transforms of the S/S and S/D ratios constitute our most accurate representation of Zara’s managerial notion of performance.

Several interesting observations can be made by comparing the impact on each type of article in Arteixo (i.e. basic or fashion) with the respective estimation errors measured in Zaragoza. For basic articles, the mean impact on the S/D ratio is positive (10.1%) and larger than the corresponding estimation error (2.6%), whereas the mean impact on the S/S ratio is negative (-2.2%) and smaller (in absolute terms) than its associated error (-5.3%). In the case of fashion articles, the mean impact on the S/S ratio is positive (6.4%) and larger than the corresponding estimation error (-0.5%), whereas the mean impact on the S/D ratio, though still positive (1.9%), is smaller than its associated error (4.6%). These results suggest that the new allocation process

¹³Average flow time is an alternative measure of overstock, and we observed that its control-adjusted value decreased by almost one week at Arteixo, whereas at Zaragoza it remained practically the same.

impacts the two main types of articles in different ways. For basic articles its benefits would mostly stem from improvements in the demand cover S/D , whereas for fashion articles they would consist of improvements in the shipment success ratio S/S . These differences are plausibly explained by the forecasting model developed as part of this project (see §1) which, even though it is unbiased overall, tends to slightly underestimate the demand of fashion articles and overestimate that of basic articles.¹⁴ Indeed, with the new inventory distribution process forecast underestimation errors conceptually generate insufficient shipments favoring the S/S ratio to the detriment of the S/D ratio, whereas overestimation errors generate excessive shipments resulting in the opposite effect. Also supporting that interpretation is the observation that the correlation between the individual S/S and S/D ratios of each article is negative for both warehouses, but significantly more so in Arteixo (-0.75) where the new inventory distribution process was tested than in Zaragoza (-0.40) which only used the manual Legacy Process. Unfortunately, we were not able to further investigate this issue because the forecasts used during the pilot were not saved, and our attempt to reconstruct them a posteriori were unsuccessful (the orders placed by the store managers were not saved either).

Other reasons besides forecast biases may also explain the different impact of the model on the primary metrics for basic and fashion articles. In the case of the S/S ratio, the apparent poor performance of the model for basic articles has at least two alternative explanations: (i) the same two (out of four) basic pilot articles that had negative S/S performance in Arteixo also performed badly (in fact, worse) in Zaragoza, indicating that the choice of the basic pilot articles was particularly adverse; and (ii) the initial values of the S/S ratio for the basic articles in the pilot was relatively high (79.0% and 76.7% on average in Arteixo and Zaragoza, respectively), making it harder for the model to introduce significant improvements (see related discussion in §4.1.2). Consistent with the latter explanation, note that for basic articles in Arteixo the changes in mean and median of the log transform of the S/S ratio relative to the control articles (a metric designed to better reflect performance independently of the starting point of the pilot) are positive and significantly different from zero, whereas the corresponding estimation errors in Zaragoza are negative, and not significantly different from zero. In the case of the demand cover ratio S/D for fashion articles, the outcome of the matching process discussed in the previous section may provide an alternative explanation for why the impact measured in Arteixo is smaller than the estimation error. In Arteixo, the initial value of the S/D ratio is larger for the fashion pilot articles compared to the respective controls, whereas in Zaragoza it is contrariwise. As in the previous case, this seemingly negative result disappears when the logarithmic transform of that ratio, which we deem to be more meaningful, is considered instead.

¹⁴This likely stems from the different demand patterns and amount of input data available for basic and fashion articles - see Correa (2007) for more details on the forecasting model as well as its accuracy.

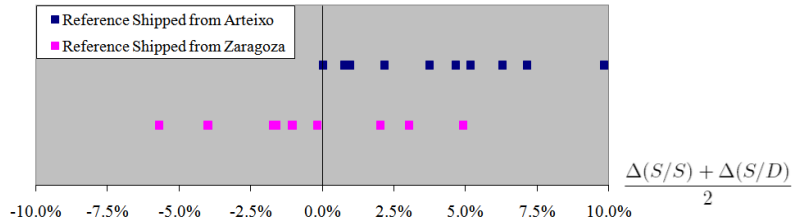


Figure 4: Estimated Relative Change in Sales for Each Pilot Article in Arteixo and Zaragoza.

We believe however that our most significant results stem from considering for each article the average of the control-adjusted impact on the S/S and S/D ratios, i.e. $\frac{\Delta(S/S) + \Delta(S/D)}{2}$. As noted in the discussion of equation (17), the control-adjusted impact estimations $\Delta(S/S)$ and $\Delta(S/D)$ each provide an estimation of the relative change in sales attributable to the model for each article in the pilot test, based on the assumption that the S/S and S/D ratios would have remained relatively unchanged over the test period under the counterfactual scenario. However, the strong negative correlation between $\Delta(S/S)$ and $\Delta(S/D)$ for each article, which is plausibly caused by forecast biases as discussed above, suggests that their average provide an estimate of the same quantity which is arguably more robust. Figure 4, which contains a plot of these averages for each pilot article distributed from Arteixo and Zaragoza, is striking in that respect: according to that measurement the relative sales impact in Arteixo is positive for every single article, with a mean across articles of 4.1% (median 4.2%), whereas the corresponding estimation error calculated using data from Zaragoza is centered around zero (mean and median across articles are 0.7% and -0.6% respectively).¹⁵ Subtracting this estimated experimental error suggests that the relative impact of the new distribution process on sales is of the order of 3 – 4%, which is substantial for a retailer like Zara given this improvement does not appear to require any major costs.

The results reported in Table 4 are similar to those discussed for the primary metrics. The impact on the stock retention ratio SR in Arteixo is larger than the estimation error, in particular for basic articles, suggesting that the model effectively reduces the level of transshipments. The measured impact of the New Process on the store cover SC and display cover DC ratios is also positive overall. However, this result is only driven by basic articles, since for fashion articles the impact on these ratios (and their log transforms) remains just under the estimation error. We note that basic articles have life-cycles that sometimes span the whole season, whereas fashion articles are by design only sold in stores for a few weeks, which may be why improving their store and display cover ratio seems more difficult. The fact that the overall impact is greater on the DC ratio than on the SC ratio is noteworthy however, as a distinguishing feature of our model is precisely

¹⁵The difference between Arteixo and Zaragoza is statistically significant with p values less than 0.05 and 0.065 for the mean and median respectively.

to capture the display of inventory on store shelves and racks (display cover), as opposed to its presence anywhere in the store including the backroom where it does not sell (store cover). As before, the results are not driven by outliers since the mean and median changes are thoroughly consistent; the results in Arteixo are significantly different from zero, whereas the estimation errors are not; and the statistical significance of the impact improves when the logarithmic transforms of the performance metrics are considered instead.

An interesting question is whether the results just described can be attributed to any consistent qualitative differences between the shipment decisions generated by the new optimization-based distribution process and those resulting from the manual Legacy Process. While we have not been able to collect sufficient data to perform a systematic study of this issue, the following are the most obvious insights that we obtained after discussing the solutions with the warehouse team.

- A typical first reaction by warehouse team members when inspecting the shipments recommended by the model was that these decisions seemed reasonable overall, and that in many cases they would have entered a quantity that differed by a couple of units at most.
- There were also situations when the solution seemed counterintuitive to them at first. One frequent source of discrepancy stemmed from the shipment decisions for different sizes of the same article. Given the data intensive decisions that the warehouse team had to make under time pressure, in some cases they would not fully account for the dependencies across sizes (see §3.1.1) and, in contrast with the model, would occasionally recommend shipments resulting in a size profile likely to be removed from display relatively quickly (if not immediately).
- We observed that when the remaining stock of an item was low, the manual practice typically consisted of trying to satisfy the requests of as many stores as possible. In contrast, the model tended then to avoid stock fragmentation by shipping to a smaller number of stores where the inventory would sell with greater probability and where full sets of sizes could be completed.
- Another source of discrepancy occurred when the shipment requests from a particular store seemed high in light of its inventory level and recent sales (perhaps that store anticipated some inventory rationing in subsequent weeks). The warehouse team would then typically cut down the shipment request but did not have a formal rule to decide by how much, and as a result tended to reduce the requests by a significantly lower amount than the model.

4.3 Financial Impact. We conclude this section with an assessment of the financial impact of the New Process.¹⁶ Following the pilot study, Zara made the decision to use our new allocation process for all its articles and stores. This large scale deployment effectively occurred over the first

¹⁶The financial impact estimations provided here were performed independently by the paper authors and did not engage the responsibility of the Inditex group.

half of 2007. On September of the same year, Zara's CFO estimated that the financial benefits specifically attributable to the use of our model were consistent with the sales increase reported here for the live pilot.

If we take 2006 as the base year, and assuming that the model increased revenues by 3.5% in 2007 (see §4.2), we thus estimate that the New Process generated about \$275M in additional revenue or \$34M in additional net income for 2007.¹⁷ ¹⁸ Given the finding from the pilot experiment that the model does not increase average inventory throughout the retail network (see §4.2), this additional net income corresponds to an increase of the return on assets by at least 3.5% (note that the New Process does not affect assets other than inventory).

The financial impact estimation just stated ignores any changes on operational costs that might have been introduced by the New Process. This seems conservative, since the software license and labor costs incurred during the implementation (see below) appear smaller than the reduction in warehouse returns and transshipments costs measured during the pilot study (§4.2), which the previous estimate based on additional revenues also ignores. We note as well that Zara's earnings before taxes and interest (EBIT) amounted to 16.5% of revenues in 2006. Ignoring again any impact on operational costs, a 3.5% revenue increase due to the New Process would imply a 21.3% increase in EBIT for 2007. The actual increase from 2006 to 2007 was 22.5%, which at least to some extent validates our estimations.

Finally, the main cost associated with this implementation stemmed from the development time spent by the project team members (in comparison, the software costs represented a minor expense, see §F in the Appendix). Although some team members had several simultaneous assignments, we estimate that the total time invested in the project is equivalent to one manager experienced in distribution and store management and two engineers experienced in databases and optimization software working together full-time during one year, which would represent less than \$1M in labor costs. Given the financial impact estimated above, this would correspond to a return on investment for the project after one year amounting to thirty times or more.

5. Conclusion

The work just presented involved the development of a new operational process to allocate scarce inventory across the store network of a fast-fashion retailer. The most salient feature of that process is arguably its reliance on an optimization model capturing inventory display policies at the store level. In addition, we also reported the implementation and test of that New Process as part of

¹⁷See §C of the Appendix for the exact calculation.

¹⁸The validity of extrapolating the specific sales increase measured during the pilot study to subsequent periods is subject to a number of assumptions, including the relative stability of Zara's products attractiveness to its customers.

a live pilot experiment, using a performance evaluation framework that may be of independent interest. The results of the live pilot test suggest that the New Process increases sales (by 3 – 4% according to our best estimate), decreases transshipments, and increases the proportion of time that the articles are on display. As of the time of writing, every item currently found in any Zara store worldwide has been shipped to that store based on the output of the optimization model described in Section §3.2 of this paper. In addition, the Inditex group is also planning to start using that New Process for its other brands in the near future.

Beyond its financial impact, this new allocation process has also had organizational implications which we believe are positive. In particular, the warehouse allocation team has seen its responsibilities shift from repetitive data entry towards exception handling, scenario analysis, process performance evaluation and improvement. That team deserves special recognition in our view, for it has played a pivotal role in the improvement and successful implementation of the new allocation model, and has demonstrated to us the importance of human experience when facing many distribution issues. To the best of our knowledge, Zara is not planning to leverage any productivity gains associated with the new allocation process through head count reductions, however we expect the New Process to generate substantial economies of scale if the company continues to grow as planned. In addition, store managers may be asked in the future to provide input to forecasts as opposed to shipments (see Correa 2007).

This project may also have had some cultural impact at Zara, a company which we believe owes part of its success to the unique intuition of its founder. We doubt that Zara will ever use advanced mathematical models to help with several of its key challenges including anticipating volatile market trends, recruiting top designers and creating fashionable clothes, and it is not clear to us that it should. In fact, we see the story of Zara’s success as a humbling one given our background, because a key aspect of its business model is to leverage the endogenous increase in demand associated with short product life-cycles, a feature not predicted or quantified by any of the current quantitative inventory purchasing models that we know of (Fisher, Raman and McClelland 2000). However, this collaborative interaction may have influenced Zara’s realization that for other processes involving large amounts of quantitative data, such as distribution and pricing, formal Operations Research models may lead to better performance and more scalable operations.

Beyond Zara, we expect that our model may also be useful to other retailers managing a network of stores, particularly those facing lost sales (as opposed to backorders) and dependencies across sizes introduced by store display policies. Indeed, the latter feature has not received much attention in the literature, and the present work suggests that accounting for it may have a significant impact on sales. In terms of future work, the methodology we applied (solving a large scale industrial op-

timization problem subject to uncertainty by embedding the linear approximation to a stochastic performance evaluation model in an MIP formulation) may be applicable to other contexts beyond retailing. Further related theoretical work could thus be interesting, for example characterizing the sub-optimality of our approximate MIP formulation, or the development of a unified framework for general allocation problems. Finally, we see other research opportunities motivated by the specific features of fast fashion retailers relative to traditional retailers. In particular, further investigations of store-level inventory display policies and warehouse ordering policies seem warranted.

Acknowledgments. We first thank our industrial partner Zara, and our two key collaborators José Antonio Ramos Calamonte and Juan Correa. We are also particularly grateful to other Zara employees, including Javier García, Miguel Díaz, and José Manuel Corredoira (Pepe). We also thank the Area Editor, the Associate Editor and two anonymous referees for many comments that helped improve the paper. We refer the reader to the online Appendix for acknowledgements of other contributing individuals and funding sources.

References

- Axsäter, S., J. Marklund and E. A. Silver. 2002. Heuristic Methods for Centralized Control of One-Warehouse, N-Retailer Inventory Systems. *MSOM* 4(1) 75-97.
- Barber, B. M. and J. D. Lyon. 1996. Detecting Abnormal Operating Performance: The empirical Power and Specification of Test Statistics. *J. Financial Economics* 41 359–399.
- Cachon, G. and M. Lariviere. 1999. Capacity Choice and Allocation: Strategic Behavior and Supply Chain Performance. *Mgmt. Sci.* 45(8) 1091-1108.
- Caro, F. and J. Gallien. 2007. Dynamic Assortment with Demand Learning for Seasonal Consumer Goods. *Mgmt. Sci.* 53(2) 276-292.
- Chen, F. 2005. Salesforce Incentives, Market Information, and Production/Inventory Planning. *Mgmt. Sci.* 51(1) 60-75.
- Correa, J. 2007. *Optimization of a Fast-Response Distribution Network*. M.S. Thesis. LFM, MIT.
- Eppen, G. and L. Schrage. 1981. Centralized Ordering Policies in a Multi-Warehouse System with Leadtimes and Random Demand. L. Schwarz, ed. *Multi-Level Production/Inventory Control Systems: Theory and Practice*. North Holland, Amsterdam, The Netherlands, 51-69.
- Federgruen, A. and P. Zipkin. 1984. Approximations of Dynamic Multilocation Production and Inventory Problems. *Mgmt. Sci.* 30 69-84.
- Federgruen, A. 1993. Centralized Planning Models for Multi-Echelon Inventory Systems under Uncertainty. S. C. Graves et al., eds. *Handbooks in OR & MS Vol. 4*, North-Holland, Amsterdam, The Netherlands, 133-173.

- Fisher, M. L., A. Raman, and A. S. McClelland. 2000. Rocket Science Retailing Is Almost Here - Are You Ready. *Harvard Business Review*. July-August 2000, 115-124.
- Fraiman, N., M. Singh L. Arrington and C. Paris. 2002. Zara. Columbia Business School Case.
- Graves, S. C. 1996. A Multiechelon Inventory Model with Fixed Replenishment Intervals. *Mgmt. Sci.* **42**(1) 1-18.
- Ghemawat, P. and J. L. Nueno. 2003. ZARA: Fast Fashion. Harvard Business School Multimedia Case 9-703-416.
- Hendricks, K. B. and V. R. Singhal. 2005. Association Between Supply Chain Glitches and Operating Performance. *Mgmt. Sci.* **51**(5) 695-711.
- Jackson, P. L. 1988. Stock Allocation in a Two-Echelon Distribution System or What to Do Until Your Ship Comes In. *Mgmt. Sci.* **34**(7) 880-895.
- Johnson, N., S. Kotz and A. Kemp, *Univariate Discrete Distributions*, Wiley Interscience, New York, NY, 1993.
- McGavin, E. J., L. B. Schwarz and J. E. Ward. 1993. Two-Interval Inventory-Allocation Policies in a One-Warehouse, N-Identical-Retailer Distribution System. *Mgmt. Sci.* **39**(9) 1092-1107.
- Kalyanam, K., S. Borle and P. Boatwright. 2005. Modeling Key Item Effects. Working Paper. Tepper School of Business. Carnegie Mellon University.
- Muckstadt, J.A. and R.O. Roundy. 1993. Analysis of Multistage Production Systems. S. C. Graves et al., eds. *Handbooks in OR & MS Vol. 4*, North-Holland, Amsterdam, The Netherlands, 59-131.
- Smith, S. A. and D. D. Achabal. 1998. Clearance Pricing and Inventory Policies for Retail Chains. *Mgmt. Sci.* **44**(3) 285-300.
- Zhang, S. and G. J. Fitzsimons. 1999. Choice-Process Satisfaction: The Influence of Attribute Alignability and Option Limitation. *Organ. Behav. Hum. Dec.* **77**(3) 192-214.

	Original Metrics		Logarithmic Transforms	
	$\Delta(S/S)$	$\Delta(S/D)$	$\Delta(-\ln(1-S/S))$	$\Delta(\ln(S/D))$
Arteixo				
Mean (median) impact on basic articles	-2.2%(-1.8%)	10.1%(8.6%)	15.0%(12.7%)	19.1%(18.8%)
Mean (median) impact on fashion articles	6.4%(7.4%)	1.9%(2.0%)	18.6%(23.4%)	15.5%(9.4%)
Mean (median) impact on all articles	3.0%(0.6%)	5.2%(7.9%)	17.1%(13.3%)	16.9%(17.9%)
t-statistic (W-statistic) on the model's impact	1.07(17)	1.82(35 $^\circ$)	3.17** (43*)	3.31*** (47**)
Zaragoza				
Mean (median) error for basic articles	-5.3%(-5.0%)	2.6%(1.9%)	-23.3%(-21.5%)	8.6%(8.0%)
Mean (median) error for fashion articles	-0.5%(-0.3%)	4.6%(5.0%)	-5.2%(-0.2%)	8.4%(17.1%)
Mean (median) error for all articles	-2.4%(-1.3%)	3.8%(2.8%)	-12.5%(-0.2%)	8.5%(12.0%)
t-statistic (W-statistic) on the error	0.76(-13)	1.55(25)	0.77(-11)	1.54(27)

Note: the p values are two-tailed, except for the correlation coefficient, and the level of statistical significance from zero is noted by $^\circ p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.025$, *** $p \leq 0.01$.

Table 3: Final Results of the Live Pilot Test for the Shipment Success (S/S) and Demand Cover (S/D) Ratios.

	Original Metrics			Logarithmic Transforms		
	$\Delta(SR)$	$\Delta(SC)$	$\Delta(DC)$	$\Delta(\ln(SR))$	$\Delta(\ln(SC))$	$\Delta(\ln(DC))$
Arteixo						
Mean (median) impact on basic articles	0.8%(0.8%)	6.1%(6.4%)	7.5%(7.6%)	0.8%(0.8%)	7.3%(8.1%)	9.4%(9.8%)
Mean (median) impact on fashion articles	0.7%(0.0%)	1.1%(0.6%)	2.1%(1.8%)	0.7%(0.0%)	1.5%(0.8%)	3.2%(2.5%)
Mean (median) impact on all articles	0.7%(0.2%)	3.1%(3.7%)	4.3%(5.0%)	0.8%(0.2%)	3.8%(4.1%)	5.7%(6.1%)
t-statistic (W-statistic) on the model's impact	2.17 $^\circ$ (31 $^\circ$)	2.19 $^\circ$ (37 $^\circ$)	2.59*(41*)	2.19 $^\circ$ (31 $^\circ$)	2.38*(39*)	2.82** (43*)
Zaragoza						
Mean (median) error for basic articles	0.0%(0.0%)	3.1%(2.7%)	3.5%(3.2%)	0.0%(0.0%)	3.8%(3.3%)	4.4%(4.0%)
Mean (median) error for fashion articles	-0.5%(-0.4%)	1.7%(1.7%)	2.5%(2.9%)	-0.5%(-0.4%)	2.0%(2.0%)	3.3%(3.6%)
Mean (median) error for all articles	-0.3%(-0.2%)	2.3%(1.8%)	2.9%(2.9%)	-0.3%(-0.2%)	2.7%(2.1%)	3.7%(3.6%)
t-statistic (W-statistic) on the error	1.24(-21)	1.65(27)	1.64(29)	1.17(-21)	1.72(27)	1.76(31)

Note: the p values are two-tailed, except for the correlation coefficient, and the level of statistical significance from zero is noted by $^\circ p \leq 0.1$, * $p \leq 0.05$, ** $p \leq 0.025$, *** $p \leq 0.01$.

Table 4: Final Results of the Live Pilot Test for the Store Retention (SR), Store Cover (SC), and Display Cover (DC) Ratios.

Appendix to “Inventory Management of a Fast-Fashion Retail Network”

Felipe Caro, UCLA Anderson School of Management, fcaro@anderson.ucla.edu

J eremie Gallien, MIT Sloan School of Management, jgallien@mit.edu

A. Acknowledgements

We would first and foremost like to thank our industrial partner Zara for providing an exciting collaboration opportunity between industry and academia, and for partly funding this project. In particular, we are most grateful to Jos e Antonio Ramos Calamonte for his uplifting friendship, tireless work, unflinching support and persuasion skills. This project owes much more to him than can be described here. A second key contributor is Juan Correa, who deserves most credit for the computer implementation of the forecasting system and optimization model, but also helped in many other ways. Other Zara employees we are particularly grateful to include Javier Garc ıa for his intuition, high energy and tough questions, Miguel D ıaz for his trust, vision and wisdom and Jos e Manuel Corredoira (Pepe) for his hard work developing the software user and data interfaces. Joaqu ın Lorenzo, Jes us Gonz alez, Juan Quintela, Mar ıa Vent ın also spent considerable time sharing their invaluable knowledge of Zara’s operations with us, and we are also grateful to Ram on Fern andez, Marcos Montes, and Francisco Babio. Don Rosenfield, Jonathan Griffith and the MIT Leaders For Manufacturing Program provided logistical help, while Serguei Netessine and Marcelo Olivares suggested useful references. We are also grateful to Steve Graves, and the participants of the Operations Management Seminar at the MIT Sloan School of Management, the DOTM Colloquium at the UCLA Anderson School of Management and other research seminars held at Stanford University, Columbia University, Northwestern University, the University of Chicago, the Wharton School at UPenn, UC Irvine, and the University of Chile for helpful feedback and discussions. We also thank the Area Editor, the Associate Editor and two anonymous referees for many comments that helped improve the paper. This work was partly funded by the Singapore-MIT Alliance and the J. Spencer Standish (1945) career development chair of the MIT Sloan School of Management.

B. Validation of the Store Inventory Display Policy

In order to validate the store inventory display policy, we collected a data set describing sales (V_{rsj}^d) and inventory shipments (X_{rsj}^d) for all stores (indexed by j) and all sizes (indexed by s) of a group of 118 articles (indexed by r) of the Women’s 2006-2007 Spring-Summer season, on every day (indexed by d) between early July and late November 2006. Note that we thus excluded the last two months of that selling season, being advised that the clearance sales period occurring then

gave rise to very distinct store execution patterns. From this data and the knowledge of the initial inventory positions at the beginning of the season (zero), we constructed, using basic inventory balance equations, the data series I_{rsj}^d of estimated store inventory positions at the end of each day during the period covered. This in turn enabled us to compute for each store j the statistics DPA_j (standing for “number of Days when the inventory display Policy was Applicable”) and DPF_j (“number of Days when the Policy was actually Followed”) defined as:

$$\left\{ \begin{array}{l} DPA_j \triangleq \sum_r \sum_d 1_{\{\min_{s \in \mathcal{S}_r^+} I_{rsj}^d = 0 \text{ and } \max_{s \in \mathcal{S}_r} I_{rsj}^d > 0 \text{ and } \max_{s \in \mathcal{S}_r^+} V_{rsj}^d = 0\}} \\ DPF_j \triangleq \sum_r \sum_d 1_{\{\min_{s \in \mathcal{S}_r^+} I_{rsj}^d = 0 \text{ and } \max_{s \in \mathcal{S}_r} I_{rsj}^d > 0 \text{ and } \max_{s \in \mathcal{S}_r} V_{rsj}^d = 0\}} \end{array} \right. ,$$

where 1_E is the indicator function associated with event E , \mathcal{S}_r is the set of size in which article r is available, and $\mathcal{S}_r^+ \subset \mathcal{S}_r$ is the subset of major sizes for that article (estimated by a Zara executive with store management experience). In words, DPA_j is the number of days, summed over all articles, when there was a stockout of a major size but there was still some inventory available in another size, and DPF_j corresponds to the subset of those days characterized by the additional requirement that no sales were observed for any size then.¹⁹ In absence of available data for whether in-store inventory is located in the display area or in the backroom, we measure the adherence by store j to the inventory display policy described in §3.1.1 by the ratio DPF_j/DPA_j ; our results are summarized by Figure 2 in the main paper, which shows the distribution of those ratios found across Zara’s entire network of approximately 900 stores (in 2006).

Note that the metric we used may be overestimating adherence somewhat, as it ignores the days when the policy was not applied if no sales were observed then. Another issue is that I_{rsj}^d represents the inventory position, as opposed to the inventory on hand at the store (the shipment lead-time ranges from 1 to 3 days across stores) – this may lead to both an underestimation of (major size) stockouts and an overestimation of inventory (of other sizes), and could thus bias our adherence ratio in either direction. Nevertheless, with less than 2% of the stores having an adherence lower than 80% and average and median across stores both equal to 89% according to that metric, we still find these results to be quite compelling.

C. Calculation of the Financial Impact

The computations leading to the financial impact reported in §4.3 of the main are described in the spreadsheet represented by Table 5 given below. Note a few important remarks that follow.

¹⁹The seemingly redundant additional condition $\max_{s \in \mathcal{S}_r^+} V_{rsj}^d = 0$ in the definition of DPA_j is meant to reduce the impact of inventory record inaccuracies whereby $V_{rsj}^d > 0$ even though $I_{rsj}^d = 0$. While this type of inaccuracy seemed infrequent, we did not change the definition of DPA_j since the correction just mentioned did not require additional effort.

		(A)	(B)
Zara/Inditex Financial Data			
	Fiscal Year	2007	2006
(1)	Inditex - Net Margin (Net Income to Sales Ratio)	13.3%	12.2%
(2)	Zara - Net Sales (M. Euros)	6,264	5,534
(3)	Zara - EBIT (M. Euros)	1,116	911
(4)	Zara - EBIT as percentage of Sales	17.8%	16.5%
Assumptions			
(5)	Exchange Rate (USD per Euro)	1.42	
(6)	Sales Increase with respect to Pre-Model (%), see §4.2	3.5%	
Model Financial Impact		Formula	Value
(7)	Model Impact on Sales (M. US Dollars)	B2*A6*A5	275
(8)	Model Impact on Income (M. US Dollars)	B7*B1	34
(9)	Estimated increase in EBIT	A6/B4	21.3%
(10)	Actual increase in EBIT	A3/B3 - 1	22.5%

Table 5: Estimation of the Model’s Financial Impact.

- The financial data in Table 5 was extracted from Inditex’s Consolidated Results for Fiscal Year 2007.²⁰
- Net sales are equal to revenues net of any consumption tax and converted to euros at the average exchange rates for the fiscal year.
- Because Zara’s net margin is not available to us, we use Inditex’s net margin instead, which we believe is a good approximation since Zara has the same EBIT as percentage of sales as Inditex (Zara represents more than 65% of Inditex’s sales).
- In order to remain conservative, we have computed the impact on sales and income using 2006 Pre-Model data. Otherwise, using figures from 2007, the impact on sales and income becomes \$311M and \$41M respectively.

D. Proof of Proposition 1

The fact that $g(\mathbf{q})$ is a non-decreasing function of each variable q_s follows directly from its definition (this is actually true for each sample path of the associated random function G defined in (1)). We will prove that the other properties stated hold for all functions of the form $h^{\mathcal{A}}(\mathbf{q}) \triangleq \mathbb{E}[\tau_{\mathcal{A}} \wedge T]$ for $\mathcal{A} \subset \mathcal{S}$ and $\tau_{\mathcal{A}} \triangleq \min_{s \in \mathcal{A}} \tau_s(q_s)$. The result will then follow because these properties are preserved by positive linear combinations, and $g = \lambda_{\mathcal{S}^+} h^{\mathcal{S}^+} + \sum_{s \in \mathcal{S}^-} \lambda_s h^{\mathcal{S}^+ \cup \{s\}}$. The proof to follow is adapted from that of Proposition 1 in Lu and Song (2003).

Define \mathbf{e}_s to be a vector with all components equal to zero except the s -th equal to one, and

²⁰Downloaded from www.inditex.com on April 14, 2008.

define $\Delta_s h^A(\mathbf{q}) \triangleq h^A(\mathbf{q} + \mathbf{e}_s) - h^A(\mathbf{q})$. Note first that

$$\begin{aligned} \Delta_s h^A(\mathbf{q}) &= \int_0^T [\mathbb{P}(\tau_s(q_s + 1) > t) - \mathbb{P}(\tau_s(q_s) > t)] \prod_{s' \in \mathcal{A} \setminus \{s\}} \mathbb{P}(\tau_{s'}(q_{s'}) > t) dt \\ &= \int_0^T \mathbb{P}(N_s(t) = q_s) \prod_{s' \in \mathcal{A} \setminus \{s\}} \mathbb{P}(\tau_{s'}(q_{s'}) > t) dt \\ &= \mathbb{P}(\tau_s(q_s) \leq \tau_{\mathcal{A} \setminus \{s\}} \wedge T). \end{aligned}$$

Because $\tau_s(q_s)$ is increasing in q_s on every sample path, this last expression is decreasing in q_s and increasing in $q_{s'}$ for $s \neq s'$, proving in particular that h^A is discretely concave in q_s .

Observe now that on every sample path the function $\tau_{\mathcal{A}} \wedge T = \min_{s \in \mathcal{A}} \tau_s(q_s) \wedge T$ is the minimum of increasing functions in each single variable and is therefore supermodular, implying that h^A is also supermodular (example 2.6.2 (f) and Corollary 2.6.2 in Topkis 1998).

E. Model Extension for the Multicolor Case

We now discuss the case of garments sold in multiple colors but which are otherwise identical, and show how our model may be extended accordingly. We emphasize however that while we have been able to solve the resulting optimization models in less than a couple minutes for several realistic data sets, the work to be described next has not yet been implemented in the field.

Multicolor articles are particularly significant for Zara, as all the colors available of these articles (for example a T-shirt or a sweater) are typically displayed together in a coordinated manner in a central location of each store, and thus account for a relatively high fraction of sales. In addition, because of the special customer appeal of these displays with assorted colors, Zara uses a specific store inventory display policy for these multicolor articles which is different from that described in §3.1.1:

- In addition to the distinction between major and minor sizes mentioned earlier, for multicolor articles store managers also distinguish between *key* colors that are particularly popular (there are always at least two such designated colors), and the other *normal* colors;
- Each article with a key color is managed as if it were displayed on its own, independently of the inventory remaining in the other colors and as described in §3.1.1. For example, if a key color article comes in sizes $\{S, M, L\}$ with M as the major size, it will remain on display as long as there is at least one unit left in size M for that color;
- Normal color articles will remain on display, independently of which of their sizes may be missing, as long as there are at least two key colors remaining on display. However, whenever the display has only one or no key color left, then all normal colors are managed again as if they were independent (as described in §3.1.1), and not part of a coordinated multicolor display.

The policy just described essentially relaxes the inventory removal rules applied to individual articles, with the goal of maintaining an assortment of as many colors as possible, and thus the attractiveness of the display, for a longer period of time. Specifically, the normal colors remain displayed longer than they would otherwise, because their presence is thought to enhance the overall display appeal, thereby contributing to the sales of the other colors. Multicolor articles still face the trade-off between sales and brand impact and labor requirements discussed in §3.1.1 however, so that key colors are not protected from early display removal in the same way that normal colors are. This is because whenever critical sizes are missing for key colors, a comparatively higher number of customers will solicit store associates or become frustrated, which (in contrast to normal colors) more than offsets their positive contribution to the overall display appeal and sales of other colors.

To extend our previous models to the case of these multicolor articles, define \mathcal{C} as the set of all available colors (e.g. {blue,black,white,red,orange,fuschia}), partitioned into a set of key colors \mathcal{C}^+ (e.g. {blue,black,white}) and a set of normal colors $\mathcal{C}^- \triangleq \mathcal{C} \setminus \mathcal{C}^+$. Considering first the case of a single store, if $q_s^c \in \mathbb{N}$ represents the inventory level of size $s \in \mathcal{S}$ in color $c \in \mathcal{C}$ immediately after replenishment, and N_s^c is the cumulative counting process of corresponding sales opportunities (with rate λ_s^c), we can define as before

$$\begin{cases} \tau_s^c(q_s^c) \triangleq \inf\{t \geq 0 : N_s^c(t) = q_s^c\} \\ \tau_{\mathcal{S}^+}^c(\mathbf{q}^c) \triangleq \min_{s \in \mathcal{S}^+} \tau_s^c(q_s^c) \end{cases}$$

as the virtual stockout time of size s in color c and the virtual removal time of color c , respectively. The last time at which at least two key colors are on display can then be expressed as

$$\tau_{\mathcal{C}^+}(\mathbf{q}) \triangleq \max_{c \in \mathcal{C}^+}^2 \tau_{\mathcal{S}^+}^c(\mathbf{q}^c),$$

where \max^2 denotes the operator returning the second highest value of a given set of numbers. According to the inventory display policy described above, the expected sales for all sizes of each color c during a replenishment period of length T starting with initial inventory vector \mathbf{q} are finally

$$g_{\lambda}^c(\mathbf{q}) = \begin{cases} \lambda_{\mathcal{S}^+}^c \mathbb{E}[\tau_{\mathcal{S}^+}^c \wedge T] + \sum_{s \in \mathcal{S}^-} \lambda_s^c \mathbb{E}[\tau_{\mathcal{S}^+ \cup \{s\}}^c \wedge T] & \text{if } c \in \mathcal{C}^+ \\ \sum_{s \in \mathcal{S}} \lambda_s^c \mathbb{E}[(\tau_{\mathcal{C}^+} \vee \tau_{\mathcal{S}^+}^c) \wedge \tau_s^c \wedge T] & \text{if } c \in \mathcal{C}^- \end{cases}, \quad (18)$$

where $\lambda_{\mathcal{S}^+}^c \triangleq \sum_{s \in \mathcal{S}^+} \lambda_s^c$ and $a \vee b \triangleq \max(a, b)$. While the case of a key color $c \in \mathcal{C}^+$ shown in (18) is the same (notation aside) as in (2) so that the analysis and approximation described in §3.1.3 readily apply, the case of a normal color $c \in \mathcal{C}^-$ slightly differs. We propose the approximation

$$\begin{aligned} \mathbb{E}[(\tau_{\mathcal{C}^+} \vee \tau_{\mathcal{S}^+}^c) \wedge \tau_s^c \wedge T] &\leq \mathbb{E}[(\tau_{\mathcal{C}^+} \vee \tau_{\mathcal{S}^+}^c) \wedge T] \wedge \mathbb{E}[\tau_s^c \wedge T] \\ &\approx (\mathbb{E}[\tau_{\mathcal{C}^+} \wedge T] \vee \mathbb{E}[\tau_{\mathcal{S}^+}^c \wedge T]) \wedge \mathbb{E}[\tau_s^c \wedge T]. \end{aligned} \quad (19)$$

Note that the first step above involves (as shown) an overestimation, while the second step involves an underestimation. Finally, while the second and third expectations in the r.h.s. of (19) can be

approximated using the same techniques as described in §3.1.3, we propose to approximate the first expectation as

$$\mathbb{E}[\tau_{\mathcal{C}^+} \wedge T] \approx \max_{c \in \mathcal{C}^+} \mathbb{E}[\tau_{\mathcal{S}^+}^c \wedge T], \quad (20)$$

then approximate the operand in (20) as before.

Turning finally to the problem of allocating inventory between all stores in our partner's distribution network, we can extend our previous optimization formulation (*MIP*) to the multicolor case by applying the above analysis and approximations as follows:

(*MIP* – *MC*)

$$\max \sum_{j \in J} P_j z_j + K \left(\sum_{c \in \mathcal{C}} \sum_{s \in \mathcal{S}} (W_s^c - \sum_{j \in J} x_{sj}^c) \right) \quad (21)$$

$$\text{s.t.} \quad \sum_{j \in J} x_{sj}^c \leq W_s^c \quad \forall c \in \mathcal{C}, s \in \mathcal{S} \quad (22)$$

$$z_j \leq \sum_{c \in \mathcal{C}^+} \left(\left(\sum_{s \in \mathcal{S}^+} \lambda_{sj}^c \right) y_j^c + \sum_{s \in \mathcal{S}^-} \lambda_{sj}^c v_{sj}^c \right) + \sum_{c \in \mathcal{C}^-} \sum_{s \in \mathcal{S}} \lambda_{sj}^c u_{sj}^c \quad \forall j \in J \quad (23)$$

$$y_j^c \leq a_i(\lambda_{sj}^c)(I_{sj}^c + x_{sj}^c - i) + b_i(\lambda_{sj}^c) \quad \forall c \in \mathcal{C}, j \in J, s \in \mathcal{S}^+, i \in \mathcal{N}(\lambda_{sj}^c) \quad (24)$$

$$v_{sj}^c \leq a_i(\lambda_{sj}^c)(I_{sj}^c + x_{sj}^c - i) + b_i(\lambda_{sj}^c) \quad \forall c \in \mathcal{C}^+, j \in J, s \in \mathcal{S}^-, i \in \mathcal{N}(\lambda_{sj}^c) \quad (25)$$

$$v_{sj}^c \leq y_j^c \quad \forall c \in \mathcal{C}^+, j \in J, s \in \mathcal{S}^- \quad (26)$$

$$u_{sj}^c \leq a_i(\lambda_{sj}^c)(I_{sj}^c + x_{sj}^c - i) + b_i(\lambda_{sj}^c) \quad \forall c \in \mathcal{C}^-, j \in J, s \in \mathcal{S}, i \in \mathcal{N}(\lambda_{sj}^c) \quad (27)$$

$$u_{sj}^c \leq r_j^c \quad \forall c \in \mathcal{C}^-, j \in J, s \in \mathcal{S} \quad (28)$$

$$r_j^c \leq y_j^c + M(1 - k_j^c) \quad \forall c \in \mathcal{C}^-, j \in J \quad (29)$$

$$r_j^c \leq \ell_j + M k_j^c \quad \forall c \in \mathcal{C}^-, j \in J \quad (30)$$

$$\ell_j \leq y_j^c + M h_j^c \quad \forall c \in \mathcal{C}^+, j \in J \quad (31)$$

$$\sum_{c \in \mathcal{C}^+} h_j^c = |\mathcal{C}^+| - 2 \quad \forall j \in J \quad (32)$$

$$y_j^c \geq 0 \quad \forall (c, j) \in \mathcal{C}^+ \times J; \quad (33)$$

$$v_{sj}^c \geq 0 \quad \forall (c, s, j) \in \mathcal{C}^+ \times \mathcal{S}^- \times J; u_{sj}^c \geq 0 \quad \forall (c, s, j) \in \mathcal{C}^- \times \mathcal{S} \times J; \quad (34)$$

$$x_{sj}^c \in \mathbb{N} \quad \forall (c, s, j) \in \mathcal{C} \times \mathcal{S} \times J; \quad (35)$$

$$k_j^c \in \{0, 1\} \quad \forall (c, j) \in \mathcal{C}^- \times J; h_j^c \in \{0, 1\} \quad \forall (c, j) \in \mathcal{C}^+ \times J \quad (36)$$

where each variable without an explicit domain is assumed to be a real number. In the formulation above, variable z_j is equal in any optimal solution, because of the maximization objective and constraint (23), to the approximation discussed above for the sum over all colors and sizes of the

expected sales in store j , $\sum_{c \in \mathcal{C}} g_{\lambda_j}^c(\mathbf{I}_j + \mathbf{x}_j)$ where g is defined in (18). Specifically, constraints (27)-(28) ensure that any optimal value of variable u_{sj}^c is equal to the minimum of r_j^c and our piecewise linear approximation for $\mathbb{E}[\tau_s^c \wedge T]$ in store j , constraints (29)-(30) and (36) likewise ensure that $r_j^c = y_j^c \vee \ell_j$, constraint (24) ensures that y_j^c is equal to our approximation for $\mathbb{E}[\tau_{S^+}^c \wedge T]$ in store j , finally constraints (31)-(32) and (36) ensure that $\ell_j = \max_{c' \in \mathcal{C}^+} y_j^{c'}$.

F. Forecast Development and Software Implementation

We only provide here a brief overview of forecast development and software implementation issues, and refer the reader to Correa (2007) for a more exhaustive discussion. As illustrated in Figure 1 (b), the new inventory allocation process starts with the calculation of demand forecasts. The primary input to the forecasting model developed includes data on past sales of each article in each store, together with the most recent shipment requests placed by store managers and their current store inventory levels. As an output, it provides an estimation of the expected sales the upcoming week for each article and size in each store (denoted by λ_{sj} in §3). This forecast is then used to calculate the parameters $a_i(\lambda_{sj})$, $b_i(\lambda_{sj})$ and $c_i(\lambda_{sj})$ characterizing the inventory-to-sales function analyzed in §3.1.3, which in turn constitute input data to the MIP described in §3.2.

This MIP was implemented in an application developed with the AMPL modeling language. It relies on direct links with Zara’s relational databases (AS400 and SQL-Server) from which it reads the relevant input parameters (store and warehouse inventory, demand forecasts, selling prices) and to which it writes the shipment recommendations generated. The optimization problem itself is solved with the optimization engine CPLEX 10.0, with a typical running time of just a few seconds to achieve full or near optimality. A graphical interface was developed so a user with no prior knowledge of modeling languages (as is the case of most members of Zara’s warehouse allocation team) could easily interact with that application, and in particular specify some of the control parameters required by the MIP (such as the set of major sizes S^+ or the valuation of units left at the warehouse K) and perform corresponding what-if scenario analysis before finalizing shipments. Some additional features were added to the application in order to make the warehouse allocation team more comfortable with the model (see Correa 2007 for details). The interface was coded using Visual Foxpro 9 and Visual Basic.

In the months preceding the live pilot we performed dry runs on a daily basis using real data. The solutions were discussed with the warehouse team, who pointed out anything that seemed “unreasonable.” Their input was used to fine tune the model, and in particular, to calibrate the K parameter. In fact, based on the interactions with the warehouse team, a default (or preset) value for the K parameter was identified, and that value was used during the entire live pilot. Currently, in most of the runs the warehouse team uses the preset value of K . The few exceptions occur when

they want to deplete the inventory in the warehouse, in which case they lower the value of K so that the model ships all the remaining stock.

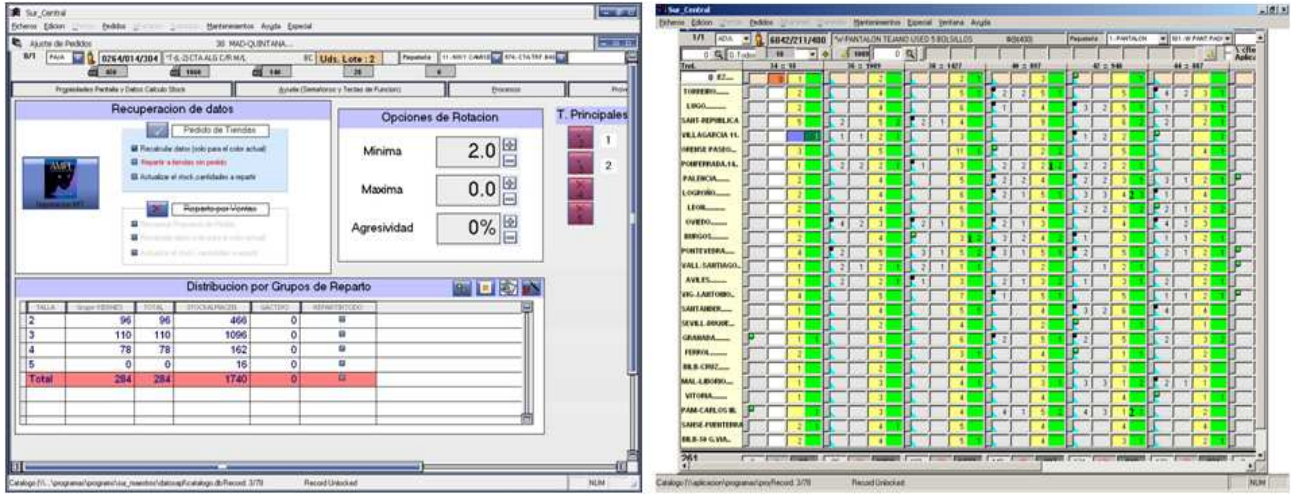


Figure 5: Screenshots of the Software Implementation.

The screenshot shown in the left of Figure 5 illustrates the part of the interface dedicated to optimization run control parameters. In particular, it displays (in the upper left) the field “Agresividad” corresponding to the warehouse unit valuation K (fittingly referred to as an “aggressiveness factor”), as well as the key sizes “T. Principales”. The bottom area displays some results summarizing an optimization run performed for what-if analysis purposes. The screenshot on the right of Figure 5 illustrates the part of the interface used to represent and potentially modify the detailed solution, i.e. each recommended shipment for each article and size to each store. Every such screen corresponds to an article, each row corresponds to a store, each group of columns refers to a size in which that article is offered, and columns in each such group contain data on the corresponding sales in the previous week, inventory currently in store, quantity requested by the store manager for the next shipment and finally the intended shipment. It should be noted that this latter screen is part of the existing application that was already used by the warehouse team before the beginning of our interaction with Zara in order to manually enter all shipment quantities, and visualize the information which they thought was most relevant to those decisions, as part of the process discussed in §1. The net impact of the new allocation process as seen by the warehouse team members through that interface was only to see default suggested values for the shipment quantities to be implemented (the output of the optimization model) in the exact location where they previously had to enter that information manually from scratch. They did retain however the ability to freely modify these suggested shipments (see discussion at the end of §4.1.1). In retrospect, we believe that the use of a pre-existing and familiar interface in order to display the model output did substantially contribute to the success of that implementation.

Additional References (not cited in the main paper)

Topkis, D. M. 1998. *Supermodularity and Complementarity*. Princeton University Press, Princeton, NJ.