The Value of Fit Information in Online Retail: Evidence from a Randomized Field Experiment

Santiago Gallino  
Tuck School of Business, Dartmouth College, Hanover, NH 03755, santiago.gallino@tuck.dartmouth.edu

Antonio Moreno  
Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston, IL 60208, a-morengarcia@kellogg.northwestern.edu

Online channels generate frictions when selling products with non-digital attributes, such as apparel. Customers may be reluctant to purchase products they have not been able to try on and those customers who do purchase may return products when they do not fit as expected. Virtual fitting-room technologies provide information about how a product fits a particular customer and promise to mitigate some of the frictions the information gap generates in the retailers’ supply chains. By implementing a series of randomized field experiments, we study the value of virtual fit information in online retail. In our experiments, customers are randomly assigned to a treatment condition where virtual fit information is available or to a control condition where virtual fit information is not available. Our results show that offering virtual fit information increases conversion rates and order value, and reduces fulfillment costs arising from returns and home try-on behavior, i.e., customers ordering multiple sizes of the same product. We explore mechanisms through which providing virtual fit information helps customers and retailers. We argue that the virtual fitting tool creates spillovers even to products that are not available for virtual try-on, increases loyalty, helps customers better parse their choice sets, and reduces uncertainty by providing size recommendation.

Key words: supply chain information; fit uncertainty; online retail; randomized field experiment; virtual fitting room.

1. Introduction

The apparel retail industry reached a size of $230 billion in 2014 in the U.S. alone. Online transactions account for an important and growing share of the total revenues (26 percent in 2014). For example, Gap Inc., the third largest apparel retailer in the world, grew its total revenues by 3.2 percent between 2012 and 2013, with online revenue increasing 21.2 percent in the same period. Nordstrom, Inc., fourth in the list of largest apparel retailers, presents a similar trend. In 2008, online revenue represented 8 percent of the company’s total revenue, while in 2013 this number increased to 13 percent. Table 1, which lists several multichannel apparel retailers and their share of online sales for the years 2009 and 2013, shows that the share of online sales has been increasing for apparel retailers across the board. Online sales have continued to grow at a faster rate than total sales. In January 2015, Abercrombie & Fitch Co. reported that online sales accounted for
more than 25 percent of its revenues, and Urban Outfitters, J.Crew and Lilly Pulitzer reported numbers closer to 30 percent. All this suggests that online transactions are very substantial already and are likely to become even more important in the near future.¹

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Online Sales as Percentage of Total Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company Name</td>
<td>Year</td>
</tr>
<tr>
<td>-----------</td>
<td>------</td>
</tr>
<tr>
<td>Gap Inc</td>
<td>14.2</td>
</tr>
<tr>
<td>Abercrombie &amp; Fitch Co.</td>
<td>18.9</td>
</tr>
<tr>
<td>Lululemon Athletica Inc.</td>
<td>16.5</td>
</tr>
<tr>
<td>Nordstrom, Inc.</td>
<td>12.9</td>
</tr>
<tr>
<td>Express Inc.</td>
<td>15.3</td>
</tr>
<tr>
<td>American Apparel Inc.</td>
<td>9.7</td>
</tr>
<tr>
<td>JC Penney Company Inc.</td>
<td>9.1</td>
</tr>
<tr>
<td>Kohl’s Corporation</td>
<td>8.9</td>
</tr>
<tr>
<td>Aeropostale Inc.</td>
<td>10.4</td>
</tr>
</tbody>
</table>

Source: Bloomberg

Brick-and-mortar and online retail channels differ in how they accomplish the two most fundamental channel functions: delivering information and fulfilling transactions (Bell et al. 2014). For apparel retailers, the traditional brick-and-mortar channel provides customers with the opportunity to physically inspect the products and assess fit information in physical stores where transactions are fulfilled as well. In contrast, the online channel typically provides product information through the web and transactions are fulfilled via direct-to-customer delivery from a centralized distribution center. The lack of physical access to inspect products prior to purchase generates friction when selling products with non-digital attributes, such as apparel (Lal and Sarvary 1999). Customers may be reluctant to purchase products they have not been able to try on and those customers who do purchase may return products when they do not fit as expected. As a consequence, the rate of returns for online channels is usually much higher than for brick-and-mortar channels.

The increasing importance of the online channel is redrawing the competitive landscape in retail. New retailers have entered the market as strictly online companies in an attempt to establish themselves in the most dynamic section of the industry. Traditional retailers have been enhancing their ability to interact with customers using multiple channels, putting customers at the center of their “omnichannel approach” and offering new combinations of information and fulfillment (Bell et al. 2014). These changes are having a profound impact on retailers’ supply chains, which originally were designed to deliver products to stores and have to be adapted to a scenario where a large fraction of the sales is delivered direct to customer. Companies also have to deal with an

increasing number of returns they often are not well equipped to handle, with substantial costs in shipping, handling, and liquidation (Tang 2016, Ng and Stevens 2015).

Online retailers differ in how they are trying to mitigate the challenge of not being able to provide customers with physical access to the products prior to making a purchase. Companies face a trade-off between the quality of the information they provide to their customers and the fulfillment costs they experience. For example, retailers like Zappos offer a very liberal return policy, so that customers may overcome their reluctance to make a purchase due to fit uncertainty. This results in low information costs but very high fulfillment costs for the retailer, as customers may order multiple sizes, try them on at home, and eventually return those products that do not satisfy them (Banjo 2013). Some online retailers, such as Bonobos and Warby Parker, have responded to the fit information challenge by creating offline showrooms where customers can “try on” products prior to placing an order. These offline showrooms provide better fit information and have been shown to reduce fulfillment costs (Bell et al. 2017).

This paper focuses on virtual fitting rooms, another innovation that tries to overcome the information gap that online customers suffer. A virtual fitting room is the online equivalent of the in-store changing room. It enables customers to see how clothes look on them and check different sizes and styles but virtually, i.e., using an Internet-connected device, rather than physically, i.e., trying them on in a store. This technology started to receive attention from the industry media around 2010, and is now available from an increasing number of companies that have developed proprietary solutions, including Fits.me, Metail, Virtusize and Shoefitr. More recently, a growing number of prominent apparel retailers—QVC, Nordstrom, and Amazon, among many others—are offering this technology in their online stores.

The objective of this paper is to study the impact of virtual fitting-room technologies on demand and fulfillment costs. We do that by collaborating with Metail, one of the pioneering companies in the virtual fitting-room space. Previous work has explored the issue of product uncertainty in e-commerce (Dimoka et al. 2012, Hong and Pavlou 2014, Kim and Krishnan 2015, De et al. 2013). However, that work has focused on observational data and is subject to potential endogeneity issues that are hard to assess. Our partnership with Metail enables us to run a series of randomized field experiments where the availability of the virtual fitting technology is randomly assigned to a fraction of the users visiting the site of a retailer that has adopted the technology. This allows us to measure the causal effect of offering fit information.

We hypothesize that the availability of virtual fit information will result in higher conversion rates and order values and lower fulfillment costs that arise from a reduction of returns and home try-on behavior (customers ordering multiple sizes of the same product to assess fit in their homes).

We conduct two large-scale randomized field experiments to test these hypotheses. Randomized field experiments have received a lot of attention recently in the economics and business literature (e.g., two interesting examples in the operations management literature are Johnson Ferreira et al. 2015 and Retana et al. 2015, and in information systems Lee and Hosanagar 2016) because they provide a very clean way to identify causal effects, overcoming issues related to confounding factors that may result in bias in observational studies.

Our approach allows us to establish that the availability of virtual fit information increases demand and reduces fulfillment costs. On the demand side, customers in the condition that had access to the tool were more likely to place an order. For customers who placed an order, those who had access to the tool had higher order value, with orders containing on average more items and more expensive products. The virtual fitting tool is particularly effective at encouraging sales of the most expensive products.

On the cost side, we find that the availability of virtual fit information decreases fulfillment costs. We observe this through a reduction in returns and a reduction in home try-on behavior for customers who have access to the virtual fitting tool. This implies that providing virtual fit information can reduce the pressure that online channel growth is putting on retail supply chains.

We explore mechanisms through which providing virtual fit information helps both customers and retailers. We argue that the virtual fitting tool creates spillovers even to products that are not available for virtual try-on, increases loyalty, helps customers better parse their choice sets, and reduces uncertainty by providing size recommendation.

To the best of our knowledge, this is the first study proposing a large-scale randomized field experiment to analyze the impact of fit information on operational metrics in online retail. Our results have important implications for retailers operating online channels. Offering richer information about fit can increase sales and customer engagement and reduce the fulfillment costs incurred by the supply chain even when serving online customers.

2. The Value of Fit Information: Theory and Hypotheses

2.1. Literature Review
Our paper builds upon previous work by researchers in marketing, information systems, and operations management.

The discrepancy in the online and offline channels’ ability to deliver product information has long been recognized as a key issue in e-commerce research. For example, Lal and Sarvary (1999) draw a distinction between digital and nondigital product attributes and how information about each is communicated in online and offline channels. Digital attributes, such as the price of a product or length of a book, suffer no loss of information when communicated online, whereas non-digital
attributes, for example, the texture of a shirt or look of a pair of glasses, may introduce significant uncertainty for some consumers when presented or characterized online.

Since consumers are uncertain about product fit before inspecting/trying the product (Ofek et al. 2011), categories where nondigital attributes are important face stronger impediments in the adoption and use of online channels. Online retailers operating in categories with nondigital attributes recognize that they face significant challenges communicating product information to customers. These companies employ a variety of strategies—e.g., providing free two-way shipping, establishing pop-up stores or physical showrooms, allowing home try-on, and related methods—to combat these challenges. Previous work has also studied information provision in online retail via offline showrooms (Bell et al. 2017, Gao and Su 2016). The focus of this paper is different. We study the impact of technologies that provide rich product-fit information without leaving the online world. In particular, we assess the value of virtual fitting-room technologies. We hypothesize that providing customers with the opportunity to try on clothes virtually will reduce the friction that arises from selling categories with important nondigital attributes.

Recent work has explored questions related to product uncertainty and its implications for e-commerce. For example, Dimoka et al. (2012) theorize that product uncertainty is a major impediment for e-commerce that can be reduced by IT solutions. In their analysis, product uncertainty is defined as “the buyer’s difficulty in assessing the product’s characteristics and predicting how the product will perform in the future.” They are concerned with two facets of product uncertainty: description uncertainty (difficulty obtaining reliable information about the product’s true quality) and performance uncertainty (difficulty predicting how the product will perform in the future). A key difference between the notion of product uncertainty those studies consider and the notion of product uncertainty we consider is that we are not concerned with an absolute measure of quality but with how close a product is to an individual’s preference. Hong and Pavlou (2014) explore the distinction between quality uncertainty and product fit uncertainty, studying the effects of pictures and product forums as a way to reduce product fit uncertainty. Kim and Krishnan (2015) examine online purchase behavior over time and, among other findings, report that providing product information in digitized commercial videos induces higher sales. Our work further explores the issue of product fit uncertainty by studying the effect of a technology specifically developed to mitigate the problem of product fit uncertainty.

Previous related work has studied how technology usage affects sales and returns. De et al. (2010) consider search and recommendation technologies and find their usage has a positive impact on product sales, while De et al. (2013) consider the effects of photo visualization and zoom technologies on product returns. Our work extends those findings, considering the effect of virtual fitting tools on sales and returns. Furthermore, we study the effects of virtual fitting tools on
certain user behaviors that had not been considered by previous literature, such as repurchase and home try-on.

Another key difference between our study and the aforementioned papers is methodological. Previous papers rely on observational data. While the datasets are interesting and detailed, it is not possible to completely rule out endogeneity issues arising from omitted variables or other confounding aspects. We are able to conduct a series of large scale randomized field experiments that allow us to rule out any of those potential issues in our assessment of the value of providing richer virtual fit information. This approach has been used to study related topics, such as the effectiveness of different recommendation systems (Lee and Hosanagar 2016). To the best of our knowledge, our study is the first to provide this type of analysis to assess the value of fit information.

Several papers in the marketing-operations interface have studied different aspects of product returns, including key drivers (Petersen and Kumar 2009) and how a return option affects consumer utility (Anderson et al. 2009). Other authors (e.g., Guide et al. 2006, Blackburn et al. 2004) focus on how to reduce the negative impact of returns by improving the reverse supply chain design or through contractual provisions, such as optimal fees or money-back guarantees (e.g., Davis et al. 1995, 1998, Shulman et al. 2009, Su 2009). Fit has also been the subject of some recent modeling research. For example, Gu and Xie (2013) study the strategic decision of retailers in a competitive market to facilitate fit revelation as a function of product quality, and Gu and Liu (2013) study how retailers can influence fit search with their shelf layout decisions. Our work complements this literature by empirically estimating the causal impact of providing virtual fit information.

Finally, our work is related to supply chain fulfillment and the emergent trend of omni-channel retail. We show that information and fulfillment can act as substitutes: investing in information can reduce fulfillment costs. Previous work considering fulfillment in e-commerce supply chains includes Netessine and Rudi (2006), who derive conditions under which retailers prefer drop-shipping to holding inventory, and Randall et al. (2006), who examine fulfillment choices by online retailers and find that firms carrying high-margin products and less variety are more likely to invest in holding inventory. Recent work exploring omni-channel retail includes Bell et al. (2017), Gallino and Moreno (2014), and Gallino et al. (2014).

2.2. Hypotheses

We develop two groups of hypotheses. The first group relates to the effect of virtual fit information on demand, including conversion and order characteristics. The second group is related to fulfillment costs and includes the effects on returns and home try-on behavior.

Regarding the effect of virtual fit information on demand, we note that consumers can obtain product information via direct, indirect, or virtual experience (Li et al. 2003). Since a virtual
fitting tool provides a type of virtual reality experience, previous research on virtual reality can inform our hypothesis development. Li et al. (2003) run a series of lab experiments and find that customers able to interactively control a 3D product-visualization system are more likely to develop positive brand attitudes and stronger purchase intentions. Building on the framework developed by Steuer (1992) that highlights virtual reality’s the vividness (ability to produce a sensorially rich mediated environment) and interactivity (degree to which users can modify and influence the form or content of the mediated environment), Jiang and Benbasat (2007) conduct a series of lab experiments confirming that both attributes influence consumers’ intentions to purchase and to return to a website. More recently, Kim and Krishnan (2015) use field data to demonstrate that showing video increases purchase probability.

The virtual fitting tool can also affect customers’ actions by encouraging what the behavioral economics literature refers to as the “endowment effect”: the phenomenon that people ascribe more value to things they own (Kahneman et al. 1990, Carmon and Ariely 2000). Experiments show that subjects tend to pay more to retain something they feel they own than to obtain something they do not own. This phenomenon can be explained by loss aversion, since relinquishing an object a person owns can be perceived as a loss (Kahneman et al. 1990). A practical consequence for retailers is that by increasing consumers’ feelings of psychological ownership of a product, it is possible to increase their perception of the product’s value, making them more willing to buy it. Retailers in the physical world sell clothes in a way that allows consumers to imagine ownership and encourages touching, trying on and interacting with garments. The act of trying on clothing is likely to increase customers’ attachment to the item and the value they ascribe to it. The virtual fitting tool provides similar stimuli. Customers create their 3D models and compose outfits and are able to rotate the model and see it from a variety of different angles. Norton et al. (2011) document an increase in the valuation of self-made products (what has been called “the IKEA effect”). In the context of the virtual fitting tool, creating outfits requires some effort that may increase the perception of ownership and valuation of the products.

In light of this discussion, we hypothesize that virtual product-fit information will increase the probability of purchase and the amount of money customers spend on the website. These hypotheses are summarized in Table 2, along with mapping to the studies that tested them (presented in Section 3) and the summary of their support, based on the empirical evidence we present in Section 4.

Turning our attention to the effect of fit information availability on fulfillment costs, previous work has explored the effect of some specific types of information, such as photo visualization and zoom technologies (De et al. 2013) or pictures and product forums (Hong and Pavlou 2014) on returns. Product returns are costly to all parties. Consumers spend time, money, and effort
returning products. Sellers experience direct fulfillment costs from shipping the products back to their distribution center and stores, as well as depreciation as product value diminishes over time (Guide et al. 2006). Previous work (e.g., Hong and Pavlou 2014) has hypothesized that having more information available before making a purchase should result in lower return rates. In the online world, customers usually form expectations of fit before placing an order. After they receive the item they assess the utility of keeping the product versus returning it. Customers will be unlikely to return the product if it turns out to be as expected or better than expected but will be likely to return the product if it does not at least confirm their expectations (Kopalle and Lehmann 1995).

Following Hong and Pavlou (2014), we build on Salop’s circular city model (Salop 1979) and assume that a consumer’s taste corresponds to a location in a circular space with a product located at the center. The distance between the customer’s taste and the product indicates the product-preference mismatch, which can lead to dissatisfaction and product returns. Customers are uncertain about their preferences, which is represented by a circular space around their actual taste. Product fit information reduces the likelihood of expectation disconfirmation, lessening the likelihood the customer will return the product (Kopalle and Lehmann 1995). Consequently, we hypothesize that the virtual fitting tool will reduce the probability of returns.

In some situations, customers may be unsure about the right size of a product for them. Absent product fit information, customers can hedge by including multiple sizes of the same product in their order, later returning those sizes that do not fit. We call this behavior “home try-on,” as the customer uses orders as a device to assess fit of the products at home. This behavior has been widely documented. For example, it has been noted that “[r]etailers are zeroing in on high-frequency returners like Paula Cuneo, a 54-year-old teacher in Ashland, Mass., who recently ordered 10 pairs of corduroy pants in varying sizes and colors on Gap Inc.’s website, only to return seven of them. Ms. Cuneo is shopping online for Christmas gifts this year, ordering coats and shoes in a range of sizes and colors. She will let her four children choose the items they want—and return the rest” (Banjo 2013). We hypothesize that this kind of behavior will be reduced when customers are offered fit information.

The hypotheses regarding the virtual fitting tool’s effects on fulfillment costs are also summarized in Table 2. If they are supported, we could interpret that fit information and fulfillment costs operate as substitutes. A firm could choose not to provide fit information, but that would result in higher fulfillment costs. Alternatively, a firm may want to reduce fulfillment costs by investing in technology to provide fit information.

3 For a more formal argument, see Hong and Pavlou (2014).
### Table 2 Summary of Main Hypotheses

<table>
<thead>
<tr>
<th>Hypotheses Regarding Demand</th>
<th>Description</th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis 1</td>
<td>Virtual product fit-information increases the probability of purchase.</td>
<td>Supported</td>
<td>Supported</td>
</tr>
<tr>
<td>Hypothesis 2</td>
<td>Virtual product fit-information increases the average order amount.</td>
<td>Supported</td>
<td>Supported</td>
</tr>
<tr>
<td>Hypotheses Regarding Fulfillment Costs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypothesis 3</td>
<td>Virtual product-fit information reduces the probability of returns.</td>
<td>Supported</td>
<td>N/A</td>
</tr>
<tr>
<td>Hypothesis 4</td>
<td>Virtual product-fit information reduces “home try-on” behavior.</td>
<td>N/A</td>
<td>Supported</td>
</tr>
</tbody>
</table>

3. **Research Setting**

To implement our study we partner with Metail (http://www.metail.com), one of the pioneers offering virtual fitting-room solutions to online retailers selling female apparel, and with a large online apparel retailer in Latin America adopting the Metail tool. We develop two independent large-scale field experiments to ascertain the effects of customized fit information and test our hypotheses. A brief description of our research setting and each of the studies follows, along with a summary of how our hypotheses are mapped to each of the studies. We also conduct two supplemental tests with other retailers to explore additional issues regarding the mechanism. We discuss these experiments in Section 5.

3.1. **Virtual Fitting-Room Technologies: Metail**

There are several technologies offering virtual fitting-room functionality, such as the ones provided by Fits.me, triMirror, and Zugara. We partner with Metail, a U.K.-based startup offering a virtual fitting tool for female apparel. The tool this company developed allows customers to create virtual models of themselves by providing some of their key measurements (height, weight, bust size, inches in waist and hips). To further enhance the virtual model, the customer is able to customize the model’s skin tone and hairstyle. Once the customer has completed these steps, she can try different outfits on her personalized model. The tool allows customers to rotate their models for better visualization and also provides size recommendations. Figure 1 shows what customers see when using the tool to “try-on” clothes.

Retailers adopting the technology need to apply a digitization process to make the try-on feature available. The efficiency and speed of this step in the process is a key component of Metail’s technology. The different virtual products customers can try on are digitized through a series of pictures that scan each product from a 360-degree perspective. The process of digitizing a garment takes less than 10 minutes and the costs are estimated in the order of 5 british pounds or about US$6.3.
Although this tool can be available for any type of apparel retailer, it is expected that strictly online retailers will benefit most from it since they do not have physical stores in which customers can try on the different garments before purchasing them. Our main retailer partner is a large online apparel retailer that operates in Latin America; we cannot disclose the retailer’s name due to a confidentiality agreement.

3.2. Study 1

Our first study focuses on evaluating the impact of the virtual fitting tool on demand and returns. To do this, we run a randomized experiment. The retailer created two versions of the site, one where the Metail tool was available (treatment condition) and one where the Metail tool was not available (control condition). We implement the experiment as an A/B test, with some customers engaging with the treatment version at the same time as other customers were engaging with the control version of the site. The retailer digitized 330 of their garments. Before this experiment was conducted, the Metail tool had not been available at this retailer and customers participating in the experiment were unlikely to have been exposed previously to the tool, since this was the first implementation of the technology in Latin America.

Customers enter the experiment when they visualize a page that contains one of the digitized garments. At that point they are assigned to either the treatment or the control version and remain in that condition until the end of that session. The experiment is designed so that 80% of the
sessions are assigned to the treatment condition and 20% to the control condition. The assignment is done by generating a random number between 0 and 1. Numbers lower than 0.8 assign customers to the condition where Metail is available, while numbers greater than or equal to 0.8 assign customers to the condition where Metail is not available.

Since we can only make the virtual tool available to customers but cannot force them to use it, our analysis explores intention-to-treat (Imbens and Rubin 2015). We mainly focus on the relationship between the assignment to a condition and the observed outcomes—i.e., we compare the outcomes of the group assigned to the Metail condition and the group assigned to the control condition, regardless of whether the virtual fitting tool is actually used. Because in this experiment treatment assignment is purely random, such comparison can be performed, with inference relying on randomization.

The field experiment was in place for 59 days (between August 23, 2013, and October 21, 2013). At the end of the experiment, we obtain data about all the orders placed in sessions that were part of the experiment, a total of 435,982. For each of those orders, we observe a time stamp indicating the date and time the order was placed, total order value, total number of items included in the order (allowing us to calculate the average price of an item in the order), and an indicator of whether the order included any products from the digitized set. We also obtain data on returns for each order, a dummy variable indicating whether any of the products included in the order had been returned. Because we collected this data at the end of the experiment and the retailer has a one-month return policy, the products bought in the second half of the experiment could still be returned. Consequently, for our analysis of returns we focus on the orders placed during the first half of the experiment, for which our returns data is complete and reliable.

Appendix 1 provides a description of the data available to us. This experiment had some limitations in terms of the elements we are able to manipulate and the data we are able to obtain. First, we do not have any information on sessions that were part of the experiment but in which an order was not placed (therefore, we cannot directly calculate conversion rate). Second, we were not able to collect any customer-specific information (for example, repeated purchases by the same customer). Third, because the assignment to a condition was only maintained during a session, a customer who revisited the site multiple times may have switched between treatment and control conditions. Finally, because we only have information about the orders placed (which are affected

---

4 One potential concern could be that because some sessions in the control condition could correspond to users who had encountered the tool in the past, those sessions could have lower sales due to a suppression effect, and this potentially could explain a superior performance of the treatment condition sessions. However, note that the chance of a customer in the control condition previously accessing the tool is minimal. The virtual fit tool was not available to retailer customers before Study 1 was run. If a customer revisited the site during the study, the chance that a customer in the control condition previously was exposed to the tool is still small. Not all revisits are potentially
by the treatment allocation) and not about the sessions assigned to the experiment, we are not able to provide a table that shows evidence of covariate balance for this experiment, although it should have been achieved by construction from the fact that assignment was purely random. Study 2 largely addressed these limitations.

3.3. Study 2

The goal of Study 2 is to dig deeper into the results obtained in Study 1 with a larger set of garments representing a broad set of product categories and with more detailed product information. The latter allows us to assess, for example, whether users who are assigned to the treatment engage in less home try-on behavior. Additionally, we track users in this experiment, assigning them to a condition when they enter the experiment and keeping them in that condition for the rest of the experiment. We also identify repurchases during the experiment.

As with Study 1, Study 2 is implemented as an A/B test. The retailer digitized 1,000 garments from different product categories, including dresses, jackets/coats, knitwear, shoes, shorts, skirts, tops, and trousers. The garments were selected among those that had high sales volume (allowing us to measure the effects without a prohibitively long experiment) and high stock levels (so that we do not face stock-outs in the garments of interest during the experiment’s execution). To drive traffic to the test, the retailer offered a series of curated collections of the Metail digitized garment range, each with a specific theme. However, the Metail technology was not specifically marketed or advertised. Although the tool had been available to some customers during Study 1, the vast majority of the participants in Study 2 had no previous exposure to the tool.

As in Study 1, customers enter the experiment when they visualize a page containing one of the digitized garments and at that point are assigned to one of the two versions of the site (treatment or control). However, unlike Study 1, in Study 2 users assigned to the treatment condition remain in the treatment condition for the whole duration of the experiment. The experiment is designed so problematic, since sessions are only assigned to one of the conditions after visiting one of the digitized garments (330 garments for Study 1). In most cases, customer revisits are not reassigned because the subset of digitized garments is small compared to the total number of available garments. Even when a customer is re-randomized, the probability of being reassigned to the same category as before is about 68%. Among the remaining 32%, suppression could only affect those who were assigned to the tool initially but are now assigned to the control (16%), and out of those, we would only be concerned about those who actually see the tool and engage with it. Given the low expected incidence of a potential suppression effect, we do not expect this to result in substantial biases. This is confirmed by the results of Study 2, which maintain the treatment allocation constant for any given customer.

Customers from Study 1 were neither excluded from Study 2 nor targeted to be a part of it. While we are unable to track customers from Study 1 to Study 2, only a very small fraction of customers in Study 2 may have participated in Study 1. The retailer was in a growth period and, based on our conversations with them, most of the customers who participated in Study 2 had been recently acquired. Study 1 had fewer garments than Study 2 (i.e., during Study 1, the Metail tool was not available for most of the garments). The experiments were more than six months apart, so even if a small fraction of customers in Study 2 had participated in Study 1, we do not expect this to significantly alter their experience.
that 50% of the users are assigned to the treatment condition and 50% to the control condition. When the user reaches a page with a digitized garment, as part of the Metail initialization code the following steps are performed: (1) Check the current state of the test, i.e., check that the test is “on,” and (2) check whether the user is already assigned to a test group for Study 2. For users already in the test, the site shows or hides the technology as appropriate. For those not in the test, the initialization code performs random assignment by generating a random number between 0 and 1. For numbers lower than 0.5, the user is added to the treatment group, and for numbers greater than or equal to 0.5, the user is added to the control group.

As in Study 1, because we can only make available the virtual fitting tool but cannot enforce usage, our analysis explores intention-to-treat (Imbens and Rubin 2015). We mainly focus on the relationship between assignment to a condition and the observed outcomes.

We run the experiment for two months (between June 3, 2104, and August 4, 2014), during which a total of 2,389,655 users visit one of the digitized garments and are assigned to one of the groups. For all those users (even for those who do not place an order), we observe the group they are assigned to, their IP address, and the list of orders they place, if any. We obtain data about all the orders placed during the experiment by users who are assigned during that time period, a total of 201,757. For each of those orders, we observe a time stamp indicating date and time when the order is placed, the total order value, the total number of items included in the order (allowing us to calculate the average price of an item in the order), an indicator of whether the order includes any products from the digitized set, and an indicator of whether the user has engaged with the Metail tool in the session in which the order is placed. We also observe a list of the items included in each order and their prices. For the products in the digitized set, we observe the category, a description, and the size. In this experiment we did not track returns. Appendix 1 provides a description of the data available to us for this experiment.

While we cannot show the covariate balance achieved by this randomization by showing extensive customer data, we can provide evidence that the randomization indeed achieves covariate balance in the variables to which we have access. In particular, because we can geolocate the sessions IP address, we can generate variables that compute the distance to different geographic points. We choose three Brazilian cities in different regions: Brasilia, S˜ao Paulo, and Belo Horizonte. Using the

---

6 As with any A/B testing study or any retargeting technology, our process can only track customers over time if they are using the same IP address and/or browser. The A/B testing worked by setting a cookie in the users browser when they first were assigned to the test. This meant that even if the user connected via a different IP, if they were still using the same browser the code would see they had been assigned to a condition already so they would remain in the same test group. If they cleared their cookies or visited the retailer on a different device, they would be considered as new users and be assigned a new condition.

7 We excluded duplicate orders and orders coming from the retailer’s IP addresses. Our results were robust to including/excluding those observations.
distance between the geolocated customer IP and those cities, we can compare them for treatment and control to show that the groups are balanced in these generated covariates. The mean distance to these cities for treated and control users is shown in Table 4. We can observe that the mean of these covariates is very similar across groups. Indeed, the standardized differences in all cases is smaller than 0.006, much below the threshold conventionally considered to indicate good covariate balance, which is typically 0.1 (Austin 2011). This shows that the randomization indeed works, and we expect it to balance observed and unobserved covariates.

3.4. Mapping Hypotheses to Studies

Study 1 and Study 2 complement each other in conducting our main analysis of the value of fit information, summarized by the hypotheses listed in Table 2. The main appeal of Study 1 is that it tracks returns, which is not possible with Study 2. Study 2 has more detailed product information for the digitized set, allowing us to measure home try-on behavior. Study 2 also tracks users, including those who did not place orders (which allows a more fine-grained analysis of conversion), follows those placing multiple orders (which allows us to study repurchase behavior), and includes information about engagement with the tool. Note that the number of observations are not directly comparable. The periods are different. Also, because the periods are different and users are assigned to the tests when they first click on a page with a garment of the digitized set, the amount of traffic assigned to the experiment depends on the site location of the garments.

Hypotheses 1 and 2, regarding the effects of virtual product-fit information on demand, can be explored in both Studies 1 and 2. The fact that we can run the same analysis with two independent studies serves as a robustness check for the specific conditions of each study. Hypothesis 3, on the effects of virtual product-fit information on returns, can only be studied with Study 1, while Hypothesis 4, on the effects of virtual product-fit information on home try-on behavior, can only be tested with Study 2 since it captures the products ordered, allowing us to check whether multiple sizes were ordered. The analyses that we do in Section 5 to dig deeper into the mechanism are largely based on Study 2 and two supplemental experiments with other retailers.

Table 3 summarizes the differences between the two studies and Figure 2 presents a schematic view of the experiments, based on the second study.

4. The Value of Fit Information: Demand and Fulfillment Costs

We divide our empirical results into two sections. In this section we analyze the results of Studies 1 and 2 and examine the effects of virtual fit information on demand (including conversion and basket characteristics) and fulfillment costs, which we proxy with the incidence of returns and home try-on behavior. In Section 5 we further explore the effects of fit information and its potential mechanisms using Studies 1 and 2 as well as other additional experiments.
Table 3  Randomized Field Experiments

<table>
<thead>
<tr>
<th></th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>Aug. 23 to Oct. 21, 2013</td>
<td>June 3 to Aug. 4, 2014</td>
</tr>
<tr>
<td>Number of digitized garments</td>
<td>330</td>
<td>1,000</td>
</tr>
<tr>
<td>% customers assigned to Metail</td>
<td>80%</td>
<td>50%</td>
</tr>
<tr>
<td>Number of users in the experiment</td>
<td>N/A</td>
<td>2,389,655</td>
</tr>
<tr>
<td>Number of orders collected</td>
<td>435,982</td>
<td>201,757</td>
</tr>
<tr>
<td>Conversion tracked</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Repurchase tracked</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Product-level information tracked</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Engagement information tracked</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Returns tracked</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Testable hypotheses</td>
<td>H1, H2, H3</td>
<td>H1, H2, H4</td>
</tr>
</tbody>
</table>

Figure 2  Study 2 — Diagram

4.1. Virtual Fitting and Demand
We start by analyzing the effects of the availability of the virtual fitting tool on the demand side. We have hypothesized above that the availability of virtual fit information will result in higher sales, coming from a higher conversion rate (larger fraction of consumers purchasing) and from larger orders. We can study these questions using any of the two independent studies (Study 1 and Study 2), although the approach in each case differs slightly since the data collected in each of the studies is different. We can also explore the effects of engaging with the Metail tool using Study 2.

4.1.1. Conversion Rate. As described above, Study 1 tracks all the orders placed during the experiment period. The specific information that we have access to is described in Appendix 1. In Study 1 we do not know whether an order includes products from the digitized set and we do not observe any nonpurchasing users. Consequently, it is not possible to obtain a direct measure of conversion. To test whether the tool’s availability results in a higher conversion rate, we compare the share of orders we observe from the Metail condition with the share of orders we would observe if the treatment had no effect. We know that 80 percent of the visiting sessions had been randomly assigned to the treatment condition. If the treatment had no effect, and sessions from the Metail and the control condition had the same probability of purchase, the share of orders from
Table 4  Summary of the Experiment Outcomes

<table>
<thead>
<tr>
<th>Study 1</th>
<th>Treatment</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total orders</td>
<td>349,688</td>
<td>86,294</td>
<td>435,982</td>
</tr>
<tr>
<td>Share of orders</td>
<td>80.21%</td>
<td>19.79%</td>
<td></td>
</tr>
<tr>
<td>Orders in the first 30 days</td>
<td>169,025</td>
<td>42,091</td>
<td>211,116</td>
</tr>
<tr>
<td>Returns from orders placed in the first 30 days</td>
<td>4,361</td>
<td>1,135</td>
<td>5,496</td>
</tr>
<tr>
<td>% of orders placed in the first 30 days that have returns</td>
<td>2.58%</td>
<td>2.70%</td>
<td></td>
</tr>
</tbody>
</table>

| Study 2 | | |
|---------| | |
| Users in the experiment | 1,218,498 | 1,171,157 | 2,389,655 |
| Total number of orders | 93,141 | 108,616 | 201,757 |
| Total number of orders with digitized products | 7,884 | 9,512 | 17,396 |
| Users placing an order | 78,159 | 68,326 | 146,485 |
| % of users placing at least one order | 6.41% | 5.83% | |
| Users placing an order with digitized products | 7,797 | 6,446 | 14,243 |
| % of users placing at least one order with dig. prod. | 0.64% | 0.55% | |
| Users placing orders, none with dig. prod. | 70,362 | 61,880 | 132,512 |
| % of users placing at least one order, none with dig. prod. | 5.77% | 5.28% | |
| Distance of user to Brasilia - mean | 942 | 942 | |
| Distance of user to Brasilia - std. dev. | 285 | 286 | |
| Distance of user to São Paulo - mean | 1,158 | 1,162 | |
| Distance of user to São Paulo - std. dev. | 702 | 705 | |
| Distance of user to Belo Horizonte - mean | 666 | 668 | |
| Distance of user to Belo Horizonte - std. dev. | 423 | 427 | |

<table>
<thead>
<tr>
<th>Study 2 - Engaged Users</th>
<th>Engaged</th>
<th>Not Engaged</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>69,593</td>
<td>1,148,905</td>
<td>1,218,498</td>
</tr>
<tr>
<td>Users with orders</td>
<td>11,914</td>
<td>66,246</td>
<td>78,160</td>
</tr>
<tr>
<td>Conversion</td>
<td>17.12%</td>
<td>5.77%</td>
<td></td>
</tr>
<tr>
<td>Users with orders including digitized products</td>
<td>3,004</td>
<td>4,794</td>
<td>7,798</td>
</tr>
<tr>
<td>Conversion</td>
<td>4.32%</td>
<td>0.42%</td>
<td></td>
</tr>
</tbody>
</table>

the Metail condition should be very close to 80 percent, up to sampling error. If the percentage of treated customers is higher in orders than in visits in a statistically significant way, we would conclude that the availability of virtual product-fit information has a positive effect on conversion. We formally test this using a binomial test (Hollander et al. 2013, Chapter 2). Table 4 shows the number of orders from the treatment and control conditions. We test the null hypothesis that the probability an order comes from the treatment group is 0.8 versus the alternative hypothesis that the probability an order comes from the treatment group is greater than 0.8. Given that we observe 435,982 orders, 349,688 of which come from the treatment condition, we can reject the hypothesis that the proportion is 0.8 (p < 0.01). A 95% confidence interval is (0.8011, 1).

Study 2 also assigns the availability of the fitting tool randomly, but unlike Study 1 it identifies whether an order contains a product from the digitized set. In addition, it tracks all users assigned during the test, including those who did not purchase. For example, as indicated in Table 4, out of the 1,218,498 users who were assigned to the Metail condition, 78,159 placed at least one order that was tracked during the experiment, 7,797 of which placed at least one order that included one or more items from the digitized set. In order to assess the impact of the tool’s availability
on the user conversion rate, we compare the proportion of users who placed at least one tracked order under the treatment \( (p_t) \) and control \( (p_c) \) conditions. We implement Fisher’s exact test (see Hollander et al. 2013, Chapter 10), which tests the null hypothesis that \( p_t = p_c \) versus the alternative \( p_t > p_c \). The test allows us to reject the null \( (p<0.01) \) and therefore we conclude that the proportion of users who place at least one order is higher under the treatment condition than the control condition.\(^8\) In a similar fashion, we compare the proportion of users who place at least one order that contains a product from the digitized set in the treatment and control units, and again we conclude that this proportion is higher for the treatment group \( (p<0.01) \).

Comparing the observed percentages gives us an idea of the magnitude of the differences. We observe at least one order for 5.83% of the users in the control group, versus 6.41% of the users in the treatment group. The difference is 0.58 percentage points, or a relative increase in the conversion rate of 10%. If we restrict our attention to orders that include at least one product from the digitized set, we observe at least one such order for 0.55% of the users in the control group, versus 0.64% of the users in the treatment group. The difference is 0.09 percentage points, or a relative increase in the conversion rate of 16%. Note that while the difference in the conversion for the digitized set may seem like a more accurate measure of the causal effect of the fitting technology (because it is based on the conversion in the products where the tool was actually available), it may overestimate the benefits of the technology to the firm. The availability of the tool for products in the digitized set could have prompted some customers to substitute non-digitized products they would have bought for products in the digitized set. Hence, not all the increase is necessarily incremental. Comparing the effects on total sales allows us to obtain a more conservative estimate that is net of any cannibalization.

Interestingly, if we restrict our attention to users who placed one or more orders, none of which included a product from the digitized set, we also find that the users assigned to the treatment condition presented a more favorable behavior. There were 70,362 such users \( (5.77\%) \) in the treatment condition versus 61,880 \( (5.28\%) \) in the control condition. Again using Fisher’s exact test (see Hollander et al. 2013, Chapter 10), which tests the null hypothesis that \( p_t = p_c \) versus the alternative \( p_t > p_c \), we are able to reject the null \( (p<0.01) \). The users in the Metail condition were also more likely than those in the control condition to purchase orders that did not include any digitized products. This suggests that there are positive spillovers of the technology that expand beyond the products in the digitized set and affect the products that were not digitized. We further analyze these spillovers in Section 5.

\(^8\) In the body of the text we focus on the Fisher exact test for ease of exposition, but we obtain the same results using a Chi-Squared test for all the analyses for which we report the results of the Fisher exact test.
In order to assess the statistical significance of the effects, we estimate a logit model of the user-level conversion (this is the fraction of the users in the experiment who placed at least one order) as a function of the randomly assigned availability of the fitting technology. A user “converts” overall \((CONV_o = 1)\) if they place an order during the experiment. We consider two types of users among those who convert. Users who “convert in the digitized set” \((CONV_d = 1)\) are those who buy at least one product from the digitized set. Users who “convert in the nondigitized set” \((CONV_n = 1)\) are the rest of the users who convert. Let \(p_x = \text{Prob}(CONV_x = 1)\). Then we can estimate the following logistic models:

\[
\logit \left( \frac{p_x}{1-p_x} \right) = \gamma_0 + \gamma_1 \text{METAIL} + \epsilon. \tag{1}
\]

where \(\text{METAIL} = 1\) for the treatment group and 0 for the control group.

Table 5 shows the results and the marginal effect of the fitting-tool availability on the conversion. Columns 1–3 consider overall conversion, conversion in the digitized set, and conversion in the nondigitized set respectively. The effects are positive and statistically significant for all three groups, with the effect being strongest in the digitized set. In Section 5 we explore potential mechanisms explaining why the effects are also positive for the conversion in the nondigitized set.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(M\text{ETAIL})</td>
<td>0.101***</td>
<td>0.006***</td>
<td>0.152***</td>
<td>0.001***</td>
<td>0.094***</td>
<td>0.005***</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.003)</td>
<td>(0.017)</td>
<td>(0.001)</td>
<td>(0.006)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>(\text{Constant})</td>
<td>-2.781***</td>
<td>-5.197***</td>
<td>-2.886***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.012)</td>
<td>(0.004)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>2,389,655</td>
<td>2,389,655</td>
<td>2,389,655</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*\(p<0.1\); **\(p<0.05\); ***\(p<0.01\)

The unit of observation is a user. Each of the three columns presents the coefficients of a logistic regression, followed by the marginal effects.

We analyze the difference between sales of the treatment and control group at different aggregation levels (e.g., by day, by product) and we find consistent patterns: the treatment group systematically has higher sales than the control group (see Appendix 2). The fact that we obtain similar

---

9 Note that we can only consider conversion at the user level, not at the session level, because we do not have traffic information. For all the users who participated in the experiment, we know the orders that they placed (or if they did not place an order at all), but we do not know how many times they visited the site. Consequently, it is more meaningful to consider conversion at the user level.
patterns in Studies 1 and 2, and also when examining the data at different aggregation levels (e.g., by day, by product) provides strong evidence supporting the hypotheses that the availability of the virtual fitting tool increases conversion rate.

4.1.2. Orders. Having established a causal effect on conversion, we turn our attention to the shopping baskets of customers who placed orders. As described above, we hypothesize that order characteristics will be affected by the availability of the virtual fitting tool (e.g., we expect customers who have access to virtual fit information to place larger orders on average). We focus on three order features: total order amount, average product price in the order, and number of products in the order. To operationalize this analysis, we rely on the fact that our data comes from a randomized experiment. This means we can compare the characteristics of orders made by customers who were randomly assigned to the treatment group (where the Metail tool was available) to those in the control group. This can be done with a two-sample t-test for each of the outcomes or with a regression model. We focus our description on the regression model, but there are no qualitative differences in the conclusions. To implement the analysis on the effect of total order amount, we use the following regression model:

\[ ORDER AMOUNT = \alpha_0 + \alpha_1 \cdot METAIL + \epsilon \]  

where \( METAIL \) is 1 for those customers who were in the treatment group and 0 for those in the control group.

The first column in Table 6 shows the results when we consider the actual order amount as the dependent variable. For Study 1 this result shows that on average customers in the treatment group spend approximately 5.39 more monetary units per order than those in the control group. For Study 2 this number is 4.92. The second column in Table 6 shows the results when we consider the log of actual order value as the dependent variable. In this case, we find that when the tool is available the order size increases by approximately 1.3 percent (Study 1) or 1.6 percent (Study 2) compared to when it is not available. Given the wide dispersion in order value, we prefer the specifications with logged dependent variables. Both specifications indicate that the availability of virtual fit information results in higher order amounts. Note that because the treatment is assigned randomly, the coefficients we obtain with these regressions are unbiased even if we do not include any additional control variables.\(^\text{10}\)

Columns 3 and 4 of Table 6 show a similar analysis, but using the average price of an item in the order as the dependent variable:

\[ AVERAGE PRICE = \beta_0 + \beta_1 \cdot METAIL + \epsilon \] 

\(^\text{10}\) Adding control variables such as day or time fixed effects does not alter the results. The results are available from the authors.
We observe that the average price of an item in the order is larger for customers in the treatment group compared to those in the control group; this result holds for a linear specification (column 3) and specification where the dependent variable is the log of the average price (column 4) for both Study 1 and Study 2. In addition to this regression analysis, we use Study 2 to explore whether the tool is particularly effective at encouraging orders of expensive products. We define a variable \( \text{HIGHPRICE} \) that we set to 1 for products with prices in the top quartile. We study the share of product sales coming from customers with/without access to the fitting tool for regular products (\( \text{HIGHPRICE} = 0 \)) and for products with prices in the top quartile (\( \text{HIGHPRICE} = 1 \)). If we use Fisher’s exact test (see Hollander et al. 2013, Chapter 10), we can reject the hypothesis that the incidence of products with \( \text{HIGHPRICE} = 1 \) is the same in the treatment and control conditions, in favor of the alternative hypothesis that the fraction of purchases with \( \text{HIGHPRICE} = 1 \) is higher in the treatment condition (\( p<0.05 \)). This suggests the tool helps when considering expensive products.
Finally, the last column in Table 6 shows evidence that orders in the treatment group include a larger number of items. For this analysis, we use a Poisson regression that takes into account the fact that the number of items in the order is a discrete variable (the results with a linear model are qualitatively similar). A potential explanation for this effect is that the virtual fitting tool may allow customers to try different products together as an outfit, which may result in including more items for some orders.

The similar patterns we obtain in Studies 1 and 2 provide strong evidence supporting the hypothesis that the availability of the virtual fitting tool results in larger orders (with more products that on average are more expensive).

Combining the effects we obtain on order amounts with the effects we discussed on conversion rate, we conclude that the virtual fitting tool results in higher revenues per user.

4.1.3. Effects of Engagement with Virtual Fitting Technology. Study 2 collected information regarding engagement with the Metail tool, which allows us to conduct an additional piece of analysis. We focus on users who had potential access to the virtual fitting tool, because only those users can engage with the tool. For each user who was part of the experiment and was randomly assigned to the treatment condition, we observe whether they engaged with the tool, which would happen if the user clicked on the “Try Me” button next to one of the digitized garments.

A total of 1,218,498 users were in the “Metail” condition and had the possibility of engaging with the tool. Of those users, 69,593 had some level of engagement with the Metail tool and the remaining 1,148,905 did not interact with the tool.

Table 4 shows the summary statistics for each of the groups. Of the users who did not engage with the Metail tool, 66,246 (5.77%) ended up placing at least one order and 4,794 (0.42%) ended up placing at least one order containing an item from the digitized set. Of the users who did engage with the Metail tool, 11,914 (17.12%) ended up placing at least one order, and 3,004 (4.32%) ended up placing at least one order containing an item from the digitized set.

The difference in the proportion of users placing an order between those who engage and those who do not is substantial (3x overall, 10x for the proportion of users who place an order from the digitized set). Using Fisher’s exact test (see Hollander et al. 2013, Chapter 10), we can reject the null ($p < 0.01$) that the proportions of purchasing users are the same for users who engage and those who do not.

The interpretation of these differences is not causal in this case. Engagement is not randomly assigned but is the result of the customer’s choice. The self-selection of customers who choose to engage with the tool will result in selection bias if the engagement variable can be correlated with variables that are unobserved to the econometrician but affect the conversion as well. It is
reasonable to think that customers who are more interested in buying a product are more likely to engage with the tool. This would result in overestimating the effect of engaging with the tool. In order to obtain a causal effect of using the tool, it is necessary to conduct an experiment that randomly assigns the possibility of using it, such as the one presented above in Section 4.1. However, the fact that we find many users who engage with the tool and those who engage have even higher conversion rates serves both as a reality check and a plausible mechanism for the aforementioned results. Interestingly, the user-level conversion rate for users assigned to the Metail condition but not engaged with the tool is not higher than the user-level conversion rate for users who did not have access to Metail. This suggests that the superior conversion rate of users assigned to the Metail condition can be attributed to those users who actually engaged with the tool.

4.2. Virtual Fitting and Fulfillment Costs

In addition to the effects on the demand side described above, we are interested in exploring whether offering a virtual fitting tool can result in cost savings for the retailer. Specifically, we focus on the impact of the virtual fitting tool on returns, which increase the average fulfillment costs through additional shipping, handling and restocking expense. If, as we have hypothesized in Section 2.2, the tool reduces product fit uncertainty, customers who can use the tool will buy on average products that fit them better, and the retailer would thus observe fewer returns and reduced associated fulfillment costs.

Study 1 directly tracked returns. Study 2 did not directly track returns. However, because Study 2 tracked product-level data for the digitized products, we are able to identify products purchased in different sizes in the same order. Ordering the same product in multiple sizes can be an indication of a customer engaging in home try-on behavior and this can provide an alternative, indirect measure of fulfillment costs. Section 4.2.1 describes our analysis of the effect of virtual fitting technology on returns, while Section 4.2.2 explores the effects on “home try-on” behavior.

4.2.1. Returns. The retailer tracked product returns using the company’s internal systems. At the end of Study 1 we obtained the list of orders for which at least one item had been returned by that time. Since returns are accepted during 30 days, we have accurate return information only for the orders placed during the initial 30 days of the experiment (211,116 orders), and we focus our analysis on that period.

Unlike the other analyses, in which being assigned to treatment and control was the result of flipping a coin with 80/20 probabilities and therefore the causal effects could be recovered using simple comparison of means, the analysis of returns is more subtle. In order to return a product, a customer must have bought it (which occurs with different probabilities for treated and untreated units). Also, product characteristics may affect the propensity to return a product. If
these characteristics are correlated with the treatment, they may confound the effects. For example, if customers are more likely to return more expensive products and if customers who use Metail are more likely to buy more expensive products (which we actually show above), a higher propensity to return in the Metail group could be caused by having bought more expensive products on average and not by merely having access to the Metail tool.

As explained above, in Study 1 we only have the orders placed by users who were assigned to the Metail versus non-Metail condition, but we do not have the entire allocation of users to the experiment (i.e., we only observe users who purchase). To recover the causal effect of the virtual fitting tool on returns, we want to compare orders that are very similar (e.g., in prices, etc.) except for the availability of the virtual fitting tool. Propensity score approaches (Rosenbaum and Rubin 1983) allow us to account for the covariates that predict receiving the treatment (placing an order after being in the “Metail” condition, in this case). One way to conduct this adjustment is by using inverse probability weighting (IPW), which weights each observation by the inverse of the probability of receiving treatment (Hirano et al. 2003). We choose this approach instead of other matching-based techniques because it allows us to use all the data available to us. IPW estimators use estimated probability weights to correct for the issue arising from the fact that customers may be affected by different factors when they choose to return products. When implementing the IPW estimators, we use a two-step approach: first, we estimate the parameters of the treatment model and compute the estimated inverse probability weights; second, we use the estimated inverse probability weights to compute weighted averages of the outcomes for each treatment level, or in a model that adjusts for additional factors. For a detailed description of this approach, see Wooldridge (2010) and Cameron and Trivedi (2005); Cattaneo (2010) presents an implementation of this technique.

To model the probability of treatment (order corresponding to the “Metail” condition), we use a logit model. The specific model we use in the results we report includes day of the week, day, and hour fixed effects (because users in the Metail condition could be more likely to buy on certain days or at particular times), the total order amount, the number of items in the order, the average price of the products in the order, and the price of the most expensive product in the order (since, as we saw above, these can differ for orders placed from the Metail and non-Metail conditions). Once we have the estimated probability as a function of those covariates, we calculate the weights as \( W_i = \frac{1}{e(x_i)} \) for the treated observations (those coming from the Metail condition) and \( W_i = \frac{1}{1 - e(x_i)} \) for the observations in the control condition, where \( e(x_i) \) is the estimated probability of treatment using the model described above (Guo and Fraser 2009). Once we have the weights, we can use them

---

11 We have estimated multiple models using different specifications (e.g., logit vs probit) and different control variables, and the results are always qualitatively consistent with those that we present.
in a comparison of means or proportions, such as the Fisher exact test (which actually suggests that the allocation to the Metail condition results in lower probability of return, \( p < 0.01 \)), or in a regression-based model, which is our preferred choice because it has a “doubly robust” property (Wooldridge 2010)—that is, the estimation is still consistent even if the regression model or the propensity score model is incorrect.

The model we consider to evaluate the impact on returns is a logit model, where the dependent variable is whether an item in an order was returned. We incorporate a number of covariates that can affect this decision, including day of the week, day, and hour fixed effects (because time of purchase could affect returns), the total order amount, the number of items in the order, the average price of the products in the order, and the price of the most expensive product. Table 7 shows the results of this analysis, with the different columns considering different subsets of the control variables.

<table>
<thead>
<tr>
<th>Table 7 Effect on Fulfillment Costs: Returns</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{DEPENDENT VARIABLE:} ) RETURN FLAG</td>
</tr>
</tbody>
</table>
| \( \begin{array}{cccc}
\text{METAIL} & (1) & (2) & (3) & (4) \\
-0.049^{**} & -0.050^{***} & -0.048^{**} & -0.047^{**} \\
(0.019) & (0.019) & (0.019) & (0.019) \\
\text{l(order amount)} & 0.502^{***} & 0.506^{***} & 0.739^{***} \\
(0.012) & (0.012) & (0.040) \\
\text{l(avg price)} & & & -0.478^{***} \\
& & & (0.059) \\
\text{l(max price)} & & & 0.188^{***} \\
& & & (0.060) \\
\text{n items} & & & -0.074^{***} \\
& & & (0.012) \\
\text{Constant} & -3.583^{***} & -6.068^{***} & -5.692^{***} & -5.454^{***} \\
(0.013) & (0.068) & (0.085) & (0.091) \\
\text{Day of the week fixed effects} & No & Yes & Yes & Yes \\
\text{Includes day fixed effects} & No & Yes & Yes & Yes \\
\text{Includes hour fixed effects} & No & No & Yes & Yes \\
\text{Observations} & 211,116 & 211,116 & 211,116 & 211,116 \\
\end{array} \) |

Notes:
- \( *p < 0.1; \quad **p < 0.05; \quad ***p < 0.01 \)
- Unit of observation is an order placed in the first 30 days of Study 1.
- All columns show logistic regression models using IPW weights.
- RETURN FLAG is 1 if at least a product in the order was returned. Order amount is the total amount of the order, avg_price and max_price are average and max prices of the items in an order, and n_items is the number of items.
Our preferred specification is Column 4 (full model), but the coefficient of interest is very similar across all specifications. In all cases, we conclude that users assigned to the Metail condition are less likely to return products, the rest being equal. The marginal effect of Metail on the probability of returning an order is -0.12 percentage points; it changes the probability of returning from 2.70% to 2.58% (note that in this market returns are much less frequent than in the US or Western Europe, where returns are often higher than 20%). In other words, providing fit information reduces average fulfillment costs.

To make sure the weights are actually correcting for observable differences in treated- and control-unit characteristics, we follow Guo and Fraser (2009) and we run weighted regressions of the continuous covariates in the treatment. If the propensity score weighting effectively removes imbalances, the coefficient of the treatment variable should not be significant. Table 15 (Appendix 3) shows it is the case.

4.2.2. Home Try-on Behavior. While Study 2 did not track returned products, it includes product information (for those products in the digitized set) that allows us to identify the orders containing multiple sizes of the same product. Some customers ordering the same product in different sizes are likely to be doing so with the intention of trying on different sizes at home before returning those that do not fit well. This behavior is very costly for the retailer.

The prevalence of returns in the region where the focal retailer is located is small compared to the U.S. and Europe, so we do not expect to find many such transactions in our data set. However, if we are able to demonstrate that using virtual fitting technologies reduces the incidence of this behavior in this market, this would suggest that savings in other markets where returns are more frequent can be very important.

We use the variable $HTO$ (for “home try-on”) to flag the products that are ordered concurrently with other sizes of the same product. Out of the 26,237 products ordered from the digitized set during the trial (14,142 when the tool was available versus 12,095 when it was not), 185 were considered part of an $HTO$ transaction. This behavior is less prevalent for customers in the treatment condition (85 products in the treatment versus 100 in the control condition, despite having more purchases from the digitized set when the virtual tool was available). Comparing the proportions using Fisher’s exact test (see Hollander et al. 2013, Chapter 10), we can reject the null ($p<0.05$) that the HTO proportion is the same in the treatment and control groups.

To assess whether the difference in HTO behavior by users in the treatment and control groups is statistically significant and get a sense of the magnitude, we estimate a logistic regression of HTO on the treatment indicator. The results are shown Table 8. These results confirm that the difference

---

12 http://blogs.forrester.com/zia_daniell_wigder/14-10-21-key_metrics_in_brazils_ecommerce_market
between the incidence of HTO in the treatment and control groups is statistically significant, with HTO behavior being less likely when the tool is available. The marginal effect is 0.2 percentage points, changing probability of HTO behavior from 0.8% when the tool is not available to 0.6% when the tool is available—a 25% reduction in the HTO incidence rate.

### Table 8 Effect of Home Try-on. Logistic Regression.

<table>
<thead>
<tr>
<th></th>
<th>Dependent Variable: HTO</th>
</tr>
</thead>
<tbody>
<tr>
<td>METAIL</td>
<td>−0.321**</td>
</tr>
<tr>
<td></td>
<td>(0.148)</td>
</tr>
<tr>
<td>Constant</td>
<td>−4.787***</td>
</tr>
<tr>
<td></td>
<td>(0.100)</td>
</tr>
<tr>
<td>Observations</td>
<td>26,237</td>
</tr>
</tbody>
</table>

*Note:* *p<0.1; **p<0.05; ***p<0.01
A unit of observation is a purchased product.
Logistic regression.

Taking together these effects of HTO and returns that we have documented in the preceding subsection, we conclude that the virtual fitting tool reduces fulfillment costs for the retailer.

5. Exploring the Value of Fit Information

So far, we have shown the main effects resulting from the availability of the virtual fitting tool on demand and fulfillment costs. We now focus on exploring different avenues through which the availability of the virtual fitting tool can help. We start by analyzing whether the effects we identified in the previous section are exclusively from the products in the digitized set (5.1). We then study how the virtual fitting technology increases customer engagement and loyalty (5.2) and how the virtual fitting tool helps customers narrow their choice sets (5.3). Finally, we study the value of the tool’s size recommendation (5.4). Taken together, these analyses allow us to gain a better understanding of the underlying drivers of the benefits described in Section 4. Unless stated otherwise in the text, the analysis in this section is based on Study 2.

5.1. Spillovers to the Nondigitized Set

As we pointed out in Section 4.1, even when we restrict our attention to users who never bought a product from the digitized set, the treatment condition seems to exhibit a more favorable behavior from the retailers’ perspective. This suggests the possibility that the virtual fitting technology may
generate spillovers affecting products beyond those that have been digitized and are available for virtual try-on. We now explore this further.

First, to formally confirm that there are indeed spillovers extending to products that were not digitized, we analyze how the number of products in an order is affected by the user being in the Metail or the control condition. To do this, we estimate a series of Poisson regressions, which we report in Table 9. Columns 1–2 simply regress the number of nondigitized and digitized products in the order on the Metail availability indicator. In Column 1 we observe that the Metail condition has a positive and significant effect on the number of nondigitized products; customers who were assigned to the Metail condition buy more products on average from the nondigitized set. Column 2 shows a positive but not significant effect on the number of digitized products (note there are fewer orders that include such products, which can explain the lack of significance).

Table 9 Spillovers

<table>
<thead>
<tr>
<th>Dependent Variable:</th>
<th>nndigitized</th>
<th>ndigitized</th>
<th>nndigitized</th>
<th>ndigitized</th>
<th>multicat</th>
<th>multicat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
</tr>
<tr>
<td>METAIL</td>
<td>0.009***</td>
<td>0.020</td>
<td>0.008**</td>
<td>-0.226***</td>
<td>0.054***</td>
<td>-0.170***</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.014)</td>
<td>(0.003)</td>
<td>(0.015)</td>
<td>(0.018)</td>
<td>(0.019)</td>
</tr>
<tr>
<td>ENGAGED</td>
<td></td>
<td></td>
<td></td>
<td>1.263***</td>
<td>1.275***</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.007)</td>
<td>(0.020)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.790***</td>
<td>-2.216***</td>
<td>0.790***</td>
<td>-2.216***</td>
<td>-2.641***</td>
<td>-2.641***</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.010)</td>
<td>(0.002)</td>
<td>(0.010)</td>
<td>(0.013)</td>
<td>(0.013)</td>
</tr>
</tbody>
</table>


Note: *p<0.1; **p<0.05; ***p<0.01

Unit of observation is an order placed in Study 2.
nndigitized is the number of digitized products in the order. nndigitized is the number of nondigitized products in the order.
multicat is 1 if the order includes multiple categories.

Columns 1-4 show results of a Poisson regression. Columns 5-6 show results of a logistic regression.

Exploring the impact of engaging with the Metail tool in the session in which the order is placed provides additional insight into the mechanisms in place. Columns 3-4 of Table 9 include the dummy variable indicating whether the user engaged with the tool. Engaging with the tool increases the number of digitized, but not nondigitized, products in the order. Note that only customers assigned to the Metail condition (METAIL=1) can engage with the tool (ENGAGED=1), so the coefficients in the two variables have to be jointly interpreted. Taking the results of columns 1–4 together, we see an interesting pattern: the Metail condition has a positive effect on the number of nondigitized
products overall and a positive effect on the number of digitized products for those customers who engage with the tool in that order.

Why do users from the Metail condition buy a higher number of products from the nondigitized set? A potential reason is an “outfit effect”: trying on a product may encourage users to buy complementary products. For example, a user trying on a pair of pants may decide to buy a top or a pair of shoes to go with it, which may or may not be from the digitized set. To test this effect, we explore how the presence of multiple product categories in an order differs for users in the treatment and control conditions. We define an indicator $multicat$ that is 1 if the order has more than one category and 0 otherwise, and we run a family of logistic regressions. The results are shown in Columns 5–6 of Table 9. Users in the Metail condition are more likely to place orders with multiple categories than users in the control group. The effect is mainly driven by users who engage with the tool, as seen in Column 6. Since the “outfit effect” may involve products that are not digitized, this can explain some of the increase in sales in the nondigitized set.

Another reason that may explain superior performance in the nondigitized set by users in the Metail condition is the effect of the tool on loyalty and revisit behavior. If the users in the Metail condition become more loyal to the firm and purchase more and more frequently, sales from customers in the Metail condition will further increase. Because most of the products are not digitized, this effect will be more noticeable in the nondigitized set. We provide further evidence of the tool’s effects on loyalty and purchase frequency in Section 5.2.

Finally, one more potential explanation for the effects in the nondigitized set is that the virtual fitting tool allows learning that generates spillovers to the products in the nondigitized set. A customer could use the Metail tool to try a (digitized) product and learn about the size that fits her best. The customer could later use that information about her size to buy another (nondigitized) product. Thus, the tool can have benefits in evaluating the customers choice set that extend beyond the digitized products. It could also happen that a customer who without the tool would have bought a product from the digitized set chooses to buy a different product from the nondigitized set when the tool is available because she does not like how the digitized product fits. This behavior can explain part of the reduction in return rates since those transactions would have been more likely to be returned. We provide further evidence on how the tool helps narrow the choice set in Section 5.3.

5.2. Loyalty and Repurchase Frequency

As mentioned above, the availability of the virtual fitting tool may result in an increased loyalty towards the firm. To test whether this is actually the case, we study repurchase behavior during the time span of our experiment. Since Study 2 tracks customers, we can generate a customer-level
variable, repurchase, that identifies customers who placed multiple orders during our experiment. To study how this variable is affected by the availability of the tool and engagement with it, we run a series of logistic regressions, where the unit of observation is a customer. Table 10 shows the results. Column 1 shows that the probability repurchase during the experiment is higher for users who were assigned to the Metail condition. Since the Metail condition was randomly assigned, we can conclude that availability of the virtual fitting tool increased repurchase behavior. Columns 2 and 3 consider the effect for customers who chose to engage with the tool at least once. The result suggests that the effects on repurchase arise mainly from users who actually engaged with the tool at some point.13

<table>
<thead>
<tr>
<th>Table 10</th>
<th>Repurchase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dependent Variable:</td>
</tr>
<tr>
<td></td>
<td>Repurchase logistic</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>METAIL</td>
<td>0.018***</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
</tr>
<tr>
<td>ENGAGED</td>
<td>1.205***</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
</tr>
<tr>
<td>Constant</td>
<td>−1.312***</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
</tr>
</tbody>
</table>

Observations: 2,389,655 2,389,655 2,389,655

Note: *p<0.1; **p<0.05; ***p<0.01
Unit of observation is a user that is part of the experiment.
All columns show logistic regression models.

Section 5.3 provides complementary evidence suggesting that availability of the virtual fitting tool also increases the average time spent on the site. Taken together, the fact that session length and purchase frequency increase suggests that the virtual fitting tool is an effective way to increase customer engagement with the online retailer.

13 Note that when we account for the ENGAGED variable, the sign of the METAIL coefficient becomes negative. Because customers can only have ENGAGED=1 if METAIL=1, we have to interpret those coefficients jointly. The fact that the sign of METAIL is negative means that customers with METAIL=1 and ENGAGED=0 are less likely to repurchase than are customers with METAIL=0. One explanation for this is that customers who are more positively inclined towards the brand (which is equally likely in both conditions) are more likely than customers who are not to engage with the tool. Therefore METAIL=1, ENGAGED=0 is a population that is less positively inclined towards the brand than the population for which METAIL=0.
5.3. Narrowing the Choice Set

Large choice sets can result in choice overload (Iyengar and Lepper 2000). We have suggested above that the virtual fitting tool may be helpful in narrowing the choice set. We now offer more formal evidence supporting this claim. First, we consider the task of purchasing a product in a specific size, narrowing the choice from the set of all available sizes for the product. This task is more complex for products that have a large number of available sizes. In Study 2, we have information about all the available sizes in the digitized set. If the virtual fitting tool is helpful in narrowing the customers choice, we would expect the tool’s availability to result in the more frequent choice of products with a high number of available sizes. We create a variable, MANYSIZES, that indicates whether a product has a broad size content. In what follows, we consider MANYSIZES = 1 for products that have four or more available sizes; we have experimented with different cuts with qualitatively similar results. Of the 26,237 products from the digitized set that were purchased during Study 2, 12,095 were purchased by customers in the control condition and 14,142 were purchased by customers in the treatment condition. We want to compare the fraction of those purchases corresponding to products with MANYSIZES = 1. This fraction is 29.6 % in the treatment condition versus 27.9 % in the control group. If we use Fisher’s exact test (see Hollander et al. 2013, Chapter 10), we can reject the hypothesis that the two proportions are the same, in favor of the alternative hypothesis that the fraction of purchases with MANYSIZES = 1 is higher in the treatment condition (p<0.01). Consequently, users assigned to the treatment condition are more likely to buy products with more available sizes. We conclude that the technology is particularly effective at helping users select between multiple competing options.

Study 1 and Study 2 did not collect rich session-level data. To complement our main studies, we conduct an additional test, Study 3, that allows us to further explore the differences in how customers parse their choice sets. The focus of this study is to obtain detailed results at the session-level data. Collaborating with Metail, we obtain data from a new A/B test run with a different retailer, based in the U.K., between June 1, 2016, and July 31, 2016. The structure of the experiment is similar to Study 2 (users were assigned to the Metail condition with p=0.5 and were maintained in the same condition in all their sessions). For each of the 479,138 sessions that we track, we are able to obtain two session-level variables: the duration of the session in minutes and the number of pages visited (which proxies the number of products viewed by the customer in the session). We conduct a regression analysis (OLS for the session duration, Poisson for the number of pages visited). The results are shown in Table 11. The results present an interesting pattern: customers in the treatment condition spend more time on the site on average, but view fewer pages.

\[\text{Note that because the treatment is assigned at random, it is not necessary to include additional control variables to obtain consistent estimates. However, if we include time fixed effects the results remain unchanged.}\]
(thus spending more time on average on each of the pages they visit). This is consistent with the idea that the tool is helpful in narrowing the choice set. Customers in the treatment condition are better able to filter their choice set and therefore can focus on a smaller number of products. For example, a customer who tries one product and does not like the fit can rule out similar products. At the same time, customers are more engaged with the retailer and end up spending more time on the site in each session.

Table 11 Effects on Session Length (from session-length A/B test)

<table>
<thead>
<tr>
<th>Dependent Variable:</th>
<th>session_length</th>
<th>pages_visited</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>OLS</strong></td>
<td><strong>Poisson</strong></td>
</tr>
<tr>
<td>METAIL</td>
<td>0.147***</td>
<td>-0.009***</td>
</tr>
<tr>
<td></td>
<td>(0.037)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>Constant</td>
<td>7.526***</td>
<td>1.253***</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>Observations</td>
<td>479,138</td>
<td>479,138</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01

Unit of observation is a session in Study 3.
Session_length is the duration of the session in minutes.
Pages_visited is the number of pages visited in the session.
Column 1 shows OLS, Column 2 shows Poisson regression.

5.4. The Impact of Size Recommendation

The tool provides two main benefits to customers: visualization and size recommendation. Most of our discussion has focused implicitly on the visualization aspect of the tool. We now study whether the size recommendation component has any significant impact on outcomes. To do that, we implement an additional A/B test (Study 4) that we conduct between December 1 and December 15, 2015, with a retailer based in India. Unlike the other studies described in this paper, all the users in this study have access to Metail. The treatment randomly assigned is not access to the Metail tool but how the size recommendation is presented to users. Users who entered the A/B test are assigned with equal probability to a treatment condition in which the size recommendation is more saliently presented or to the control condition in which size recommendation is not emphasized. Another key difference with the other studies is that we only have access to aggregate-level data from the retailer. We do not observe each transaction but instead look at the total revenues obtained each of the 15 days of the test for the treatment and control groups, for web and mobile
visitors. Thus, we have 15 days x 2 platforms (web/mobile)=30 cells, for each of which we have 2 observations: one treated (salient size recommendation) and one control. Given the small sample size, we use the Wilcoxon rank sum test, which is nonparametric and robust to violations of the normality assumptions of a t-test (see Hollander et al. 2013, Chapter 4). This analysis rejects the null hypothesis ($p<0.05$) that the median difference between pairs of observations is 0, in favor of the alternative hypothesis that the median difference between pairs of observations is positive (salient size recommendation has higher revenues).

Since we find that the salient size recommendation condition results in higher revenues than the control condition, we conclude that size recommendation plays a positive role on its own and that the results we described above do not arise exclusively from the visualization aspect of the tool.

6. Implications and Conclusions

6.1. Economic Implications
From the retailers’ perspective, it is important to understand the economic impact of virtual fit technology and whether it has a positive return on investment.

To address this question, we implement the following analysis, where we start with a base case in which there is no virtual fit technology. We consider that 1,000 customers visit the company’s website and 5.83% convert (this number is consistent with the data we observed in our retailer partner). We then assume an average order of US$120, cost of goods sold (COGS) of 30% of the revenue, shipping costs of 5% of the revenue, and a return rate of 20%. We then construct an alternative scenario using effects sizes consistent with the results from our analyses. Under the Metail scenario (where the virtual fit technology is available), the conversion is 6.41%. Compared to the base scenario, the average order is 1.6% larger and the return rate is 5.2% smaller. In this scenario we include a 1% increase in the COGS that capture the extra cost of implementing the virtual fitting technology. Table 12 shows the calculations obtained under these assumptions for both scenarios.

When the virtual fit technology is implemented, there is a positive net impact in profit of US$426.2, a 13.5% increase with respect to the base case. If we consider the total incremental benefit under the Metail scenario, 71.8% of it comes from a higher conversion rate, 11.6% comes from larger orders and 16.6% from lower returns.

Alternatively, it is possible to use a randomization test of no effect (Rosenbaum 2010, Chapter 2). Under the null hypothesis that the treatment has no effects, if one randomly reassigns the labels of the condition (switching some treatment and control cells), the distribution of outcomes should not change substantially. The test uses all the possible permutations of the treatment allocation to construct the distribution of outcomes. When the observed outcomes are extreme in terms of that distribution, one can reject the null of no effect. We can reject the null of no effect ($p<0.1$) in favor of the alternative hypothesis that the revenues in the salient size recommendation condition are higher.
We now look at how sensitive these results are under different base return-rate assumptions. This analysis is presented in Figure 3a. We observe that when the return rate doubles and reaches 40%, the incremental profit obtained under the Metail scenario increases from 13.5% to 19.6%. This analysis supports the intuition that virtual fitting technology works better when the return rates that the retailer experiences are higher.

Another relevant factor to consider is how the cost of the virtual technology affects the profitability of the Metail scenario. This analysis is presented in Figure 3b. We observe that even under high implementation costs (e.g., 5% of COGS), the profit increase is still positive and economically meaningful, at 10.6% in favor of the Metail scenario.

Lastly, we look at an alternative way to increase profits by increasing traffic to the retailer’s website. We implement a breakeven analysis where we find answers to the following question: What increase in traffic at the retailer’s website would make the base case obtain the same profit as the Metail scenario? The results of this analysis are presented in Figure 3c. This analysis shows that the traffic increase that would provide the same impact on profits as the implementation of the technology is above 18% for any reasonable range of implementation costs.
This series of analyses does not prove the viability of a virtual fitting tool. However, the economic results obtained—under a wide range of reasonable parameters and considering the impact observed in our main analysis—suggest that virtual fitting tools can be a promising investment.

6.2. Conclusions

By comparing the outcomes for customers assigned to the treatment and control groups, we have obtained evidence of a significant impact of virtual fitting tools on demand and fulfillment costs. On the demand side, we observe that customers with access to the tool are more likely to place an order. Those customers who place an order spend more money on the site, purchasing on average more and more expensive products. On the cost side, we find that fulfillment costs arising from returns and home try-on behavior are lower for customers in the treatment group. We have also explored the mechanisms through which virtual fit information helps both customers and retailers. The virtual fitting tool creates spillovers that extend to products not available for try-on, increases loyalty, helps customers better parse their choice sets, and reduces uncertainty by providing size recommendation.

The overall impact of virtual fit information on online apparel retailers and their supply chains is important. For apparel retailers, the information uncertainty online customers experience is one of the main barriers to growing their online business efficiently. Our analysis shows that virtual fitting technologies can be an effective way to deal with the information uncertainty challenge. When a virtual fitting tool is available, customers become more inclined to convert their browsing experience into a purchase, and that purchase is larger on average than when the virtual fitting technology is not available to them. Furthermore, the additional online fit information translates to a customer who is less likely to return purchased items or engage in home try-on behavior, reducing overall fulfillment costs.

By modeling how each clothing item fits a customer’s body and providing size recommendations, online retailers reduce the information friction in the last step of their supply chain and enhance their customers’ experience in the virtual store. We show that a virtual fitting room provides retailers with benefits similar to those provided by physical fitting rooms. Customers who have the opportunity to virtually try-on clothes when deciding to buy an item benefit from the additional information they receive, more efficiently parsing their choice set. In addition, as with the physical fitting room, a customer who receives a size recommendation or feedback on how things look on her will feel better served and more engaged with the retailer and hence more likely to buy in the future. While it is expected that the availability of a virtual fitting room will increase online sales and engagement, our analysis shows that the benefits of a virtual fitting room come from more than an increase in the engagement level. They also are driven by the fact that the virtual fitting room
effectively reduces information uncertainty, as evidenced by the reduction in returns we observed for customers who had access to the virtual fitting tool. Reducing returns is a key benefit of a brick-and-mortar fitting room. By offering a fitting room, retailers avoid two undesired behaviors: First, customers do not buy a product that does not match their needs, either because they do not like how the product fits or simply because it is the wrong size. Second, the retailer avoids “home try-ons,” i.e., a situation where customers “buy” multiple sizes of the same product and try them on at home. We have shown that virtual fitting technology helps reduce returns coming from these two sources.

The focus on sales growth is commonplace in retailing. However, the impact and relevance of returns can be easily overlooked. Returns are a very important issue for any retailer but are particularly impactful for apparel online retailers since they find themselves affected at multiple levels: net sales go down, customers returning items signals that something went wrong in their interaction with the company, and, more importantly, returns add an extra burden to the retailer’s supply chain. In the U.S., returns for online retailers have been reported to be close to 30 percent and these return rates are no different in other developed e-commerce markets, like the U.K. In less-developed online markets, such as the one in our empirical setting, return rates tend to be smaller, but analysts expect this situation will change as online retailers streamline and promote their return policies. The increasing return rates in online channels affect retailers’ supply chains in a new and expensive way. An online purchase that is returned will be shipped back to the retailer’s warehouse, where it needs to be inspected and reallocated to be assigned in a future order. The lag that lapses until the product can be sold again can be very long and result in some products losing their value (Guide et al. 2006). In other cases, the product is sold to a liquidator who pays only a small fraction of the retail value (Ng and Stevens 2015). Traditionally, apparel supply chains were designed to bring products to the customers in a fast and efficient way, but now retailers are facing the challenge of handling a reverse flow of nearly one-third of their orders. This adds the direct cost associated with shipping back the product—nowadays, most competitive retailers will offer “free” (to the customer) returns—and the cost of processing the return once it arrives at the warehouse. In addition, it adds the indirect cost associated with the complexity of handling a reverse flow equivalent to one third of the sales volume.

Our results provide strong evidence that retailers that incorporate virtual fitting technologies to reduce the level of information uncertainty can mitigate the challenges of selling apparel products

16 http://www.wsj.com/articles/SB10001424052702304773104579270260683155216
18 http://blogs.forrester.com/zia_daniell_wigder/14-10-21-key_metrics_in_brazils_ecommerce_market
19 Craig et al. (2014) discuss how complexity can drive execution quality issues that impact retailers’ performance.
online. We expect the evidence we provide with our analysis will motivate retailers to further explore and enhance their virtual fitting capabilities and obtain the dual benefit of sales growth and reduced operational costs of their online channels.
References


Craig, Nathan, Nicole DeHoratius, Yan Jiang, Diego Klabjan. 2014. Execution quality: An analysis of fulfillment errors in a retail distribution center. *Available at SSRN 1971500*.


Johnson Ferreira, Kris, Bin Hong Alex Lee, David Simchi-Levi. 2015. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management* Forthcoming(0) null.


Retana, German F., Chris Forman, D. J. Wu. 2015. Proactive customer education, customer retention, and demand for technology support: Evidence from a field experiment. Manufacturing & Service Operations Management Forthcoming(0) null.


Appendix. 1. Detailed data description.

A. Study 1.

Order level data. We have an entry for each of the orders placed during the test after a first visit to one of the pages displaying digitized garments.

- Order id: Retailer order id
- Order time: Time stamp for the date and time the order was placed.
- Condition: 1 if the session was assigned to the Metail condition, 0 if assigned to the control condition.
- Order value: Total amount spent in the order.
- Number of products: Total number of products included in the order.
- Average price: Average price of products included in the order, calculated as Order value/Number of products.
- Indicator of digitized set: 1 if the order contained any items from the digitized set, 0 otherwise.
- Order return status: 1 if one or more items in the order was returned to the retailer, 0 otherwise. This variable was collected at the end of the experiment. Because the retailer accepted returns during 30 days, the variable is only accurate for the first month of the experiment.

B. Study 2.

Customer A/B assignment data. We have an entry for each of the customers who clicked on one of the digitized garments and was assigned to one of the conditions, regardless of whether they purchased. We observe:

- Browser fingerprint
- IP address
- Condition: 1 if the customer is assigned to the Metail condition, 0 if assigned to the control condition.
- Order id: Retailer order id if that user made an order, null if not. If the user made multiple orders during the test, we see multiple entries. The orders may or may not contain digitized items.

Order level data. We have an entry for each of the orders placed during the test by customers who were assigned to one of the conditions after a first visit to one of the pages displaying digitized garments.

- Order id: Retailer order id
- Order time: Time stamp for the date and time the order was placed.
- Condition: 1 if the customer is assigned to the Metail condition, 0 if assigned to the control condition.
- Engagement indicator: 1 if the user used the Metail Technology in the corresponding order id, 0 if not
- Order level statistics (e.g., total order value, number of items, average price) is generated using the product level data

Item level data. We have an entry for each of the items included in orders placed during the test by customers who were assigned to one of the conditions after a first visit to one of the pages displaying digitized garments.

- Order id: Retailer order id for the order in which the item was purchased. This allows to inherit all the order level data described above.
• SKU id: Identifier of the item that was purchased
• Size: Size of the item that was purchased (e.g., L, M, etc.)
• Price: Price of the item that was purchased
• Quantity: Number of units of the item purchased.

**Product information.** For those SKUs that were digitized, we have the following additional information:

• Description: String that describes the product (e.g., “Saia Recorte Preta”)
• Available sizes (e.g., P, M, G, etc.)
• Category: Type of garment (e.g., tops, dresses, trousers, etc.)
Appendix. 2. Product-Level Analysis.

To confirm the robustness of our results, we run several additional analyses. There are two main goals to these analyses. First, we want to confirm that the results we find are not an artifact of some design decisions, such as the level at which we have conducted the main analysis. Second, we want to explore systematic differences in the effects observed. Since the products that were digitized were not chosen at random, it could be that the effects we find are very specific to those particular products. By focusing on different categories, we get a better sense of the potential generalizability of our main insights.

We conduct our robustness studies using detailed product-level sales, which we are able to track in Study 2. We conduct three main analyses. In the first analysis, we reproduce the analysis described in Section 4.1 but at the product level instead of the order level. In the second analysis, we analyze how sales evolve over time for customers with and without access to the fitting tool. Finally, we conduct a more granular analysis that considers sales in each of the available product categories.

For the first analysis, we break down the products sold during the trial to users who had access to the tool and those who did not have access to the tool. During the trial, 385,384 products were sold to customers who had been assigned to the treatment condition, while 315,050 products were sold to customers who had been assigned to the control condition. This suggests that sales for users with access to the tool were substantially higher overall (22.32 percent larger). If we restrict our attention to just products belonging to the digitized set, 14,142 of those products were sold to customers who had access to the tool, representing a 16.9 percent increase over the 12,095 products sold to customers who did not have access to the tool. We implement a binomial test (see Hollander et al. 2013, Chapter 2), to test whether share of sales that corresponds to the treatment condition exceeds 50% in a statistically significant way. The test allows us to reject the null in favor of the alternative hypothesis that the share of sales from the treatment condition exceeds 50% (p <0.01) and therefore we conclude that the positive effects of the tool documented in Section 4 persist even when we consider product-level sales.

Our second analysis studies how sales evolve over time for customers with and without access to the fitting tool. To do that, we compute total number of sold products per week for both the treatment and control groups. If the effects are robust, we would expect sales for the treatment group to be consistently larger than sales for the control group. Figure 4 plots the evolution of both groups, and we can observe that sales to customers in the treatment group (i.e., those customers who have access to the fitting tool) are systematically higher. Table 13, column 1 presents the results in a regression format, where the coefficient of the virtual fitting tool is positive and significant.

The aforementioned discussion considers the aggregate effects at the product level but does not consider the effects in the different product categories. Finally, we conduct a more granular analysis that considers sales in each of the available product categories. Table 14 shows the total number of units sold for products in the digitized set for both for treatment and control customers. If there was no effect in a category, we would expect a very similar share of treatment and control customers in the category. If the virtual fitting tool has an effect on sales in a given category, we would expect the share of treated customers among the purchased products of that category to exceed 50 percent. We formally test this using a binomial test (Hollander et al. 2013, Chapter 2), to test whether share of sales that corresponds to the treatment condition exceeds 50% in a statistically significant way. The test allows us to reject the null in favor of the alternative hypothesis that the share of sales from the treatment condition exceeds 50% (p <0.01) and therefore we conclude that the positive effects of the tool documented in Section 4 persist even when we consider product-level sales.

Our second analysis studies how sales evolve over time for customers with and without access to the fitting tool. To do that, we compute total number of sold products per week for both the treatment and control groups. If the effects are robust, we would expect sales for the treatment group to be consistently larger than sales for the control group. Figure 4 plots the evolution of both groups, and we can observe that sales to customers in the treatment group (i.e., those customers who have access to the fitting tool) are systematically higher. Table 13, column 1 presents the results in a regression format, where the coefficient of the virtual fitting tool is positive and significant.

The aforementioned discussion considers the aggregate effects at the product level but does not consider the effects in the different product categories. Finally, we conduct a more granular analysis that considers sales in each of the available product categories. Table 14 shows the total number of units sold for products in the digitized set for both for treatment and control customers. If there was no effect in a category, we would expect a very similar share of treatment and control customers in the category. If the virtual fitting tool has an effect on sales in a given category, we would expect the share of treated customers among the purchased products of that category to exceed 50 percent. We formally test this using a binomial test (Hollander et al. 2013, Chapter 2), to test whether share of sales that corresponds to the treatment condition exceeds 50% in a statistically significant way. The test allows us to reject the null in favor of the alternative hypothesis that the share of sales from the treatment condition exceeds 50% (p <0.01) and therefore we conclude that the positive effects of the tool documented in Section 4 persist even when we consider product-level sales.
Table 13  Effect of Virtual Fitting Tool on Sales Over Time

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log(Sales)</td>
<td></td>
</tr>
<tr>
<td>METAIL</td>
<td>0.205***</td>
</tr>
<tr>
<td>(0.015)</td>
<td></td>
</tr>
</tbody>
</table>

Time effects   Week
Observations   20
R-squared      0.998

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 4  Sales During the Test Weeks

2013, Chapter 2) for each category. The test allows us to reject the null for all categories (p<0.05). This suggests that while the precise effects can be slightly different for different products, the benefits of virtual fit information are likely to be materialized for a wide range of products.

Table 14  Sales by Category (Digitized Set)

<table>
<thead>
<tr>
<th></th>
<th>Dresses</th>
<th>Jackets</th>
<th>Knitwear</th>
<th>Shoes</th>
<th>Shorts</th>
<th>Skirts</th>
<th>Tops</th>
<th>Trousers</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control</strong></td>
<td>4,262</td>
<td>1,532</td>
<td>180</td>
<td>334</td>
<td>393</td>
<td>510</td>
<td>3,724</td>
<td>1,160</td>
<td>12,095</td>
</tr>
<tr>
<td></td>
<td>47.38</td>
<td>45.39</td>
<td>41.96</td>
<td>46.98</td>
<td>42.72</td>
<td>42.82</td>
<td>45.82</td>
<td>46.62</td>
<td>46.10</td>
</tr>
<tr>
<td><strong>Treatment</strong></td>
<td>4,734</td>
<td>1,843</td>
<td>249</td>
<td>377</td>
<td>527</td>
<td>681</td>
<td>4,403</td>
<td>1,328</td>
<td>14,142</td>
</tr>
<tr>
<td></td>
<td>52.62</td>
<td>54.61</td>
<td>58.04</td>
<td>53.02</td>
<td>57.28</td>
<td>57.18</td>
<td>54.18</td>
<td>53.38</td>
<td>53.90</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>8,996</td>
<td>3,375</td>
<td>429</td>
<td>711</td>
<td>920</td>
<td>1,191</td>
<td>8,127</td>
<td>2,488</td>
<td>26,237</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Appendix. 3. Additional Tables.

To make sure the weights are actually correcting for observable differences in treated- and control-unit characteristics, we follow Guo and Fraser (2009) and we run weighted regressions of the continuous covariates in the treatment. If the propensity score weighting effectively removes imbalances, the coefficient of the treatment variable should not be significant. The following table (Table 15) shows it is the case.

<table>
<thead>
<tr>
<th>Table 15</th>
<th>Evidence that Weighting Removes Imbalances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dependent Variable:</td>
</tr>
<tr>
<td></td>
<td>order_amount</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>METAIL</td>
<td>0.305</td>
</tr>
<tr>
<td></td>
<td>(0.921)</td>
</tr>
<tr>
<td>Constant</td>
<td>187.324***</td>
</tr>
<tr>
<td></td>
<td>(0.651)</td>
</tr>
<tr>
<td>Observations</td>
<td>211,116</td>
</tr>
<tr>
<td>R²</td>
<td>0.00000</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>-0.00000</td>
</tr>
<tr>
<td>Residual Std. Error (df = 211114)</td>
<td>299.098</td>
</tr>
<tr>
<td>F Statistic (df = 1; 211114)</td>
<td>0.110</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01