Inventory Management of a Fast-Fashion Retail Network

Felipe Caro * Jérémie Gallien †

August 2, 2007

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 a reference is removed from display whenever one of its *key sizes* stocks out. We first formulate and analyze a stochastic model predicting the sales of a reference 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. Secondly, we formulate a mixed-integer program embedding a piece-wise linear approximation of the first model applied to every store in the network and allowing 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 field experiment performed to assess the impact of that model relative to the prior procedure used to determine weekly shipment quantities. The results of that experiment suggest that the new allocation process tested increases sales, reduces transhipments, and increases the proportion of time that an important category of Zara's products spends on display.

1. Introduction

The recent impressive financial performance of the spanish group Inditex (its 2005 income-to-sales ratio of 12% was among the highest in the retail industry) shows the promise of the *fast-fashion* model adopted by its flagship brand Zara, but also other retailers that 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 from 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 references produced in smaller series, continuously changing the assortment of products displayed in their stores (Ghemawat and Nueno 2003 report that Zara offers on average $11,000$ references in a given season, compared to only $2,000 - 4,000$ items for key competitors) in order to increase their appeal to customers (a top Zara executive quoted in

*UCLA Anderson School of Management, Los Angeles, CA 90095, fcaro@anderson.ucla.edu
†MIT Sloan School of Management, Cambridge, MA 02142, jgallien@mit.edu

Fraiman et al. 2002 states that Zara customers in Spain make on average 17 store visits per year). In addition, products offered by fast fashion retailers during the selling season may result from design changes decided after the season has started as a response to actual sales information, 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 to be described here, which has been conducted in collaboration with Zara, addresses in particular the 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, store inventory display policies) do 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 reference to be included in the weekly shipment from the warehouse to each store (see Figure 1 (a) for an illustration).

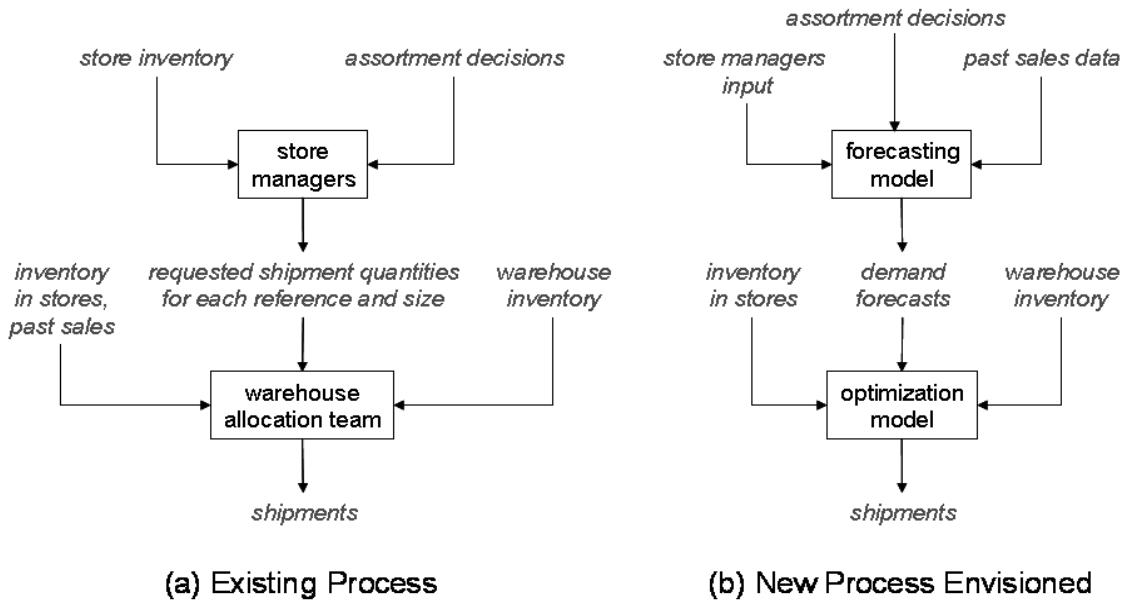

Figure 1: Existing Process and New Process Envisioned to Determine Weekly Shipments to Stores.

According to that process, each store manager would receive a weekly statement of the subset of references 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 assortment decision made by Zara's headquarters for that particular store. It would not mention

however the total quantity of inventory available in the warehouse for each reference 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 references. A team of employees at the warehouse would then reconcile all those requests by modifying (typically lowering) these shipment quantities so that the overall quantity shipped for each reference 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) may justify a more scalable process. A first issue centered on the incentives of store managers, who were primarily rewarded for the total sales achieved in their stores. We believe that as a consequence store managers would frequently request quantities exceeding their true needs, particularly when suspecting that the warehouse may not hold enough inventory of a top-selling reference to satisfy all stores (among others, Cachon and Lariviere 1999 study a stock rationing model capturing this behavior). Another issue was that store managers are responsible for a large set of responsibilities beyond determining shipment quantities, including building, sustaining and managing a team of several dozen sales associates in environments with high employee turnover. Finally, we also believe that the very large amount of data that the warehouse allocation team was responsible for reviewing (shipments of several hundred references offered in several sizes to more than a thousand stores) made it challenging to balance inventory allocations across stores and references in a consistent way, let alone one that would globally maximize sales. Motivated by these observations, we started discussing with Zara the alternative process for determining these weekly shipment quantities that is illustrated in Figure 1 (b). The new process envisioned consists of using some input from store managers along with past historical sales to build demand forecasts, then use these forecasts, the inventory of each reference 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 remainder of this paper discusses the work we have performed in order to develop and test the optimization model supporting the new inventory allocation process just described – we do not further discuss the associated forecasting model here, details of which may be found in Correa (2007). After a discussion of the relevant literature in §2, we discuss in §3 the successive steps we followed to develop that optimization model, specifically the analysis of a stochastic model predicting the weekly sales to be expected from a single store with given starting inventory profile and merchandise display policy (in §3.1) and the formulation of an optimization model for

the distribution of a single reference over the entire network embedding an approximation of the stochastic model just described (in §3.2). Section 4 discusses a field experiment we have conducted with Zara in order to assess the likely impact of a potential large-scale implementation of our proposed inventory allocation process. Finally, we offer concluding remarks in §5. The Appendix contains a technical proof (§5.1), an extension of the model just mentioned to the case of references offered in multiple colors (§5.2), and some material related to the software implementation of this work (§5.3).

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), Ferdows et al. (2003), McAfee et al. (2004) 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, which is motivated by the frequent assortment changes seen in fast-fashion outlets. In the present paper, the product assortment of each store is considered an exogenous input to the inventory allocation problem. This is justified by Zara's current operations, whereby inventory shipments are subordinated to assortment decisions in a hierarchical manner (see Figure 1). In that sense, the present paper constitutes a logical continuation to Caro and Gallien (2007).

The generic problem of allocating inventory from a central warehouse to several locations satisfying separate demand streams has received much attention in the literature. Remarkably, the optimal allotment of limited stock over time in common models of such a distribution system is still an open question. When demand is assumed to be deterministic however, there are very effective heuristics with data-independent worst case 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, inventory policies described in the literature are based on approximate analysis, and bounds on their performance, when they are available, depend on problem data. Focusing on stochastic periodic-review models (store inventory replenishment occurs on a fixed weekly schedule at Zara), Table 1 summarizes the main features of representative existing studies along with that of the present one – for a more exhaustive description of this body of literature, see the recent paper by Axsäter, Marklund and Silver (2002) or the earlier survey by Federgruen (1993).

	decision scope			time horizon		shortage model		retailers	
	ordering	*withdrawal*	*allocation*	*finite*	*infinite*	*backorder*	*lost sales*	*identical*	*non-identical*
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.

A first differentiating feature in Table 1 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 entire network of retailers; and *allocation* to the split of any inventory withdrawn from the warehouse between individual retailers. Note that some papers (e.g. Eppen and Schrage 1981, Federgruen and Zipkin 1984) assume that the warehouse does not hold inventory, in which case only the ordering and allocation decisions are relevant.[1] Other differentiating features include the time horizon (finite or infinite), shortage model (backlog or lost sales) and retailer types (identical or not) considered.

We observe that the operational strategy of fast-fashion retailers consists of offering through the selling season a large number of different references, 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, at Zara a single manufacturing order for each reference is typically placed with suppliers (or its internal manufacturing group), 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 Zara uses in fact separate organizational processes for purchasing from its suppliers and distributing warehouse inventory to its stores. Consequently, we have chosen to not consider here 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 the input by the user of a valuation associated with warehouse inventory, and that we do not provide any rigorous methodology for determining the value of that parameter (see §3.2 for more details and discussion). We also point out that Zara

[1] In such models, the warehouse may represent a cross-docking facility or central ordering function.

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.

But 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 fast-fashion retail industry, explicitly captures some dependencies across different sizes and colors of the same reference. Specifically, in Zara stores (and we believe many other fast-fashion retail stores) a stock-out of some selected *key* sizes or colors of a given reference triggers the removal from display of the entire set of sizes or colors. While we refer the reader to §3.1.1 and §5.2 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 (e.g. customers having selected an article that is not available in their size are more likely to solicit a store associate and/or leave the store in frustration). 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 every reference on display will have a minimum number of units exposed (the number of key sizes). In that sense, the existing studies on marketing inventory are also relevant (see Smith and Achabal 1998 and references therein for the description and discussion of demand models whereby the sales rate decreases when the inventory displayed goes under a threshold).

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 to 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 field implementation experiment for an inventory allocation model in a large distribution network (see §4). We also believe that the simple performance evaluation framework we developed when designing that experiment 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 relationship between the inventory of a specific reference 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 reference they would like to buy, but cannot find their fitting 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 is thus particularly critical to the brand perception of fast-fashion retailers such as Zara, who offer a large number of references produced in small series throughout the season: customers would quickly assimilate a store containing many references with missing sizes to nothing more than a thrift shop.

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 reference, store associates move all of the remaining inventory of that reference from the display shelf/rack to the backroom, thus effectively removing the entire reference from customers' sight. In contrast, no such action is taken when the store runs out of one of the minor sizes. 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 reference, and

do not expect references 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 reference 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 reference were made in size XS in the first place (not all references 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 in fact 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, and constituted rather an empirical observation of common store behavior. Because this seemed a key modeling issue, we decided to verify through analysis how prevalent that policy was. Specifically, we collected a data set describing sales (V^d_{rsj}) and inventory shipments (X^d_{rsj}) for all stores (indexed by j) and all sizes (indexed by s) of a group of 118 references (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 occuring 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^d_{rsj} 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:

$$\begin{cases} DPA_j \triangleq \sum_r \sum_d 1_{\{\min\limits_{s\in\mathcal{S}_r^+} I^d_{rsj}=0 \text{ and } \forall s\in\mathcal{S}_r^+ \ I^d_{rsj}\geq V^d_{rsj} \text{ and } \max\limits_{s\in\mathcal{S}_r} I^d_{rsj}>0\}} \\ DPF_j \triangleq \sum_r \sum_d 1_{\{\min\limits_{s\in\mathcal{S}_r^+} I^d_{rsj}=0 \text{ and } \max\limits_{s\in\mathcal{S}_r} I^d_{rsj}>0 \text{ and } \max\limits_{s\in\mathcal{S}_r} V^d_{rsj}=0\}} \end{cases},$$

where 1_E is the indicator function associated with event E, $\mathcal{S}_r$ is the set of size in which reference

r is available, and $\mathcal{S}_r^+ \subset \mathcal{S}_r$ is the subset of major sizes for that reference (estimated by a Zara executive with store management experience). In words, DPA_j is the number of days, summed over all references, 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. In absence of available data for whether in-store inventory is located in the display area or in the backroom, we propose to measure the adherence by store j to the inventory display policy described earlier by the ratio DPF_j/DPA_j; our results are summarized by 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).

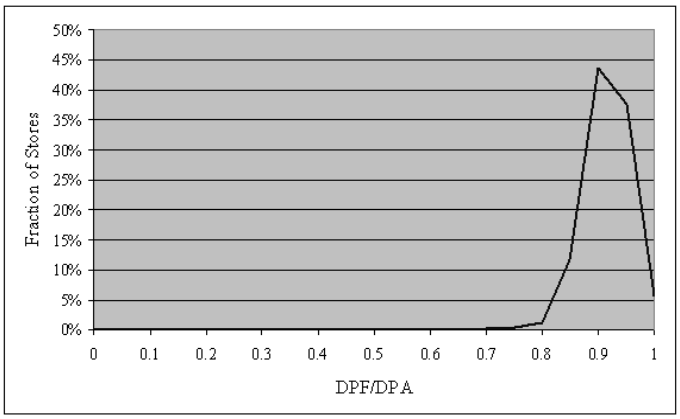

Figure 2: Histogram of the Adherence Ratios DPF_j/DPA_j Across All Stores.

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 striking. In particular, they justify in our view that the inventory display policy based on major sizes be used as a representation of store execution behavior for modeling purposes.

We describe next a stochastic model developed to answer the following question: Given the dependency between inventory and sales of different sizes introduced by the store inventory management policy based on major sizes described above, how many sales of each reference 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 references. 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 references. In section §5.2 of the Appendix however, we discuss how our model may be extended to the case of products offered in multiple colors (although we still assume that demand streams across colors are independent, so that the dependencies considered only stem from store execution policies).

3.1.2 Model Description.

Consider a reference 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 occured between 0 and t). 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

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

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 every 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. 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 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

$$g_{\boldsymbol{\lambda}}(\mathbf{q}) = \lambda_{\mathcal{S}^+} \mathbb{E}[\tau_{\mathcal{S}^+} \wedge T] + \sum_{s \in \mathcal{S}^-} \lambda_s \mathbb{E}[\tau_{\mathcal{S}^+ \cup \{s\}} \wedge T], \tag{2}$$

where $\lambda_{\mathcal{S}^+} \triangleq \sum_{s \in \mathcal{S}^+} \lambda_s$ (see e.g. Karatzas and Shreve 1991).

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

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

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

$$\begin{aligned} \mathbb{E}[\tau_s \wedge T] &= \frac{1}{\lambda_s} \sum_{k=1}^{q_s} \mathbb{P}(N_s(T) \geq k) & (4) \\ &= \sum_{k=1}^{q_s} \frac{\gamma(k, \lambda_s T)}{\lambda_s \Gamma(k)}, & (5) \end{aligned}$$

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. As is clear from (4) and (5), 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

$$\mathbb{E}[\tau_s \wedge T] = \min_{i \in \mathbb{N}} \{a_i(\lambda_s)(q_s - i) + b_i(\lambda_s)\} \tag{6}$$

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 (6) over a (small) finite subset $\mathcal{N} \subset \mathbb{N}$ instead of the entire set of non-negative integers $\mathbb{N}$. That 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

$$\mathcal{N}(\lambda_s) \triangleq \{i \in \mathbb{N} \cup \{\infty\} : b_i(\lambda_s) \approx 0, 0.3T, 0.6T, 0.8T, 0.9T, T\} \tag{7}$$

and thus straightforward to compute numerically. 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 $\tilde{g}_{\boldsymbol{\lambda}}$ for our original expected sales function $g_{\boldsymbol{\lambda}}$:

$$\tilde{g}_{\boldsymbol{\lambda}}(\mathbf{q}) = \lambda_{\mathcal{S}^+} \min_{s\in\mathcal{S}^+} \min_{i\in\mathcal{N}(\lambda_s)} \{a_i(\lambda_s)(q_s - i) + b_i(\lambda_s)\} + \sum_{s\in\mathcal{S}^-} \lambda_s \min_{s'\in\mathcal{S}^+\cup\{s\}} \min_{i\in\mathcal{N}(\lambda_{s'})} \{a_i(\lambda_{s'})(q_{s'} - i) + b_i(\lambda_{s'})\}. \tag{8}$$

Note that $\tilde{g}_{\boldsymbol{\lambda}}$ 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 (7) results in an upper bound of the original value, so that $\tilde{g}_{\boldsymbol{\lambda}}(\mathbf{q})$ is also an upper bound for $g_{\boldsymbol{\lambda}}(\mathbf{q})$. Finally, it is easy to see that the approximating function $\tilde{g}_{\boldsymbol{\lambda}}(\mathbf{q})$ still exhibits the desirable qualitative properties of the original function $g_{\boldsymbol{\lambda}}(\mathbf{q})$ stated in Proposition 1.

3.2 Network Sales Optimization Model.

As stated in §1, our main purpose here is to develop an operational 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. That 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 (decision variables) and sales (rewards) in subsequent weeks. Ignoring for now any dependences across distinct references (we return to this issue in §5.2), a possible dynamic programming (DP) formulation of this problem involving the characterization of a profit-to-go function V_t (Bertsekas 2005) would be

$$(DP) \qquad V_t(\mathbf{W}^t, \mathbf{I}^t) = \max_{\mathbf{x}\in\mathbb{N}^{\mathcal{S}\times J}} \quad \left(\sum_{j\in J} P_j g_{\boldsymbol{\lambda}_j}(\mathbf{x}_j + \mathbf{I}_j^t) + \mathbb{E}_{\boldsymbol{\omega}}[V_{t+1}(\mathbf{W}^{t+1}, \mathbf{I}^{t+1})] \right) \tag{9}$$

$$s.t. \quad \sum_{j\in J} x_{sj} \leq W_s^t \quad \forall s \in \mathcal{S} \tag{10}$$

$$W_s^{t+1} = W_s^t - \sum_{j\in J} x_{sj} \quad \forall s \in \mathcal{S} \tag{11}$$

$$I_{sj}^{t+1} = I_{sj}^t + x_{sj} - \omega_{sj} \quad \forall (s,j) \in \mathcal{S}\times J, \tag{12}$$

where the decision variables are noted $\mathbf{x} \triangleq (\mathbf{x}_j)_{j\in J} \triangleq (x_{sj})_{(s,j)\in\mathcal{S}\times J}$ and defined as the vector of shipment quantities of each size s to each store j to be determined for the current replenishment period t. The state in the DP formulation just stated includes $\mathbf{W}^t \triangleq \left(W_s^t\right)_{s\in\mathcal{S}}$ the vector of inventory at the warehouse in each size s available for shipments in the current period t, and $\mathbf{I}^t \triangleq (I_{sj}^t)_{(s,j)\in\mathcal{S}\times J}$ the vector of inventory of each size s remaining in each store j from all previous periods up to the beginning of the current one. Other problem data includes: P_j the unit selling

price at store j, assumed exogenous here (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); and $(\boldsymbol{\lambda}_j)_{j\in J} \triangleq (\lambda_{sj})_{(s,j)\in\mathcal{S}\times J}$ the vector of demand rates for each size s at each store j. Finally, the expectation in (9) and second state dynamics equation (12) feature a discrete random component $\boldsymbol{\omega} \triangleq (\omega_{sj})_{(s,j)\in\mathcal{S}\times J}$ representing the number of units of each size s that will be sold in each store j between the receipt of the shipment being determined in period t and the next replenishment in the following period $t+1$. Note that as described in §3.1.1 the distribution of ω_{sj} depends on the entire vector $\mathbf{x}_j + \mathbf{I}_j^t$ of inventory of all sizes available in store j, and by definition of the expected sales function g introduced in §3.1.2 and appearing in (9), $g_{\boldsymbol{\lambda}_j}(\mathbf{x}_j + \mathbf{I}_j^t) \triangleq \mathbb{E}[\sum_{s\in\mathcal{S}} \omega_{sj}]$ for every store j.

While the formulation (DP) just stated in (9)-(12) turned out to be an insightful conceptual modeling step, several considerations made us question the feasibility and soundness of actually using it as the core of an operational decision support system:

- Even though (DP) only considers a single reference, the curse of dimensionality is still quite severe here, with the state space dimension of the order of the number of stores (more than a thousand) times the number of sizes (up to eight), the number of possible state values yet considerably larger, and the relevant time horizon possibly extending to the 24 weeks of a selling season. In addition, because of the large number of references shipped simultaneously (possibly more than a hundred) and the little time available between the availability of the demand forecast information and the time at which shipment decisions must be determined, any operational implementation must satisfy very stringent running time constraints (at most a few seconds). For more details on implementation issues, see §4 and Correa (2007).

- Perhaps more fundamentally, formulation (DP) assumes all demand forecast information to be both exogenous and appropriately summarized by the demand rate vector $\boldsymbol{\lambda}$. In reality, similar mean demand forecast estimates may have very different accuracy levels, for example demand forecasts for a given reference are substantially less reliable in the first week or two after its market introduction than later in its life cycle, after many weeks of demand have been observed. Besides, these different forecast accuracy levels, which affect the effective predicted variability of demand, should thus seemingly impact shipment decisions (as suggested by the classical newsvendor model, see further discussion below). Conversely shipment decisions, in that they impact whether and how many sales may be observed in a given store, also directly affect the process of demand learning driving how forecast quality changes over time. In fact, a prior study of a stylized dynamic assortment model (Caro and Gallien 2007) suggests that such demand learning dynamics may have a first-order effect. As a result, the modeling choice of a DP formulation explicitly capturing one type of dynamics (inventory) while ignoring another

one possibly just as important (demand learning) may seem questionable. Unfortunately, the data immediately available to us in this specific operational setting did not allow to construct or validate a descriptive model of forecast evolution. Even if that data had been available however, it is not clear at all that the likely more complex DP resulting from the addition of demand learning dynamics would have been easy to solve (even approximately) within the running time constraints mentioned above. We believe that this issue warrants further investigation beyond the limited time window that was available to us for this collaborative project, as discussed in §5.

- Finally, observe that the objective (9) represents total expected sales, as opposed to profits. The core issue here is not that the formulation features the unit selling price P_j instead of (say) the gross unit margin, as this data could be easily substituted in the objective definition. The problem is rather that (9) ignores important additional costs incurred when shipping either not enough inventory (missed sales resulting from inaccurate forecast) or too much inventory (store transhipments, returns to warehouse, markdowns) to each store – the usual underage and overage costs defined in the classical newsvendor framework. In the present setting however, each store manager determines transhipments and warehouse returns independently (see §1), and they do not appear to follow any formal guidelines when doing so. In addition, our attempts to identify systematic patterns for these warehouse returns and transhipment decisions have so far been unsuccessful. This also complicates any explicit modeling of price markdowns at the end of the selling season, because units that are unsold by then may have been offered for sale in several different stores at a regular price beforehand.

The alternative problem formulation we have used instead in order to partly circumvent the difficulties listed above is the following:

$$(MIP) \qquad \max \quad \sum_{j\in J} P_j z_j + K\big(\sum_{s\in\mathcal{S}}(W_s - \sum_{j\in J} x_{sj})\big) \tag{13}$$

$$s.t. \quad \sum_{j\in J} x_{sj} \le W_s \quad \forall s \in \mathcal{S} \tag{14}$$

$$z_j \le \big(\sum_{s\in\mathcal{S}^+} \lambda_{sj}\big) y_j + \sum_{s\in\mathcal{S}^-} \lambda_{sj} v_{sj} \quad \forall j \in J \tag{15}$$

$$y_j \le a_i(\lambda_{sj})(I_{sj} + x_{sj} - i) + b_i(\lambda_{sj}) \quad \forall j \in J, s \in \mathcal{S}^+, i \in \mathcal{N}(\lambda_{sj}) \tag{16}$$

$$v_{sj} \le a_i(\lambda_{sj})(I_{sj} + x_{sj} - i) + b_i(\lambda_{sj}) \quad \forall j \in J, s \in \mathcal{S}^-, i \in \mathcal{N}(\lambda_{sj}) \tag{17}$$

$$v_{sj} \le y_j \quad \forall j \in J, s \in \mathcal{S}^- \tag{18}$$

$$z_j, y_j \ge 0 \quad \forall j \in J; v_{sj} \ge 0 \quad \forall (s,j) \in \mathcal{S}^- \times J; x_{sj} \in \mathbb{N} \quad \forall (s,j) \in \mathcal{S} \times J \tag{19}$$

In the formulation (MIP) just stated, the primary decision variables x_{sj} represent as before the shipment quantities of each size s to each store j, which are still subjected to the warehouse inventory constraint (14). 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 that result from 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 data $\boldsymbol{\lambda}_j \triangleq (\lambda_{sj})_{s\in\mathcal{S}}$. Constraints (15)-(18) satisfied by z_j and the auxiliary variables y_j and v_{sj} along with the maximization objective (13) ensure indeed that in any optimal solution to (MIP), $z_j = \tilde{g}_{\boldsymbol{\lambda}_j}(\mathbf{x}_j + \mathbf{I}_j)$ where $\tilde{g}$ is the approximate expected sales function defined in (8).

While the first term in the objective (13) represents as before the expected sales revenue in the current period, the second term is a departure from the previous formulation. Specifically, it provides an evaluation for the total inventory remaining in the warehouse after the shipment decisions considered are executed, using an exogenous unit value K for that inventory which 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 transhipment 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 towards the end of the planned shelf life of a reference.

The warehouse inventory value K appearing in (13) thus effectively allows the shipment output 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 in practice the determination of that control to the users' appreciation and experience with the model. In addition, it is clear that the second term in (13) is only a very simple approximation of the expected revenue-to-go function $\mathbb{E}_{\boldsymbol{\omega}}[V_{t+1}(\mathbf{W}^{t+1}, \mathbf{I}^{t+1})]$ appearing in the objective (9) of the DP formulation discussed above (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. Specifically, Federgruen and Zipkin (1984) in particular 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 (the present finite horizon setting arguably makes this first problem harder), 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 $\sum_{s\in\mathcal{S}}\sum_{j\in J} x_{sj}$ 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 $\sum_{s\in\mathcal{S}}\sum_{j\in J} x_{sj}$ is constrained to be equal to that value.

4. Field Implementation Experiment

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 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 performance metrics used (in §4.1.2) and our impact assessment method (in §4.1.3). Results are then discussed in §4.2.

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 experiment would be best organized within a single one of them. Despite initial plans to test the model within the Children's section (the smallest of the three), after a few successful runs of the optimization model using historical data it was decided that the new process would be tested using fifteen references of the (largest) Women's section. These references were chosen by a Zara manager knowledgeable with the entire Women's collection, with the goal of selecting as representative a sample as possible. Because references 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 reference 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[2]. Specifically, the new inventory allocation process was only implemented in Arteixo at some point during the life-cycle of the fifteen references mentioned above, while the old process was still used to distribute all references (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 references and a carefully selected control group of paired references, conducted with data from stores assigned to Arteixo. Because both the sample and the control groups of references were distributed using the old 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 reference in the sample group, the warehouse team in Arteixo thus switched from the old to the new process at some point after July 2006, possibly after that reference had been offered in stores for one or several weeks already, 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 that 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, and we are now deeply convinced that this implementation strategy turned out to be the right decision. Unfortunately, we were not able to choose the exact week when the model would be used for the first time for each reference. 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 references 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 references, which corresponds to approximately the same proportions as in the overall assortment.

[2]Because 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.

4.1.2 Performance Metrics.

We now present the framework we developed to measure the 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 references across a network of many selling locations.

The primary data available to us included sales (V^d_{rsj}), shipments to store (X^d_{rsj}) and returns to warehouse/transhipments from store (R^d_{rsj}), all expressed in number of units, on each day d of the entire experiment period, for each available size s of each reference r in a group of 118 (including the pilot references described in §4.1.1), for every Zara store j in the world. From these we could derive the corresponding daily inventory positions (I^d_{rsj}) as described in §3.1.1, and compute the corresponding weekly sales $Sales^w_{rsj}$, shipments $Shipments^w_{rsj}$ and returns/transhipments $Returns^w_{rsj}$ series, by summing up the daily data series over each day d in each (calendar) week w. Finally, we computed the corresponding network-wide cumulative weekly series $Sales^t_r$, $Shipments^t_r$ and $Returns^t_r$ for each reference r, by summing up the previous series over all stores j in the network, sizes s of reference r, and weeks w in the selling season up to and including week t.

In order to quantify missed sales, we constructed data series $Demand^w_{rsj}$ and $Demand^t_r$, 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

$$DND^w_{rsj} \triangleq \sum_{d \in w} 1_{\{I^d_{rsj}=0\} \text{ or } \{\min_{\tilde{s} \in \mathcal{S}^+_r} I^d_{r\tilde{s}j}=0 \text{ and } \max_{\tilde{s} \in \mathcal{S}_r} V^d_{r\tilde{s}j}=0\}},$$

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

if $Sales^w_{rsj} > 0$ *and* $DND^w_{rsj} < 7$ **then**
  $Demand^w_{rsj} = Sales^w_{rsj}\left(\frac{7}{7-DND^w_{rsj}}\right)$
else
  $Demand^w_{rsj}$ = 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). 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), and therefore feel that this simple approach is appropriate given our purposes.

We used the ratio of cumulative sales to cumulative shipments $S/S^t_r \triangleq Sales^t_r/Shipments^t_r$, or *shipment success ratio*, as our primary metric for quantifying any excess inventory (i.e. overstock) in Zara's network. It represents the fraction of all units of a given reference shipped to stores since the beginning of the season that have actually been sold to date. That metric was actually used and closely monitored by Zara long before our collaboration, in particular when assessing the performance of product managers and planning the clearance sales period.

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 $S/D^t_r \triangleq Sales^t_r/Demand^t_r$, or *demand cover ratio*, where the cumulative weekly demand series $Demand^t_r$ is calculated analogously to $Sales^t_r$ and $Shipments^t_r$. 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.

Because cumulative sales $Sales^t_r$ are always lower than both cumulative shipments $Shipments^t_r$ (by constraint) and cumulative demand $Demand^t_r$ (by construction) for any week t and reference r,

$$Sales^t_r \leq \min\left\{Shipments^t_r,\ Demand^t_r\right\} \tag{20}$$

so that both metrics defined above are dimensionless fractions, in contrasts with (say) the objective of the newsvendor model, which involves the weighed sum of any overstock and understock quantities. This feature facilitates a comparison of distribution performance across references having possibly very different characteristics. In particular, the process scope considered here does not include any design, purchasing, promotion or advertising decisions (see §1) affecting how much sup-

ply is available initially, or what demand will be for any given reference. It thus seems appropriate that our proposed metrics focus more on how well supply is matched with demand, as opposed to what supply and demand are. Finally, 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 Figure 3.

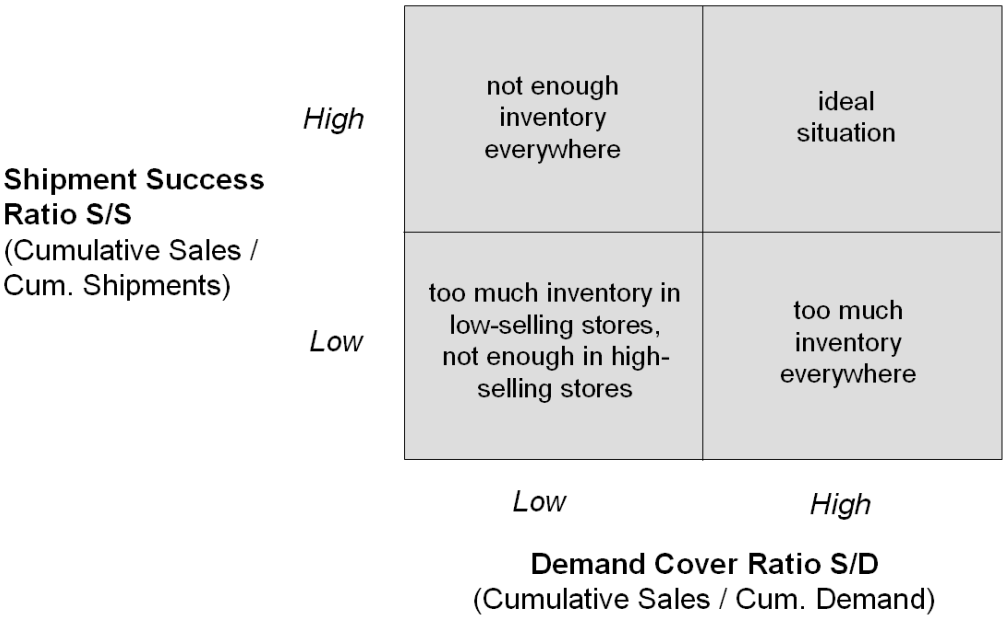

Figure 3: Proposed Evaluation Framework for Zara's Distribution Performance.

Besides the two primary metrics just discussed, we used three additional secondary metrics. The first is the *Stock Retention ratio*, defined as $SR_r^t \triangleq 1 - Returns_r^t/Shipments_r^t$. It thus represents the fraction of units shipped to date that were not sent to another store or sent back to the warehouse 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).[3] Our last two metrics are the *Store Cover ratio* (SC_r^t) and the *Display Cover ratio* (DC_r^t), formally defined as

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

The store cover ratio is thus 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

[3] At Zara, store transhipments and returns to the warehouse also require the approval of the regional manager.

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. Besides, the shipment success ratio S/S of a given reference 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 [4]. The increasing inventory scarcity over a typical life cycle of a Zara reference 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. However, several distribution managers at Zara emphasized to us that in their experience increasing the shipment success ratio of a given reference by a given amount was increasingly difficult for higher initial values of that metric. Likewise, limiting the decrease of the demand cover ratio and all other metrics from a given point in time was perceived to be considerably more challenging when the starting value of these ratios is closer to one. In order to mitigate the non-linear relationships between all the metrics just defined and these managerial notions of Zara's distribution performance, we thus also consider the logarithmic transformations $-\ln(1 - S/S)$ and $\ln(M)$, where M is any of the other metrics S/D, SR, SC and DC.

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 references over the same period of time had the new inventory process not been employed (i.e. the counterfactual), may not 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 references included in the pilot

[4] From now on we omit the indices t and r when no ambiguity arises.

experiment) with others in the population at large (the references still distributed using the old process).

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

$$\frac{|S/S_r^t - S/S_{\tilde{r}}^t|}{\max\{S/S_r^t, S/S_{\tilde{r}}^t\}} + \frac{|S/D_r^t - S/D_{\tilde{r}}^t|}{\max\{S/D_r^t, S/D_{\tilde{r}}^t\}}, \tag{21}$$

where r is the pilot reference considered, $\tilde{r}$ its matched control reference, and t is the week before the new process was used for the first time to distribute reference r. That is, the notion of proximity across references 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 on this matching procedure independently in the Arteixo and Zaragoza warehouses for the ten pilot references (although the new process was only used in Arteixo, see §4.1.1). For one reference in Arteixo the control reference initially selected was a clear outlier with an unusual 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 reference and repeated the matching procedure. Its final outcome is summarized in Table 2.

Note that both performance ratios show significant correlation among pilot and control references in Arteixo. For that warehouse the mean and median of the S/S ratios across references 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 references. 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 references, 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

	Arteixo	**Zaragoza**
Number of pilot references matched	10	10
Mean (median) shipment success ratio S/S of pilot references	52.3%(46.8%)	50.0%(46.7%)
Mean (median) shipment success ratio S/S of control references	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)	$2.00^{\diamond}(33)$
Mean (median) demand cover ratio S/D of pilot references	62.0%(63.4%)	61.1%(55.4%)
Mean (median) demand cover ratio S/D of control references	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 $^{\diamond}p \leq 0.1$, $^{*}p \leq 0.05$, $^{**}p \leq 0.025$, $^{***}p \leq 0.01$.

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

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 old, because this leaves less room for improvement to the pilot references – 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.

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

$$\Delta(M) \triangleq \left(M_r^{End} - M_r^{Start}\right) - \left(M_{\tilde{r}}^{End} - M_{\tilde{r}}^{Start}\right), \tag{22}$$

where M_r^{Start} (resp. $M_{\tilde{r}}^{Start}$) is the value of the metric considered for the pilot (resp. control) reference 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 (22) thus provides an estimate for the specific impact of the new process employed on the metric considered: the differences within parentheses excludes any time period other than that when the new process was used from consideration, 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 references. Because the old process was used for both pilot and control references in Zaragoza, expression (22), 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 (22), 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 the new process, relative to either shipments (S/S) or demand (S/D). Rearranging the terms defining $\Delta(S/S)$ for example yields

$$\Delta(S/S) = \left(\frac{Sales_r^{End} - \frac{Sales_r^{Start}}{Shipments_r^{Start}} \cdot Shipments_r^{End}}{Shipments_r^{End}} \right) - \left(\frac{Sales_{\tilde{r}}^{End} - \frac{Sales_{\tilde{r}}^{Start}}{Shipments_{\tilde{r}}^{Start}} \cdot Shipments_{\tilde{r}}^{End}}{Shipments_{\tilde{r}}^{End}} \right); \tag{23}$$

while the two terms in parenthesis in (23) respectively correspond to the pilot reference and the control reference 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 old process, based on conditions immediately preceding the implementation of the new one. Because of inequality (20), 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 references of the values obtained for each metric using equation (22), 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 references is justified by the need to factor out the randomness (noise) that we cannot control (indeed, the focus on such 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 references), a difference from zero has to be fairly large for it to be statistically significant.

Overall, Table 3 indicates that the new allocation process has a positive impact on the primary metrics defined in §4.1.2. These results are not driven by outliers since the mean and median changes have all the same sign, and the statistical significance is also consistent. Considering all pilot references, 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%. 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 reference in Arteixo (i.e. basic or fashion) with the respective estimation errors measured in Zaragoza. For basic references, 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 references, 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 impacts the two main types of references in different ways. For basic references its benefits would mostly stem from improvements in the demand cover S/D, whereas for fashion references they would consist of improvements in the shipment success ratio S/S. These differences are plausibly explained by forecast biases. Indeed, systematic forecast underestimation errors would generate shipments favoring the S/S ratio to the detriment of the S/D ratio, whereas overestimation errors would do the opposite. Posterior conversations with Zara indicate that the forecasting method does seem to behave differently depending on the type of reference. Also supporting that interpretation is the observation that the correlation between the individual S/S and S/D ratios of each reference is negative for both warehouses, but significantly more so in Arteixo (-0.75) than in Zaragoza (-0.40). 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 references. In the case of the S/S ratio, the apparent poor performance of the model for basic references has at least two alternative explanations: (i) the same two (out of four) basic pilot references that had negative S/S performance in Arteixo also performed badly (in fact, worse) in Zaragoza, indicating that the choice of the basic pilot references was particularly adverse; and (ii) the initial values of the S/S ratio for the basic references 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 references in Arteixo the changes in mean and median of the log transform of the S/S ratio 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 references, 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 references 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.

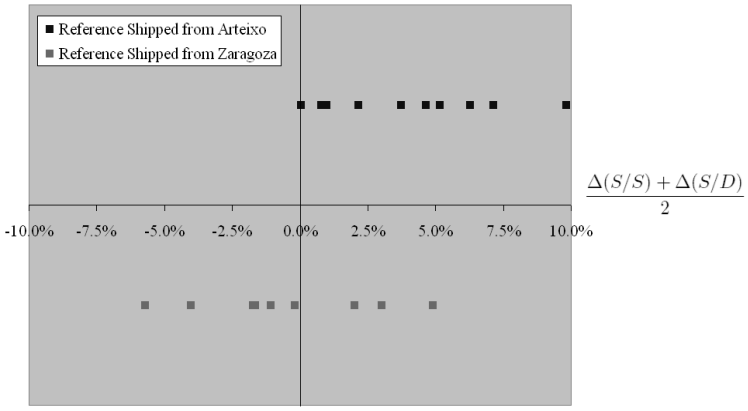

Figure 4: Estimated Relative Change in Sales for Each Pilot Reference in Arteixo and Zaragoza.

We believe however that our most significant results stem from considering for each reference 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 (23), 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 reference 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 reference, 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 reference 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 reference, with a mean across references of 4.1% (median 4.2%), whereas the corresponding estimation error calculated using data from Zaragoza is centered around zero (mean and median across references are 0.7% and -0.6% respectively).[5] 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

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

retailer like Zara given this improvement does not appear to require any major costs.[6] While this is the best impact estimate we were able to compute from the data available to us, we acknowledge however that it is based on a reduced number of references and assumes a particular method to predict the counterfactual. These limitations prevent us in particular from constructing a rigorous confidence interval around that estimate, which we must accept as a limitation of our findings.

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 references, suggesting that the model effectively reduces the level of transhipments. 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 references, since for fashion references the impact on these ratios (and their log transforms) remains just under the estimation error. We note that basic references have life-cycles that sometimes span the whole season, whereas fashion references 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 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.

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 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 transhipments, and increases the proportion of time that the references are on display. Following this pilot test, Zara made the decision to use this new allocation process for all its references and stores. As of the time of writing, this large scale deployment is

[6] If revenues increase 3% as well, then under a 12% net margin (as Inditex had in 2005), the corresponding increase in profits would be 25%.

now complete, and 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 also 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 allocation process through head count reductions, however we expect that 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 optimization 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, an analysis of the DP formulation described in §3.2, 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 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 José Antonio Ramos Calamonte for his uplifting friendship, tireless work, unfailing 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 García for his intuition, high energy and tough questions, Miguel Díaz for his trust, vision and wisdom and José Manuel Corredoira (Pepe) for his hard work developing the software user and data interfaces. Joaquín Lorenzo, Jesús González, Juan Quintela, María Ventín also spent considerable time sharing their invaluable knowledge of Zara's operations with us, and we are also grateful to Ramón Fernández, 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 and the University of Chicago for helpful feedback and discussions. 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.

Appendix

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

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

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 $h^{\mathcal{A}}$ 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 $h^{\mathcal{A}}$ is also supermodular (example 2.6.2 (f) and Corollary 2.6.2 in Topkis 1998).

5.2 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 references are particularly significant for Zara, as all the colors available of these references (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 references which is different from that described in §3.1.1:

- In addition to the distinction between major and minor sizes mentioned earlier, for multicolor references 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 reference 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 reference 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 references will remain on display, independently of which of their sizes may be missing, as long as there at 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 references, 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 references 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 references, 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 $\mathcal{C}^- \triangleq \mathcal{C} \backslash \mathcal{C}^+$. 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

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

where $\overset{2}{\max}$ 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

$$g_{\boldsymbol{\lambda}}^c(\mathbf{q}) = \begin{cases} \lambda_{\mathcal{S}^+}^c \mathbb{E}[\tau_{\mathcal{S}^+}^c \wedge T] + \sum_{s \in \mathcal{S}^-} \lambda_s^c \mathbb{E}[\tau_{\mathcal{S}^+ \cup \{s\}}^c \wedge T] \text{ if } c \in \mathcal{C}^+ \\ \sum_{s \in \mathcal{S}} \lambda_s^c \mathbb{E}[(\tau_{\mathcal{C}^+} \vee \tau_{\mathcal{S}^+}^c) \wedge \tau_s^c \wedge T] \text{ if } c \in \mathcal{C}^- \end{cases}, \tag{24}$$

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 (24) 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

$$\begin{aligned} \mathbb{E}[(\tau_{\mathcal{C}^+} \vee \tau^c_{\mathcal{S}^+}) \wedge \tau^c_s \wedge T] &\leq \mathbb{E}[(\tau_{\mathcal{C}^+} \vee \tau^c_{\mathcal{S}^+}) \wedge T] \wedge \mathbb{E}[\tau^c_s \wedge T] \\ &\approx (\mathbb{E}[\tau_{\mathcal{C}^+} \wedge T] \vee \mathbb{E}[\tau^c_{\mathcal{S}^+} \wedge T]) \wedge \mathbb{E}[\tau^c_s \wedge T]. \end{aligned} \tag{25}$$

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 (25) can be approximated using the same techniques as described in §3.1.3, we propose to approximate the first expectation as

$$\mathbb{E}[\tau_{\mathcal{C}^+} \wedge T] \approx \max^2_{c \in \mathcal{C}^+} \mathbb{E}[\tau^c_{\mathcal{S}^+} \wedge T], \tag{26}$$

then approximate the operand in (26) 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:

$(MIP-MC)$

$$\max \quad \sum_{j\in J} P_j z_j + K\big(\sum_{c\in\mathcal{C}}\sum_{s\in\mathcal{S}}(W_s^c - \sum_{j\in J} x_{sj}^c)\big) \tag{27}$$

$$s.t. \quad \sum_{j\in J} x_{sj}^c \le W_s^c \quad \forall c\in\mathcal{C}, s\in\mathcal{S} \tag{28}$$

$$z_j \le \sum_{c\in\mathcal{C}^+}\left(\big(\sum_{s\in\mathcal{S}^+}\lambda_{sj}^c\big)y_j^c + \sum_{s\in\mathcal{S}^-}\lambda_{sj}^c v_{sj}^c\right) + \sum_{c\in\mathcal{C}^-}\sum_{s\in\mathcal{S}}\lambda_{sj}^c u_{sj}^c \quad \forall j\in J \tag{29}$$

$$y_j^c \le a_i(\lambda_{sj}^c)(I_{sj}^c + x_{sj}^c - i) + b_i(\lambda_{sj}^c) \quad \forall c\in\mathcal{C}, j\in J, s\in\mathcal{S}^+, i\in\mathcal{N}(\lambda_{sj}^c) \tag{30}$$

$$v_{sj}^c \le a_i(\lambda_{sj}^c)(I_{sj}^c + x_{sj}^c - i) + b_i(\lambda_{sj}^c) \quad \forall c\in\mathcal{C}^+, j\in J, s\in\mathcal{S}^-, i\in\mathcal{N}(\lambda_{sj}^c) \tag{31}$$

$$v_{sj}^c \le y_j^c \quad \forall c\in\mathcal{C}^+, j\in J, s\in\mathcal{S}^- \tag{32}$$

$$u_{sj}^c \le a_i(\lambda_{sj}^c)(I_{sj}^c + x_{sj}^c - i) + b_i(\lambda_{sj}^c) \quad \forall c\in\mathcal{C}^-, j\in J, s\in\mathcal{S}, i\in\mathcal{N}(\lambda_{sj}^c) \tag{33}$$

$$u_{sj}^c \le r_j^c \quad \forall c\in\mathcal{C}^-, j\in J, s\in\mathcal{S} \tag{34}$$

$$r_j^c \le y_j^c + M(1-k_j^c) \quad \forall c\in\mathcal{C}^-, j\in J \tag{35}$$

$$r_j^c \le \ell_j + Mk_j^c \quad \forall c\in\mathcal{C}^-, j\in J \tag{36}$$

$$\ell_j \le y_j^c + Mh_j^c \quad \forall c\in\mathcal{C}^+, j\in J \tag{37}$$

$$\sum_{c\in\mathcal{C}^+} h_j^c = |\mathcal{C}^+| - 2 \quad \forall j\in J \tag{38}$$

$$y_j^c \ge 0 \quad \forall (c,j)\in\mathcal{C}^+\times J; \tag{39}$$

$$v_{sj}^c \ge 0 \quad \forall (c,s,j)\in\mathcal{C}^+\times\mathcal{S}^-\times J; u_{sj}^c \ge 0 \quad \forall (c,s,j)\in\mathcal{C}^-\times\mathcal{S}\times J; \tag{40}$$

$$x_{sj}^c \in \mathbb{N} \quad \forall (c,s,j)\in\mathcal{C}\times\mathcal{S}\times J; \tag{41}$$

$$k_j^c \in \{0,1\} \quad \forall (c,j)\in\mathcal{C}^-\times J; h_j^c \in \{0,1\} \quad \forall (c,j)\in\mathcal{C}^+\times J \tag{42}$$

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 (29), to the approximation discussed above for the sum over all colors and sizes of the expected sales in store j, $\sum_{c\in\mathcal{C}} g_{\boldsymbol{\lambda}_j}^c(\mathbf{I}_j + \mathbf{x}_j)$ where g is defined in (24). Specifically, constraints (33)-(34) 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 (35)-(36) and (42) likewise ensure that $r_j^c = y_j^c \vee \ell_j$, constraint (30) ensures that y_j^c is equal to our approximation for $\mathbb{E}[\tau_{\mathcal{S}^+}^c \wedge T]$ in store j, finally constraints (37)-(38) and (42) ensure that $\ell_j = \max_{c'\in\mathcal{C}^+}^{2} y_j^{c'}$.

5.3 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 reference 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 reference 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 databases 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 $\mathcal{S}^+$ 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).

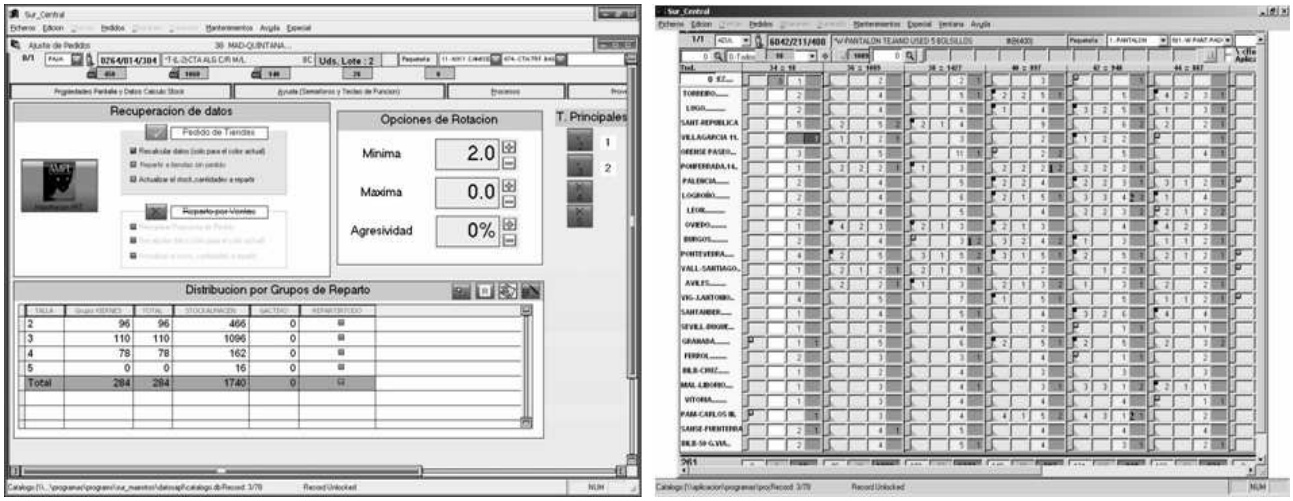

Figure 5: Screenshots of the Software Implementation.

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 refered 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 reference and size to each store. Every such screen corresponds to a reference, each row corresponds to a store, each group of columns refers to a size in which that reference 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 these 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 he 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.

References

Axsäter, S., J. Marklund and E. A. Silver. 2002. Heuristic Methods for Centralized Control of One-Warehouse, N-Retailer Inventory Systems. *Manufacturing & Service Operations Management* **4**(1) 75-97.

Barber, B. M. and J. D. Lyon. 1996. Detecting Abnormal Operating Performance: The empirical Power and Specification of Test Statistics. *Journal of Financial Economics* **41** 359–399.

Bertsekas, D. 2005. *Dynamic Programming and Optimal Control (Vol. I)*. Athena Scientific, Nashua, NH.

Cachon, G. and M. Lariviere. 1999. Capacity Choice and Allocation: Strategic Behavior and Supply Chain Performance. *Management Science* **45**(8) 1091-1108.

Caro, F. and J. Gallien. 2007. Dynamic Assortment with Demand Learning for Seasonal Consumer Goods. *Management Science* **53**(2) 276-292.

Correa, J. 2007. *Optimization of a Fast-Response Distribution Network.* Masters Thesis. Leaders for Manufacturing, 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. *Management Science* **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.

Ferdows, K., J. AD Machuca and M. Lewis. 2003. Zara. The European Case Clearing House. Case 603-002-1.

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. *Management Science* **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. *Management Science* **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. *Management Science* **34**(7) 880-895.

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. *Management Science* **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.

Karatzas, I. and S. E. Shreve. 1991. *Brownian Motion and Stochastic Calculus.* Springer-Verlag, New York, Second Edition.

McAfee, A., V. Dessain, and A. Sjöman. 2004. ZARA: IT for Fast Fashion. Harvard Business School Case 9-604-081.

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. *Management Science* **44**(3) 285-300.

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

Zhang, S. and G. J. Fitzsimons. 1999. Choice-Process Satisfaction: The Influence of Attribute Alignability and Option Limitation. *Organizational Behavior and Human Decision Processes* **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 references	-2.2%(-1.8%)	10.1%(8.6%)	15.0%(12.7%)	19.1%(18.8%)
Mean (median) impact on fashion references	6.4%(7.4%)	1.9%(2.0%)	18.6%(23.4%)	15.5%(9.4%)
Mean (median) impact on all references	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($35^{\diamond}$)	$3.17^{**}(43^{*})$	$3.31^{***}(47^{**})$
Zaragoza				
Mean (median) error for basic references	-5.3%(-5.0%)	2.6%(1.9%)	-23.3%(-21.5%)	8.6%(8.0%)
Mean (median) error for fashion references	-0.5%(-0.3%)	4.6%(5.0%)	-5.2%(-0.2%)	8.4%(17.1%)
Mean (median) error for all references	-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 $^{\diamond}p \leq 0.1$, $^{*}p \leq 0.05$, $^{**}p \leq 0.025$, $^{***}p \leq 0.01$.

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 references	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 references	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 references	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	$2.17^{\diamond}(31^{\diamond})$	$2.19^{\diamond}(37^{\diamond})$	$2.59^{*}(41^{*})$	$2.19^{\diamond}(31^{\diamond})$	$2.38^{*}(39^{*})$	$2.82^{**}(43^{*})$
Zaragoza						
Mean (median) error for basic references	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 references	-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 references	-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 $^{\diamond}p \leq 0.1$, $^{*}p \leq 0.05$, $^{**}p \leq 0.025$, $^{***}p \leq 0.01$.

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