The Causal Impact of Fit Valence and Fit Reference on Online Product Returns

Completed Research Paper

Yang Wang
University of Utah
Salt Lake City, UT, 84112
yang.wang@eccles.utah.edu

Vandana Ramachandran
University of Utah
Salt Lake City, UT, 84112
vandana.ramachandran@eccles.utah.edu

Olivia Sheng
University of Utah
Salt Lake City, UT, 84112
olivia.sheng@eccles.utah.edu

Abstract

We investigate the causal impact of two types of product fit-related information – fit valence and fit reference – on online product return rate by leveraging a change in the product review system that took place at an online retailer. This quasi-experiment in the apparel product category allows us to examine the importance of fit information. We find that the mere presentation of either fit-valence (e.g. "true to size") or fit-reference information (e.g. body size) by itself does not help reduce purchase errors. Rather, it is the combination of the two types of fit information in a review that drives the drop in product return rate. We employ the lens of semantic relativism to illustrate how customers interpret fit-valence expressions by using the fit-reference information provided by the same reviewer. Our findings offer useful business implications to online retailers grappling with high product return rates for merchandise where fit matters.

Keywords: Quasi-experiment, product fit opinions, fit reference, online product returns

Introduction

According to eMarketer (Jackson 2015), the apparel category accounted for 17.6% of all online sales in 2013, and its total sales amount is projected to reach nearly $86 billion in 2018, second only to the sales of consumer electronics. Yet troublingly, the online apparel return rate is also substantially high: averaging around 30%-40%, and running as high as 50% for some retailers (The Economist 2013). Recent industry studies (e.g. Trueship.com 2016) report that, while poor quality and items being shipped incorrectly lead consumers to return products, the top reason for online apparel returns is the lack of product fit.

Fit describes how well a product suits a consumer's preferences. For clothing, fit may encompass size, cut and shape; for shoes, it may refer to width, arch support and flexibility; for hotels, it may include noise levels of rooms and proximity to resources, to name a few. For products purchased online, fit poses unique informational challenges that are different from quality. In many product categories where non-digital attributes are important (Lal and Sarvary 2009), fit is difficult to assess prior to purchase, and can usually only be determined upon physical product inspection. This is because determining product fit depends crucially on the highly subjective matching of one's unique preferences to a product's horizontal attributes[1].

Moreover, due to its idiosyncratic nature, information about fit found in online reviews left by previous customers need not be meaningful for other consumers. In contrast to quality attributes, where consumers generally tend to agree on what is good versus poor, their opinions on fit are relatively more divergent[2]. For instance, an item of clothing that suits a taller (or athletic) customer may not be as dapper on a shorter (or stockier) customer. A pair of shoes that fits a runner with high arches or narrow feet may not fit a customer with different foot shape, even if the shoes are made well. Cruise ship rooms proximal to elevators offer different value than rooms close to the recreation centers, and consumers have heterogeneous valuations for these attributes. In all of these examples, customers may agree in their online reviews on the level of product quality, but disagree on whether the product fits their idiosyncratic preferences. Thus, when it comes to fit, one man's treasure may be another's trash.

More importantly, when fit is an important driver of a product's utility, ill-fitting products may be of little use to consumers even if the quality were acceptable or better. A high quality jacket or shoe that does not "fit" provides little value to customers, and leads to it being returned. In realization of the unique challenge posed by fit uncertainty online, online retailers are increasingly seeking to implement new solutions to ease the challenge of lack of fit information online. One approach has been to introduce fit-related product review functions. For example, Zappos.com and Nordstrom.com request their customers to rate the item's fit on a 5-point scale from runs small (narrow) to runs large (wide) in addition to the overall rating; Urban Outfitters encourages reviewers to use a graphic scale to indicate their size evaluation as ranging from small to true-to-size to large. Hotels.com allows customers to rate the hotel in terms of its location and service in addition to an overall score. We refer to consumers' subjective evaluations of the fit of a product's horizontal attributes as *fit-valence expression*.

However, one customer's fit evaluation may not be meaningful to another without additional referential information that allows the latter to determine the likeness of the former's preferences to their own. In the case of clothing and shoes, this reference information refers to the reviewers' body type; in the case of hotels or cruises, the trip goals of the traveler, etc. We refer to this as *fit-reference* information, which is defined as a customer's self-description or self-categorization of their fit-relevant preferences. Interestingly, online retailers vary in which of the two types of fit-related information they implement on their website for product categories that are sensitive to fit. Some retailers' online reviews systems allow customers to systematically provide either fit-valence and/or fit-reference information. Reviewers may also choose to report this information embedded in their review text itself. This heterogeneity in the availability of fit-

[1] Product variety or differentiation can occur along three dimensions defined by the way a consumer's evaluation of a product changes as a function of changes in an attribute (Ulrich 2011): *fit* where consumers' preferences have an ideal point (or single strong peak) and deviations from it lower the product's value precipitously; *taste* where consumers have multimodal preferences and the value function has peaks and valleys – and is less sharply defined than fit attributes; and finally, *quality* where consumers prefer more of the positive and less of the negative attributes, controlling for price. In this study, we focus on fit variety.

[2] We follow Lancaster (1971, 1979) in economics literature in distinguishing the product space into horizontal attributes, where preferences vary across consumers and results in a distribution (variety) of ideal characteristics, and vertical attributes, where consumers agree that more of that characteristic is always better, but differ in their willingness to pay for more. Economic models of product differentiation commonly use this notion to model fit as a horizontal attribute and quality as a vertical attribute (see Economides 1989; Neven and Thisse 1990; Bohlmann et al. 2002). Horizontal would include Ulrich's fit and taste varieties, and vertical refers to quality variety.

related information offers us as a testing ground to study the impacts of different combinations of fit-valence and fit-reference information on consumer's product return behaviors. We do not know whether and how consumers use the new fit information made available to them, especially given that they have varying product fit preferences. Specifically, is it the fit-valence or fit-reference information itself or the combination of them that helps customers select the right product?

Our goal in this study is to examine the effectiveness of such newly-added product fit review functions, and whether they are able to reduce incorrect purchase decisions (those culminating in products being returned) made by online consumers. In particular, we leverage an online review system change in the apparel product category – a treatment - that took place at a leading online e-commerce retailer in the outdoor goods industry in the United States to examine the causal effects of providing fit-related information on product returns. This setting allows us to specifically tease apart the value of the two types of fit information using the theoretical lens of *semantic value relativism* in linguistics and philosophy of language. This lens when applied to *predicates of personal taste* (such as fit) gives us a framework to help understand how consumers may accurately extract value from a piece of subjective information (such as fit valence) when appropriate knowledge about the circumstance of evaluation (or referential knowledge about fit) is available.

We discovered that by itself, fit reference does not produce a significant drop in product return rate. Nor does the mere existence of fit valence. Rather, it is the combination of the two types of fit information that drives the decrease in product return rate. We estimate that an additional 10% increase in the availability of fit-valence and fit-reference information in reviews leads to a 1.6% drop in apparel return rates for our retailer. The fit-reference information provides the circumstance of evaluation that helps customers interpret the fit-valence expression, i.e. customers can make meaningful inferences from the fit-valence opinions provided by other customers by comparing the level of match between their own and other's preferences. In the context of apparel purchase in our study, we find that customers are able to make more accurate purchase decisions based on the size of the product, and thus subsequently return fewer products. Also interestingly, but not so surprisingly, we find that for apparel products, the impact of overall rating measures on product return rate is less important than that of product fit opinions, further highlighting the relevance of fit information in online markets.

Our study makes several contributions to the literature of online product reviews and product fit uncertainty. While the impact of quality opinions has received much attention in the academic literature on online reviews, the impact of fit opinions is comparatively less understood, and we attempt to fill this gap. By leveraging a quasi-experiment, we show that two types of fit opinions in online reviews have unique causal impacts on outcomes different from those of quality opinions. We apply a theory of the philosophy of language to help guide us in our understanding of how others' fit information acquires meaning for a consumer, and in doing so shed light on the plausible underlying mechanism. Our results are useful for online retailers that struggle with returns caused by fit uncertainty. Based on our findings, we recommend retailers to implement both fit-valence and fit-reference information in their online review systems, and more importantly, encourage or incentivize customers to provide both types fit opinions.

The rest of the paper is organized as follows: we review related literature and present our theoretical bases in §2. In §3, we describe the setting of our quasi-experiment and define the treatment. We discuss the data and empirical identification strategy in §4. §5 reports the main estimation, while §6 shows several robustness checks that explicitly deal with potential threats to internal validity. We conclude in §7 by offering several business implications and also discuss some future directions of our work.

Related Literature and Theory

Three streams of work are relevant to our work: the expansive literature on the economic impacts of online product reviews, the emerging literature on fit uncertainty in online markets, and the literature on the theory of semantic relativism.

Economic Impact of Online Product Reviews

Much of the prior work in the domain of online product reviews focuses on the economic impact of product reviews on sales (e.g. Chevalier and Mayzlin 2006, Liu 2006, Duan et al.2008, Chintagunta et al. 2010, Archak et al. 2011, Gopinath et al.2014; Rosario et al. 2016). Product return is nevertheless an important aspect of the sales cycle. Returns have an economic effect on net profit (Guide et al. 2006, De et al. 2013), and are reflective of consumers' immediate post-purchase satisfaction (Kopalle and Lehmann 1995, Hong and Pavlou 2014), and may also affect consumers' long-term relationships with a retailer (Petersen and Kumar 2010). While the role of products return has received much attention in several streams of business literature, only recently have researchers begun to examine the effect of online product reviews on product returns (Sahoo et al. 2015). We add to this stream of work by studying how online product opinions affect one particular type of product return – that primarily driven by lack of (or poor) fit – and whether introduction of new types of fit information in online review system can help consumers choose better suited (fitting) products and thereby, reduce purchase errors.

Product Fit Uncertainty

Long recognized as an important driver of firms' positioning and product differentiation in the theoretical economics literature (Economides 1989; Neven and Thisse 1990; Bohlmann et al. 2002), fit uncertainty has recently received attention in studies of online markets. Kwark et al. (2014) analytically examines how quality and fit information in online reviews impact interactions between upstream and downstream partners. Directly relevant for our study is the small but growing empirical literature on the role of product fit uncertainty in online consumers' decision-making. Hong and Pavlou (2014) theorize and measure the construct of product fit uncertainty using surveys research method, and find that fit uncertainty is distinct from quality uncertainty, and the former has stronger positive effects on online product returns. Sahoo et al. (2015) study the impact of product rating and return costs on returns through their effects on consumers' uncertainty about expected utility. Their study implicitly treats rating as a reflection of product quality – a notion that is commonplace as demonstrated in a recent meta-analysis of this literature (Rosario et al. 2016). However, researchers also recognize that a single overall rating conflates consumers' evaluations about quality and price (e.g., Li and Hitt 2010), and quality and fit (e.g., Sun 2012) – the latter of which is of concern in this study. For products where fit uncertainty plays a critical role, online reviews may be based on consumers' fit assessments rather than or in addition to their quality assessments. Thus, it is important to separately examine the role of fit evaluations in online reviews.

In order to mitigate the unique challenges posed by fit uncertainty in product categories where non-digital attributes are important, online retailers have implemented a variety of solutions. A handful of research studies examine the effects of specific web technologies directed at reducing fit uncertainty, such as zoom functions (De et al. 2013) and virtual fitting rooms (Gallino and Moreno 2015), and find that they may lower returns. Another solution is offered by a multi-dimensional rating system – and researchers have found that when allowed to provide multiple ratings, consumer rate differently than when they can only provide a single rating (Liu et al. 2014). Yet, the effects of multi-dimensional rating system on product sales and returns is not well understood. Shulman et al. (2015) interestingly finds that provision of additional information meant to reduce fit uncertainty, but which does not fully do so, can actually increase purchase decision reversals or returns. These varying empirical effects of fit information motivate our study.

We focus on examining the impact of allowing consumers to separately report fit-related evaluations, in addition to overall rating scores. Product fit is, however, less easily understood from other consumers' online reviews than product quality. While customers tend to value high over low quality (at the same price), they have different preferences for horizontal attributes and may not share common criteria in judging product fit (Hong and Pavlou 2014). Thus, the valence of product fit opinions from past consumers, by itself, may not be very meaningful to future customers. This raises the next question – what type of fit information is useful to online customers? Taking advantage of a quasi-experiment in the field where our focal retailer introduced a change in the review system to allow its customers to report fit-reference information (i.e. information about the size of product purchased and a consumer's body measurements) - we are the able to study the mechanism by which fit information causally influences product returns.

Semantics and Online Product Opinions

The vast volume and variety of online reviews offers researchers a powerful new tool for studying consumers' experiences with products, and to better understand why they rate as they do. Analyses of textual narratives left by consumers has delivered key insights above and beyond that provided by numerical rating scores in studies across several disciplines. This growing area uses techniques at the intersection of computer science and linguistics and has been successfully applied to understand the characteristics of online reviews that are not only helpful (e.g., Ghose and Ipeirotis 2011; Mudambi and Schuff 2010) but also influential on sales (Ludwig et al. 2013 *inter alia*). In addition to popularly applied sentiment analysis techniques, researchers have shown that semantics play an important role in making sense of expressions found in online reviews (Turney 2002, Dave et a. 2003, Liu et al. 2005). In recent work, Qi et al. (2015) use a combination of sentiment mining and semantic ontology to build a knowledge base from online reviews that can be reused for decision making. In similar vein, we apply the lens of semantics to decipher how fit expressions in online product reviews are used for decision-making by online shoppers.

Semantic Relativism and Fit

Fit-valence expressions are provided by consumers who have purchased a product and consumed by subsequent online shoppers. In this manner, each customer processes expressions containing assertions about fit (that are sometimes divergent) made by other consumers. For example, one consumer may write that an apparel product fits true to size whereas another may report it to be oversized, relative to the size that they have bought. How then should a potential consumer make sense of these fit valence expressions?

Literature in linguistics and philosophy of language related to predicates of personal taste (PPT) provides us with a useful framework for doing so (Lasersohn 2005, Stephenson 2007). PPT express matters of subjectivity or taste such as tasty, fun, enjoyable, which are personal or idiosyncratic, unlike objective facts such as height and weight of a product (Lasersohn 2005). Traditional semantics literature holds that the truth of a proposition at most varies with possible worlds and time (Kaplan 1989) but is otherwise absolute. This has been found to be insufficient to explain the truth value of expressions containing PPT ("this food is tasty" "rollercoasters are fun"), especially where two individuals could express opposing sentences about the same product and yet, both be true (Stephenson 2007). Contemporary semantic linguistic literature proposes that subjective expressions that contain PPT cannot be judged absolutely as true or false by themselves, and need an additional factor in order to evaluate their truth value. To the extent that the fit attributes that we discuss in this study are taste-dependent horizontal attributes of a product, and therefore subjective and capable of producing disagreeable expressions in online reviews, we believe that the semantics of predicates of personal taste would be applicable in our context.

In contemporary semantics literature, relativist theories hold that the truth value of some expressions are circumstance-sensitive, where the *circumstance of evaluation* is defined as all parameters used to evaluate an expression (Macfarlane 2005, 2007, Kolbel 2008). Traditionally, these parameters consisted of world and time (Kaplan 1989)[3]. In his work on assessing the accuracy of statements that contain PPT expressions, Lasersohn (2005) introduced the notion of a "judge" or assessor as an additional parameter in the *circumstance of evaluation*. This judge refers to the author of the expression, for whom (or relative to whose standard of taste) the PPT expression is true. Consider that in the context of online reviews, reviewers often make assertions about product fit ("this shirt is too big", "these pants are true-to-size"), which should be, according to relativism, interpreted as being true relative to the taste of the author (judge or assessor) of the review, but not necessarily for others. A reader of these reviews should then correctly interpret the fit-valence expression to mean that "this shirt is too big" *for the author* or "these shoes are true-to-size" *for the author*[4]. Therefore, the accurate interpretation of a fit valence expression depends on or is relative to the author or judge of the expression (MacFarlane 2007).

Now we go back to the question of how a potential customer of a product faced with multiple (sometimes disagreeing) fit-valence expressions from previous consumers should interpret that information for accurate decision-making. It follows from the above discussion that an expression of fit valence cannot be meaningful on its own. Rather, only in the presence of information about the circumstance of evaluation (provided by characteristics of the judge in addition to world and time) may a potential customer be able to assess the relevance of the fit valence expression for his or her own purpose. For example, knowing that a medium-sized jacket runs large for a prior consumer (judge) with a specific height and weight (circumstance of evaluation) will allow the potential customer to assess how well the jacket might fit herself by comparing her own height and weight with the circumstance of evaluation (the weight and height and size (purchased) of reviewer) of a review. Whereas in the absence of information about the circumstance of evaluation, a

[3] Subjective expressions may hold relative to a particular norm of world (cultural, geopolitical, social, etc) and time period.

[4] This view is autocentric, but Lasersohn (2005) also allows for an exocentric view where the PPT may be evaluated relative to an appropriate referent other (e.g. "the shirt is too big for her").

potential customer would not be able to accurately make sense of an expression such as “runs big” as this would leave several unanswered questions – runs big for what kind of body and product size?

For products where fit plays a critical role, consumers may suffer from fit uncertainty online, or not being able to assess whether a product matches his or her preference (Hong and Pavlou 2014). In the presence of fit-reference information – which in our study is given by the reviewer’s body size and the product size purchased – customers that are shopping online can better evaluate the suitability of fit-valence expression found in online reviews. Consequently, the presence of both types of fit information can lower online consumers’ fit uncertainty by offering them tools to help make better decisions. We therefore expect that it is not the mere presence of valence information, rather the combination of fit valence and fit-reference information that would reduce the product return rate. We illustrate our research framework in Figure 1.

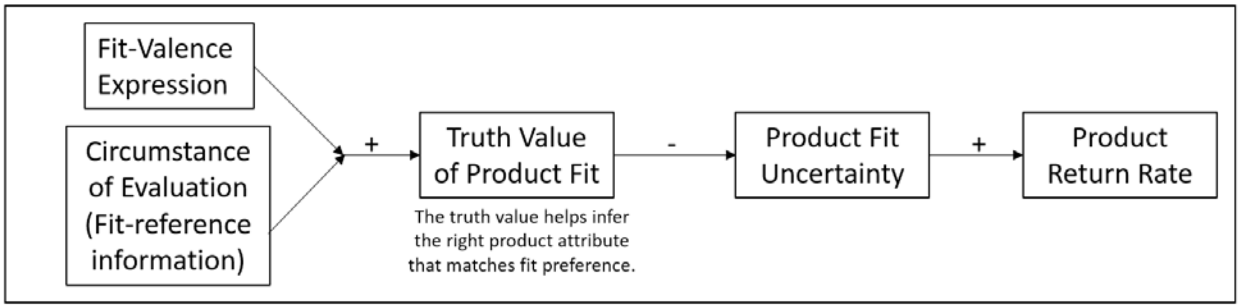
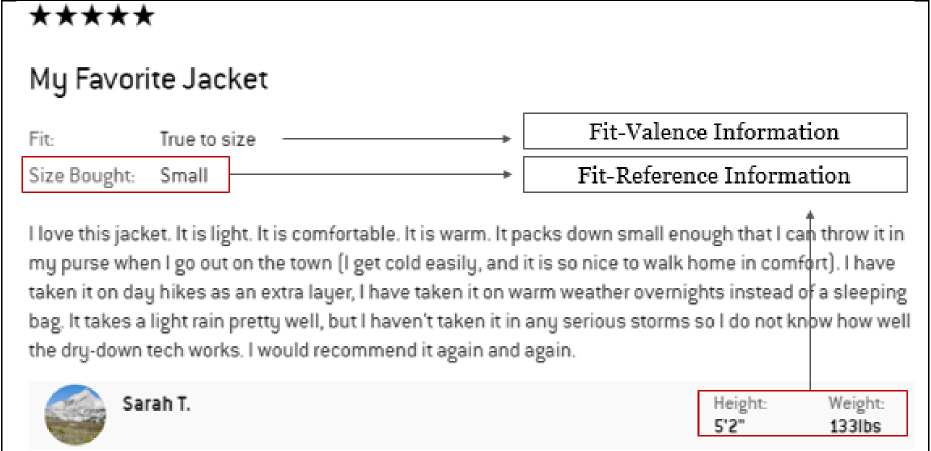

Fit-Valence Expression
Circumstance of Evaluation (Fit-reference information)
+
Truth Value of Product Fit
-
Product Fit Uncertainty
+
Product Return Rate
The truth value helps infer the right product attribute that matches fit preference.

Figure 1. Impact of Semantic Relativism and Fit Information on Return Rate

Background and Treatment

Natural or quasi-experiments allow us to examine the effects of a single treatment or change in a system that researchers would find difficult to otherwise manipulate in the field (Shadish et al. 2002). Although the literature of online product reviews has evolved for more than a decade, the number of natural experiment studies is limited. Chen et al. (2011) is one such paper that exploits an online review system change implemented by Amazon.com in 2009 when it published information about what customers ultimately purchased after looking at a focal product, thereby allowing the authors to examine the effect of observational learning and online word of mouth on sales rank. Liu et al. (2014) leverage a natural experiment to investigate how a newly adopted multi-dimensional rating system on TripAdvisor.com affects consumers’ rating behaviors. We describe the background of our natural experiment next.



★★★★★

My Favorite Jacket

Fit: True to size

Size Bought: Small

I love this jacket. It is light. It is comfortable. It is warm. It packs down small enough that I can throw it in my purse when I go out on the town (I get cold easily, and it is so nice to walk home in comfort). I have taken it on day hikes as an extra layer, I have taken it on warm weather overnights instead of a sleeping bag. It takes a light rain pretty well, but I haven't taken it in any serious storms so I do not know how well the dry-down tech works. I would recommend it again and again.

Sarah T.

Height: 5'2" Weight: 133lbs

Figure 2. Fit-related Information in an Actual Online Review (The newly-added fit-reference information is shown in red)

In this paper, we study the causal impact of a change in the review system at our online retail partner’s website - that provided fit-reference information to consumers – on product return rate. The setting for our study is one of the largest online-only specialty retailer of premium outdoor goods located in western USA. This retailer sells products in several categories; in 2015, it carried more than 50,000 products from over 900 brands. Our focus is the apparel category (jackets, shirts, pants etc.), whose products are especially vulnerable to fit-related pre-purchase uncertainty, and therefore provides an apt context for us to examine our research question. The retailer introduced a change in the online review system on February 18, 2015 – which is the treatment of interest to us. In addition to the overall rating and textual review on post-purchase product experiences, reviewers could now reveal information on their body size (i.e. height and weight) and the size of product that they have purchased - which we refer to as *fit-reference information*. Prior to this, in June, 2013, the retailer had implemented an earlier review system change that allowed reviewers to provide a summary evaluation of their overall fit experience with the product that they purchased. This *fit-valence* expression ranges from *runs-small* to *true-to-fit* to *runs-large*. Figure 2 illustrates the two types of fit information using an actual consumer review.

This innovation in the review system diffuses over time, creating variation in treatment levels both across products and within a product across time, as illustrated in Figure 3. Because it is not mandatory for customers to disclose fit-reference information after the change, some products receive new reviews with fit-reference information sooner than others, and not all new reviews for a product use the function. This is similar to Sun and Zhu's (2015) study where bloggers on a Chinese portal participate in an ad-revenue sharing program (the treatment) at different times after its introduction, and Goldfarb and Tucker's (2015) study where the actual usage of standardized ads occurred over time after a new standard for banner ads (the treatment) was implemented in the industry. We illustrate this diffusion of the new reviews containing fit-reference information in Figure 3 for 4,605 unique apparel products that have at least one new review generated following the introduction of the fit-reference function. By the end of September, 2015 – when our data collection ended, there were a total of 8,344 new reviews posted, out of which 6,006 adopted the new fit-reference function, and were spread across 3,616 unique products.

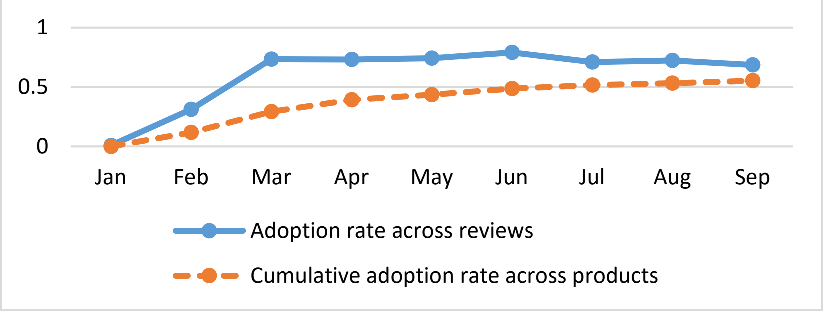


Figure 3. Monthly Adoption Rate of Fit-Reference Function (Year 2015)

In Figure 3, the solid line depicts the monthly adoption rate of the fit-reference function across all newly arriving apparel reviews. It is calculated as the percentage of newly generated reviews in each month that adopts the fit-reference function. As observed in the graph, the new review function became widely-adopted very quickly: since March, 2015, the adoption rate across incoming apparel reviews has remained rather stable around its mean value of 71.98%. The dashed line shows the cumulative adoption of the new fit-reference function across apparel products, and is increasing steadily over time. About 55% of the products available for sale in 2015 had received reviews that use the fit-reference function by September 2015. This distribution across products alleviates concern that the newly-posted fit-reference reviews are confined to certain types of products that may not be representative of the apparel category.

We then identify two groups of products after the review system update– the treatment group which includes products with fit-reference reviews, and the comparison group of products without fit-reference reviews. Recall that the membership in both groups was varying across time as treatment diffuses. This unique setting allows each treated product to serve as its own counterfactual in time periods when it does not receive treatment reviews. Such a setting makes it feasible to examine the distinct impact of product fit expressions with fit-reference information versus those without fit-reference information on product return rates. The treatment is implemented uniformly across all apparel products, and no incentives were provided by the retailer to customers to use the new function. Thus, reviews that utilize the new fit-reference function appear to arrive randomly across these products, and we will test this in detail in the next section.

Data

To conduct our analyses, we focus on apparel products sold by the retailer during a two-year time range between September 18, 2013 to September 18, 2015 that received reviews both before and after the system update, which results in 1,052 products. In order to aid in the calculation of return rates, we further drop products that do not have sales both before and after the treatment, which filters out products that are only released after the system update or eliminated after it. This ensures that our sample contains products that are actively selling before and after introduction of the treatment. We also do not include orders placed without the product page visit information, suggesting that the consumer may not have read product reviews prior to purchase. The final sample contains 942 products[5], among which 696 of them receive fit-reference information in the very first review after the system update, while the remaining 246 products serve as the comparison group.

Measures

The unit of analysis in our study is a unique apparel product. We define two variables to measure fit-related information available for a product: percentage of reviews that contain fit valence (*Fit_val%*) and percentage of reviews that contain fit reference (*Fit_ref%*) information. Fit_ref% measures the main treatment effect - the level of newly-generated fit-reference information. We apply percentage measures for review information variables based on the belief that the impact of the new review content is marginal as the review volume increases, similar to the percentage of positive/negative ratings used in Chevalier and Mayzlin (2006), Liu (2006), and Chen et al. (2011).

Further, to help account for conditions where fit-reference information occurs alone versus when it is combined with fit-valence expressions in a new review, we develop a set of three mutually exclusive measures: percentage of reviews that only contain fit-valence expressions (*Only_fit_val%*), percentage of reviews that only contain fit-reference information (*Only_fit_ref%*), and percentage of reviews that contain both types of fit information (*Fit_comb%*). Note that it is not suitable to use the overall interaction between fit valence and fit reference as the measure of combined fit information, because only when these two types of fit information are provided by the same reviewer is it meaningful for subsequent customers.

Table 1. List of Variables in the Main Analysis

Variables	**Explanation**
Return_rate	return rate of products purchased in the time period
Fit_val%	percentage of reviews that contain fit-valence information
Fit_ref%	percentage of reviews that contain fit-reference information
Only_fit_val%	percentage of reviews that only contain fit-valence information
Only_fit_ref%	percentage of reviews that only contain fit-reference information
Fit_comb%	percentage of reviews that contain both fit-valence and fit-reference information
Avg_fit_val	average fit valence
Avg_rating	average rating
Rev_volume	number of reviews (in natural log)
Avg_rev_length	average number of words in each review (in natural log)
Avg_price	average prices (in natural log)
Avg_disc	average product discount
Length_period	Inter-arrival time between two reviews in days (natural log)
Holidays	number of holidays during the period
Avg_Age	average number of days between the first sales date and the date of each order (natural log)

We included a number of pertinent control variables to account for any observable differences among treatment and comparison products that may affect return rates. We measure the average fit valence (*Avg_fit_val*), which is calculated as percentage of positive fit valence (i.e. "true to size") minus percentage of negative fit valence (i.e. "runs large" or "runs small"). Also, we followed the extant literature (e.g. Chevalier and Mayzlin 2006, Chen et al. 2011, Archak et al. 2011, Sahoo et al. 2015 etc.) and controlled for

[5] In unreported analyses, we find no significant differences across product review measures between the 942 included products and the 110 products that do not have sales information either before or after the system update. Hence, sample attrition is not a problem.

basic rating features, i.e. average rating (*Avg_rating*), number of reviews (*Rev_volume*), and average number of words in each review (*Rev_length*). Other control variables include average price (*Avg_price*), average product discount information (*Avg_disc*), number of days during the period as explained later (*Length_period*), and number of holidays (i.e. *Holidays*, which includes Thanksgiving Day, Christmas, and Valentine's Day) during the period. Similar to past studies (e.g. Chevalier and Mayzlin 2006, Chen et al. 2011, Sun 2012 etc.), natural logarithm is taken for price, review volume, and average review length. Table 1 describes all variables used in the main estimation.

Model-Free Evidence

Before proceeding to the formal analysis, we present some model-free evidence on the impact of treatment reviews on return rates. In Figure 4A, we sort the products into three groups that are mutually exclusive in each month. a) The leftmost blue bar with solid color in each bar cluster shows the return rate for products that do not receive fit-reference reviews during the period for analysis – they are *never treated* (they are always in the comparison group). Since the treatment in our context is introduced dynamically across products, we use two more bars to represent eventual participants consisting of b) *already treated* products that receive treatment before or by the focal month (the middle orange bar with horizontal stripes in each bar cluster) and c) *not yet treated* products that have not received treatment by that month (the rightmost grey bar with vertical stripes in each bar cluster).

The key observations from Figure 4A are: a) after accounting for the time trend, the average return rate for products already in the treatment group (orange bar) is the lowest among the three groups across all months, and b) the return rates for comparison products that are never treated (blue bar) and eventual participants that are not yet treated (grey bar), are closer to each other and higher than the return rate for the already treated products. We observe that prior to treatment, products show no difference in outcomes.

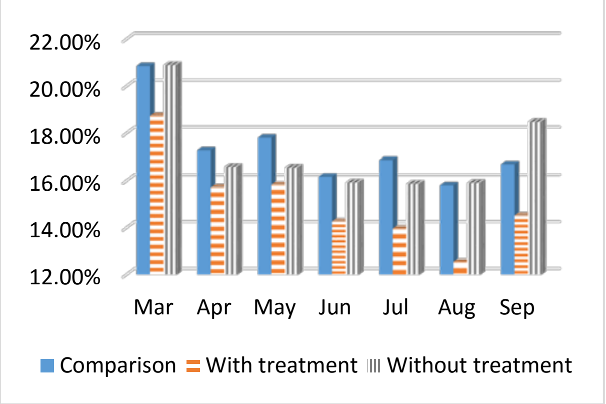


Figure 4A. Across-Group Return Rate (Year 2015)

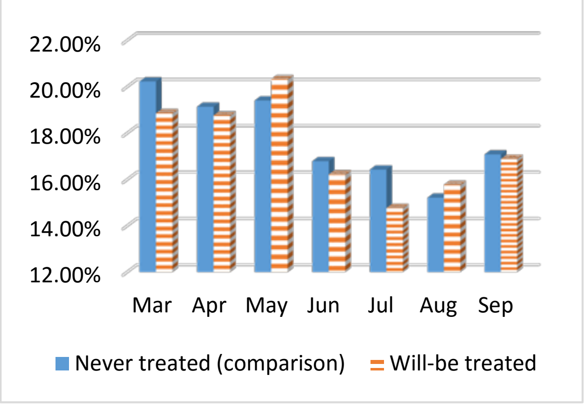


Figure 4B. Across-Group Return Rate (Year 2014)

Figure 4B plots the monthly product return rate for the same products in the previous year 2014. Note that none of the products received treatment reviews during this period, so we use the red bar to represent the return rate for all eventual participants that will be treated after the system update. We see that before the system update, there is no systematic pre-treatment differences in return rate across the products that receive fit-reference reviews versus those that do not These results give us strong model free evidence of the impact of the treatment on lowering return rates.

Next, in Table 2, panel A, we compare the means of average return rates across the comparison and treated groups of products in the pre-treatment period. The average return rate was 22% for the comparison products versus 24% for the treated group, which is only marginally significant at the .10 level, suggesting that treated products had a slightly higher return rate prior to the review system change. In panel B, we examine the differences in means of covariates and controls between the comparison and treated products in the pre-treatment period. Panel B in Table 2 shows that none of those measures differs significantly (at .05 level) between the two groups, with the exception that treated products receive slightly longer reviews. Most importantly, the two groups do not differ in the average percentage of pre-treatment reviews that contained fit-valence expressions or the average value of fit valence. This demonstrates that the comparison group is likely to be a good proxy for the counterfactual (Goldfarb and Tucker 2014).

Table 2. Comparison of Means

Variables	Mean (Comparison) N = 696	Mean (Treated) N = 246	T-statistic (P-value)
A. Outcome:			
Pre-treatment return rate	0.22	0.24	-1.80 (.0721)
B. Covariates and Controls:			
Fit_val%	0.76	0.74	0.58 (.5643)
Avg_fit_val	0.47	0.44	0.81 (.4207)
Avg_rating	4.42	4.46	-0.87 (.3856)
Rev_volume	8.67	7.75	0.84 (.4021)
Avg_rev_length	82.92	90.11	-1.88 (.0608)
Avg_price	119.65	133.97	-1.45 (.1468)
Avg_disc	-0.14	-0.23	1.03 (.3056)
Length_period	153.46	154.04	-0.61 (.9515)
Holidays	1.82	1.82	-0.04 (.9653)
Avg_Age	419.13	436.19	-0.93(.3544)

Endogenous Selection

It is crucial to understand the source of variation in the treatment effect in order to be confident of the treatment effect estimates. One important concern is that incidence of the treatment is affected by the very outcome we wish to test (i.e. product return rate). The problem of endogenous selection could arise from one of several (unobserved) reasons. The firm may have selectively induced or encouraged usage of the new review function for products with high return rate rather than those with low return rate, if the fit-reference function is especially designed to deal with high return rate. Also, customers who return a product may tend to post reviews by using the new review function. The two scenarios would result in the problem of regression to the mean, i.e. the drop of product return rate of the treatment products is likely to be a result of high pre-treatment product return rate rather than a result of the treatment effect. An opposite situation occurs if fit-reference reviews are more likely to appear when product return rate is low. Whereas it does not seem logical that the online retailer would induce treatment reviews for products with low return rate, it is entirely possible that customers who purchase products that have low historical return rate (or those who do not return products) are more likely to write reviews by using the fit-reference function. Such a case would lead to the issue of reverse causality.

Our data provider's original goal of updating the review system was to help increase the conversion rate for the apparel products, but not to reduce the product return rate. Even though, the firm did not intentionally induce reviews for products with low conversion rate. We verified this by conducting several analyses at the product level. First, we gathered all the reviews generated for the apparel category in the first month after the system update and examined whether the lagged product return measures would: a) hasten the appearance of new reviews by examining the impact of time to first review, b) would attract fit-reference information in the very first review after the update; and b) affect the total number of fit-reference reviews generated after the system update. If products with low/high return rate were likely to receive fit-reference reviews sooner and received more of them, we should observe a significant coefficient for the lagged return rate in the results presented in Table 3. The dependent variable in Models 1 and 2 is time measures in days; in Models 3 and 4, the outcome is a dummy indicating whether the first review generated after the system update is a fit-reference review or not; in Models 5 and 6, the dependent varable is the count of fit-reference reviews generated within a month after the system update. In Models 1, 3 and 5, we use a one-month lagged product return rate, whereas in Models 2, 4 and 6 we add the one-year lagged return rate. Specifications for main models are listed in Table3. For robustness checks, we use logit specification for Models 3 and 4, and QMLE specification for Models 1, 2, 5 and 6. Across all models, there is no significant relationship between any lagged product return measures and the incidence of treatment reviews.

Table 3. Product Level Estimation on Endogenous Selection

Dependent Variable	Time to First Review		Dummy Treatment		# of Treatment Reviews	
	Model 1 (log-linear)	Model 2 (log-linear)	Model 3 (probit)	Model 4 (probit)	Model 5 (log-linear)	Model 6 (log-linear)
Lag Avg. Rating	.0014 (.0564)	.0043 (.0567)	.1963* (.0933)	.2002* (.0938)	.0306 (.0252)	.0308 (.0254)
Lag Rating Volume	-.0876* (.0438)	-.0871* (.0439)	-.0524 (.0733)	-.0520 (.0733)	.0376 (.0197)	.0377 (.0197)
Rating	.0335 (.0289)	.0343 (.0289)	.0215 (.0476)	.0226 (.0477)	NA	NA
Lag # of Sales	-.0792 (.0728)	-.0732 (.0736)	-.0096 (.1236)	-.0037 (.1248)	.0665* (.0325)	.0669* (.0329)
Lag Conversion Rate	-.3390 (.3900)	-.3334 (.3903)	-1.0745 (.6358)	-1.0669 (.6361)	-.2599 (.1742)	-.2594 (.1744)
Lag # of Returns	.1155 (.0918)	.1129 (.0920)	-.0209 (.1556)	-.0229 (.1557)	-.0093 (.0412)	-.0095 (.0413)
Lag Month Return Rate	-.0403 (.1729)	-.0514 (.1742)	-.1946 (.2897)	.2073 (.2919)	-.0527 (.0770)	-.0535 (.0776)
Lag Year Return Rate		.1300 (.2367)		.1461 (.4011)		.0092 (.1063)
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
AIC	2020.7	2032.2	1003.4	1005.2	727.4	729.4
# of Observations	815	815	815	815	815	815
**** p< .001, ** p< .01, * p< .05*						

We also analyzed at the order level whether purchasers who returned the products are more (or less) likely to post a fit-reference review. We tested this a month after the system update. While this data is typically hard to obtain for researchers, the rich dataset from our data provider allows us to link customer IDs in the sales table with those in the review table. The extracted sample consists of 44,015 orders, among which 627 lead to a post-purchase product review[6].

Table 4. Order-Level Estimation on Endogenous Selection

	First Stage	Second Stage
Dependent Variable	Consumer Posts a Review?	Review Contains Treatment?
# of Items Purchased in the Order	-.3589* (.1728)	4.1801 (117.0212)
Sales Amount of the Order ($)	-.1019*** (.0193)	.1150 (.1298)
Order Shipping Fee ($)	.0020 (.0048)	-.0356* (.0164)
Lag Rating Volume Posted by Customer	.1172*** (.0060)	.0365 (.1056)
Product Returned	.0642 (.0399)	.0199 (.1522)
Inverse Mills Ratio		.6042 (1.1309)
AIC	6049.4	714.3
# of Observations	44,015	627
**** p< .001, ** p< .01, * p< .05*		

We adopt a two-stage model (Heckman 1979), where the outcome in the first stage measures whether the customer posts a review or not, and at the second stage, the outcome indicates if the reviewer adopts the fit-reference function. We applied probit estimation in the first stage, and then obtained and included the inverse Mills ratio, along with the main parameter of interest – *Product Returned* dummy, in the second

[6] Reviews posted within the first month after the system update also include sales made before the update, hence the difference in N.

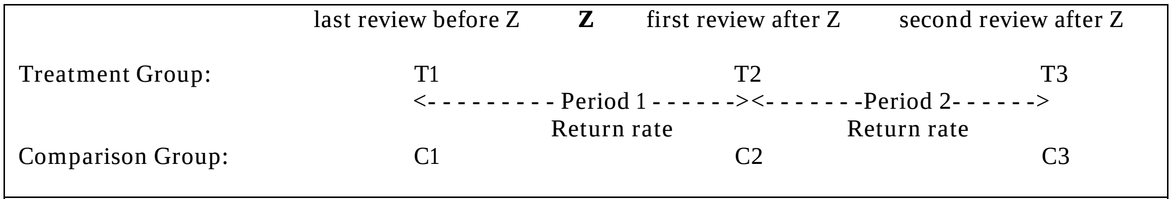
stage. The impact of sales, order characteristics, and past rating behavior of the customer are controlled. Table 4 reports the results. We find that the two stages do not appear to be correlated (i.e. the estimate of *Inverse Mills Ratio* is insignificant); it means that the decision to adopt the new fit-reference function is independent of the decision to post a review. Second, a consumer's decision to return the product does not affect her adoption of the fit-reference function (i.e. coefficient of *Product Returned* is not significant). Overall, the analyses at the product- and order-level are consistent. They help mitigate the concern of endogenous selection, which increases our confidence that treatment is random.

Empirical Analyses

To analyze the effect on return rates of the introduction of fit-reference information in online reviews, we apply the difference-in-differences estimation technique. We take the timing of the introduction of the review system change as exogenous, similar to previous studies (Chen et al. 2011, Sun and Zhu 2013). Identification of the treatment effect then relies on the assumption that treatment is randomized across products (as verified above). Recall that treatment effects are staggered in our data, with different products receiving their first treatment reviews at different times after the reviews system change.

We define a panel where the time dimension is measured using review posting date corresponding to the arrival of any review. Our focal variables of interest – those related to fit information – only change when new reviews are received. On a day that a focal review was posted; cumulative review information received prior to that day is modeled as influencing the product return rate of purchases made during the following period, which is the time between the focal review and its following review. For example, assume a review is posted on Jan 18, 2015 and the next one is posted on Feb 18, 2015. According to our design, we would examine the impact of all reviews posted by Jan 18, 2015 on the product return rate of all orders placed between Jan 18, 2015 and Feb 18, 2015.

We present the estimation of a two-period panel, where the first period is the time between the last review posted before system update and the first review posted after the system update, and the second period is between the first and second reviews posted after the system update. In this model, the treated products are those whose first review after Feb 18, 2015 contains fit reference information, and the comparison group consist of the remaining products that don't receive fit reference information after the treatment. Figure 4 illustrates the experimental design for the two-period estimation. We borrow the notation from Shadish et al. (2002): T_t denotes the reviews received by products in the treatment group and C_t are reviews received by products in the comparison group at time t and Z denotes the update of the review system.



last review before Z **Z** first review after Z second review after Z

Treatment Group: T1 T2 T3

<- - - - - - - - - - Period 1 - - - - - -><- - - - - - - -Period 2- - - - - ->

Return rate Return rate

Comparison Group: C1 C2 C3

Figure 5. The Quasi-Experimental Design (Z: introduction of treatment, i.e. system update)

Model Specifications

We employ a series of model specifications to examine the return rate for product *i* at time *t* as a function of fit information. In equation [1] the dummy *After* represents the post treatment period, and captures the effect of time-variant unobservable that occur after the system change but not before. α_1 is the effect of fit-valence expressions and α_3 is difference-in-differences estimate of fit-reference information. *R* includes review-related time-varying controls such as *Avg_rating*, *Rev_volume*, and *Avg_rev_length* as described in Table 1. We use product fixed effects $\theta_{i,}$ to control for product-level unobserved heterogeneity and τ_t controls for quarterly effects (i.e. seasonality). We also use product-level cluster standard errors to help address heteroscedasticity (Long et al. 2000) and serial correlation (Bertrand et al. 2004).

$$Return\_rate_{it} = \alpha_1 Fit\_val\%_{it} + \alpha_2 After_{it} + \alpha_3 After_{it} \times Fit\_ref\%_{it} + \alpha_4 R_{it} + \theta_i + \varepsilon_{it} \qquad [1]$$

Next, we replace *Fit_val%* and *Fit_ref%* with *Only_fit_val%, Only_fit_ref%* and *Fit_comb%* in equation [2]. The difference-in-differences estimate of the treatment effect of interest is given by $\beta_{4.}$ We include additional control variables as described in Table 1. In equation [3], we further decompose the main treatment effect into positive valence expressions combined with fit reference (i.e. *Fit_comb_pos%*), and negative valence expressions combined with fit reference (i.e. *Fit_comb_neg%*).

$$Return\_rate_{it} = \beta_1 Only\_Fit\_val\%_{it} + \beta_2 After_{it} + \beta_3 After_{it} \times Only\_fit\_ref\%_{it} + \beta_4 After_{it} \times Fit\_comb\%_{it} + \beta_5 R_{it} + \beta_6 X_{it} + \theta_i + \tau_t + \varepsilon_{it} \quad [2]$$

$$Return\_rate_{it} = \gamma_1 OnlyFit\_val\%_{it} + \gamma_2 After_{it} + \gamma_3 After_{it} \times Only\_fit\_ref\%_{it} + \gamma_4 After_{it} \times Fit\_comb\_pos\%_{it} + \gamma_5 After_{it} \times Fit\_comb\_neg\%_{it} + \gamma_6 R_{it} + \gamma_7 X_{it} + \theta_i + \tau_t + \varepsilon_{it} \quad [3]$$

Empirical Findings

Table 5. Impact of Fit Information on Product Return Rate

	Model 1	Model 2a	Model 2b	Model 3
Fit_val%	-.0775 (.0593)			
After × Fit_ref%	-.0924 (.0551)			
Only_fit_val%		-.0699 (.0662)	-.0794 (.0670)	-.0983 (.0638)
After × Only_fit_ref%		-.0572 (.1377)	-.0676 (.1393)	-.0737 (.1361)
After × Fit_comb%		-.1661* (.0804)	-.1612* (.0794)	
After × Fit_comb_pos%				-.2067** (.0772)
After × Fit_comb_neg%				-.0968 (.0847)
Avg_rating	-.0332 (.0230)	-.0323 (.0231)	-.0268 (.0225)	-.0277 (.0196)
Rev_volume	.0513 (.0399)	.0524 (.0398)	.0546 (.0427)	.0584 (.0453)
Avg_rev_length	.0264 (.0241)	.0264 (.0241)	.0304 (.0240)	.0297 (.0219)
Avg_fit_val	.0327 (.0332)	.0323 (.0331)	.0312 (.0326)	.0608 (.0350)
Avg_price			.1554** (.0484)	.1572*** (.0385)
Avg_disc			.0034*** (.0007)	.0034 (.0019)
Age			-.0238 (.0263)	-0.247 (.0290)
Length_period			.0058 (.0097)	.0057 (.0086)
Holidays			-.0201* (.0079)	-.0202* (.0081)
Quarter Dummies	No	No	Yes	Yes
After Dummy	Yes	Yes	Yes	Yes
R-squared	0.033	0.033	0.060	0.061
F-statistic	4.582***	4.015***	3.675***	3.561***
# Products	942	942	942	942
# Observations	1,884	1,884	1,884	1,884

*** $p < .001$, ** $p < .01$, * $p < .05$

In Model 1, although the estimate of *After × Fit_ref%* is negative, it is marginally significant at the 0.1 level. It shows that the newly-added fit-reference information may help reduce the product return rate. In Models 2a and 2b, which differ only in the inclusion of additional non-review controls, we find that while the fit-reference information itself (i.e. *After × Only_fit_ref%*) does not have an impact, but the combination of fit-reference and fit-valence expressions (i.e. *After × Fit_comb%*) has a significant negative impact on product return rate. The estimate shows that, on average, when the percentage of reviews that contain both fit reference and fit valence increases by 10% (e.g. adding one such review to a product that already has ten reviews without both types of fit information), the subsequent product return rate would decrease by around 1.6%. Considering the huge transaction volume (more than a million orders per year) of apparel products on the website of our data provider, this could potentially generate huge cost savings.

Finally, in Model 3, the results suggest that both positive and negative fit valence expressions, when combined with fit reference information, have a negative relationship with product return rate, although only positive valence has a significant impact. The finding of insignificant estimates of all instances measuring fit valence alone suggests that valence itself does not really matter in the context of product fit. It becomes meaningful for customers only when the relevant fit-reference information is also provided.

Robustness Checks and Inference Validity

In this section, we address several concerns about potential threats to internal validity. All robustness check results are reported in Table 6, which include checks for different functional forms, controls, and samples.

Table 6. Robustness Checks for Different Functional Specifications, Controls, and Samples

	Functional Forms			**Controls**		**Samples**		
	Censored Regression	Log - linear	Binary Treatment	Cust. Avg. Past Returns	Textual Fit Info	One-time Buyers	Sufficient Sales	Extend Sample
Only_fit_val%	-.1074 (.0857)	-.0649 (.0500)	-.0419 (.0276)	-.0691 (.0632)	-.0761 (.0620)	-.0104 (.0391)	-.0023 (.0393)	-.0739 (.0624)
After × Only_fit_ref%	.0462 (.1816)	-.0692 (.1109)	-.0204 (.0505)	-.0805 (.1351)	-.0859 (.1323)	-.0700 (.0883)	-.1277 (.0931)	-.0775 (.1353)
After × Fit_comb%	-.1745* (.0853)	-.1301* (.0593)	-.0447 (.0254)	-.1539* (.0737)	-.1652* (.0726)	-.1372** (.0476)	-.1007* (.0470)	-.1743* (.0742)
Customer Avg. Past Returns				.0656*** (.0135)	.0666*** (.0135)			.0578*** (.0072)
Only_fit_val_text%					.0414 (.0601)			.0510 (.0640)
Only_fit_ref_text%					.0835 (.0495)			.0819 (.0494)
Fit_comb_text%					.0833 (.0555)			.0785 (.0560)
Quarter Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
After Dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	NA	0.073	0.057	.0959	.0999	0.259	0.073	.0730
F-statistic/Wald-statistic	26.490***	4.534***	3.529***	5.772***	5.119***	63.843***	4.534***	12.160***
# of Products	942	942	942	942	942	931	790	2,486
# of Observations	1,884	1,884	1,884	1,884	1,884	1,862	1,580	4,972

***** *p< .001, ** p< .01, * p< .05*

The first and second columns use a different functional form, namely censored and log-linear regression to better handle a percentage outcome. The third column uses a dummy to represent the treatment – presence or absence of combined fit information. These three models change the functional forms of the estimation.

Our next goal is to control several important omitted factors in affecting product return rate. The fourth column controls for potential unobserved heterogeneity stemming from customer's self-selection bias that arises if early purchasers are different than later purchasers in their return behaviors. We test and control for the possibility that customers who purchase earlier in the lifecycle have greater product fit uncertainty, and rely more on in-home trials, than customers who buy later, thereby driving the reduction in return rate over time. To deal with the challenge, we use the average number of past returns of purchasers in each period for a product as additional variables in the model to proxy for their tendency to rely on fit trials. In the fifth column, we address the possibility that the review text may contain information about fit, which is overlooked in our model. Specifically, we apply text mining techniques to extract the fit-valence opinions and body size information found in the text, corresponding to those presented in the review function.

The last three models change the sample size for estimations, based on different rationales. The sixth restrict our sample to only those purchases from customers who made one-time purchases of products in a brand, dropping our sample to 931 products. The elimination of customers who made repeat purchases of a brand at the retailer helps address the concern that brand familiarity may be causing the reduced return rate. A further restriction is taken to only include those products with sufficient sales by eliminating products in the lower 25th percentile of sales. The relevant estimation results are presented in the seventh column. This sample of 790 products is likely to have more stable return rates. Last, in the eighth column, we conduct estimation on an extended sample that includes products with no reviews posted after the system update, raising sample size to 2,486 products

All the results of robustness checks are shown in Table 6. And the models include all relevant controls as in the main models. Due to space limitation, we do not report the results of main controls in this table. The treatment effect given by *After* × *Fit_comb%* remains significant in all models, and the coefficient of *After* × *Only_fit_ref%* is not in any model, further supporting our main hypothesis.

Overall, we have addressed most concerns about the threats to internal validity during the entire process of our quasi-experimental analysis. Table 7 records the detail about how we dealt with each potential issue. We believe our estimation procedure is rigorous enough to infer a solid causal relationship between the combination of fit-valence and fit-reference review information and the subsequent product return rate.

Table 7. How We Handle the Potential Threats to Internal Validity

Potential Threats	Problem Description	How We Handle the Problems
Selection Bias	Treatment is not random	▪ Our data provider claims that there is no selective inducement of treatment reviews for specific products. ▪ There is no evidence showing that past return outcomes would affect the appearance of treatment reviews. Other potential drivers of treatment reviews are controlled in the main models.
	Customers in two periods are different in return behaviors	▪ We use number of past returned orders as a proxy to control customers' self-selection bias.
	Estimation sample is not representative.	▪ Our main estimation sample includes all apparel products that have reviews posted both before and after the system update. ▪ In the extended sample, we include 1,544 more products that do not have new review generated after the system update.
Selection Maturation	Estimated result is due to experimental groups maturing overtime.	▪ The main estimation sample contains products at different lifetime stages. The extended sample even covers a broader range of the product life cycle. ▪ Product age is controlled in the models.
Mortality	Estimated result is due to sample attrition.	▪ There is no significant difference between review measures for products in the estimation sample and products dropped due to no sales in either of the two estimation periods.
History / Events	Omitted variables that co-vary across periods drive the change in outcomes.	▪ There is no return policy change or other review system update events during the time period for data analysis. ▪ Seasonality and holiday events are controlled. ▪ Textual review information about product fit is controlled.
Regression toward the Mean	Treatment is a result of extreme levels of past outcomes.	▪ Our preliminary analyses indicate that past product return rate does not have significant relationship with the incidence (or volume) of treatment reviews.
Reverse Causality	Causal direction may be wrong.	▪ There is time lag between the treatment and the outcome, which clearly shows the direction of the causal relationship.

As for external validity, our analyses include all the different apparel products sold by the online retailer who offers us data. We do note, however, that the main business of our data provider is outdoor products.

Thus, we caution that our findings may be more representative of outdoor apparel. Future studies can extend the analysis to a broader range of apparel goods and even to the category of footwear.

Conclusion

We leverage a natural experiment that introduced new fit review functions at an online retailer to examine the impact of different types of product fit information in online reviews on the subsequent product return rate for apparel goods. Our findings provide evidence that mere fit-valence expressions or fit-reference information by themselves do not lower product return rate. Rather, it is the combination of the two types of product fit information that matters. Specifically, when a review contains both fit-valence expression and fit-reference information (such as the reviewer's body size and size purchased), it assists a subsequent customer to infer the right size that would fits her preferences, thereby enabling her to make more accurate product purchases and reducing her need to return products. We also demonstrate the mechanism by which the new fit review functions improve consumers' decision-making. Specifically, we find strong support for the theory of semantic relativism that explains why and how both types of product fit information (i.e. fit valence and fit reference) are important for online consumers' purchase-related decision-making. Finally, our sample includes a variety of apparel products (shirts, jackets, pants, shorts etc.), therefore providing external validity and increasing the generalizability of our results. We believe that our study adds valuable contributions into the growing literature of product fit uncertainty in online market (Hong and Pavlou 2014), especially to the specific context of online product reviews (Kwark et al. 2014). While most of the literature has focused on quality dimensions in reviews, for products where fit is key, we show that fit information in reviews plays a more important role in affecting online product returns.

Our findings offer several important implications for firms. First, our study provides guidance to firms that are looking to improve their online review systems. For product categories that are highly sensitive to fit, we recommend that retailers implement review functions that allow reviewers to provide both a summary of fit-valence expression and their fit-relevant preferences such as body type, measurements etc. This will enable future customers to better interpret the fit-valence information found in reviews of previous customers. In line with this logic, we suggest that the review model used by retailers such as Zappos.com, that only uses a fit-valence function, can be improved by allowing users to add structured fit-reference information as well. Similarly, online retailers that only provide the fit-reference information (e.g. Urban Outfitters) can also improve the effectiveness of their review system by allowing reviewers to provide fit-valence information. Second, our results show that fit information found embedded in review text does not have the same impacts that structured fit-valence and fit-reference information do, highlighting the need for firms to provide an avenue for reviewers to offer this information, rather than relying on them to include that in their textual reviews. Alternately, it is worthwhile for retailers to mine this information from review text and make it salient by presenting it in a structured way. The summarized product fit information makes it easy for future customers to judge the product fit, and can help reduce purchase errors. Third, our findings demonstrate the value of using multi-dimensional rating systems that separate quality opinions from fit opinions, rather than single ratings or scores in the review systems.

We propose several directions for future work. Fit to one's taste or needs is important for both products and services, however, in this study we focus only on products that consumers can return upon experiencing a lack of fit. In contrast, services such as a hotel room cannot usually be returned even if consumers are dissatisfied with its fit with their preferences. It would be useful to examine the impact of new fit review functions in the context of services. We can extend our analysis to multi-period panels to examine how the impact of new fit information in reviews evolve over time. We also plan to look in greater detail at the differing impacts of positive vs negative fit-valence expressions.

References

Anderson, E. T., and Simester, D. I. 2014. "Reviews Without a Purchase: Low Ratings, Loyal Customers, and Deception," *Journal of Marketing Research* (51:3), pp. 249-269.

Anderson, E. T., Hansen, K., and Simester, D. 2009. "The Option Value of Returns: Theory and Empirical Evidence," *Marketing Science* (28:3), pp. 405-423.

Archak, N., Ghose, A., and Ipeirotis, P. G. 2011. "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Management Science* (57:8), pp. 1485-1509.

Bertrand, M., Duflo, E., and Mullainathan S. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics* (119:1), pp. 249-275.

Besley, T., and Case, A. 2000. "Unnatural Experiments? Estimating the Incidence of Endogenous Policies," *The Economic Journal* (110:467), pp. 672-694.

Chen, Y., Wang, Q., and Xie, J. 2011. "Online Social Interactions: A Natural Experiment on Word of Mouth Versus Observational Learning," *Journal of Marketing Research* (48:2), pp. 238-254.

Chevalier, J., and Mayzlin, D. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research* (43:3), pp. 345-354.

Chintagunta, P. K., Gopinath, S., and Venkataraman, S. 2010. "The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets," *Marketing Science* (29:5), pp. 944-957.

Dave, K., Lawrence, S. and Pennock, D.M. 2003. "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," in *Proceedings of the 12th International Conference on World Wide Web*, Budapest, Hungary

De, P., Hu, Y. J., and Rahman, M. S. 2013. "Product-Oriented Web Technologies and Product Returns: An Exploratory Study," *Information Systems Research* (24:4), pp. 998-1010.

Duan, W., Gu, B., and Whinston, A. B. 2008. "Do Online Reviews Matter? – An Empirical Investigation of Panel Data," *Decision Support Systems* (45:4), pp. 1007-1016.

Garvin, D. A. 1984. "What Does Product Quality Really Mean?" *Slogan Management Review* (26:10), pp. 25-43.

Gallino, S., and Moreno, A. 2015. "The Value of Fit Information in Online Retail: Evidence from a Randomized Field Experiment," Available at SSRN: http://ssrn.com/abstract=2677404

Ghose, A., and Ipeirotis, P. G. 2011. "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics," *IEEE Transactions on Knowledge and Data Engineering* (23:10), pp. 1498-1512.

Godes, D., and Mayzlin, D. 2004. "Using Online Conversations to Study Word-of-Mouth Communication," *Marketing Science* (23:4), pp. 545-560.

Godes, D., and Silvva, J. C. 2012. "Sequential and Temporal Dynamics of Online Opinion," *Marketing Science* (31:3), pp. 448-473.

Goldfarb, A., and Tucker, C. 2014. "Conducting Research with Quasi-Experiments: A Guide for Marketers," *Rotman School of Management Working Paper*, Available at SSRN: http://ssrn.com/abstract=2420920

Gopinath, S., Jacquelyn, S. T., and Krishnamurthi L. 2014. "Investigating the Relationship Between the Content of Online Word of Mouth, Advertising, and Brand Performance," *Marketing Science* (33:2), pp. 241-258.

Guide, V., Souza, G., Wassenhove, L., and Blackburn, J. 2006. "Time Value of Commercial Product Returns," *Management Science* (52:8), pp. 1-15.

Heckman, J. J. 1979. "Sample Selection Bias as Specification Error," *Econometrica* (47:1), pp. 153-161.

Hempel, C. G. 1950. "Problems and Changes in the Empiricist Criterion of Meaning," *Revue Internationale de Philosophie* (41:11), pp. 41-63.

Hong, Y., and Pavlou, P. A. 2014. "Product Fit Uncertainty in Online Markets: Nature, Effects and Antecedents," *Information Systems Research* (25:2), pp. 328-344.

Imbens, G. W., and Wooldridge, J. 2009. "Recent Developments in the Econometrics of Program Evaluation," (47:1), pp. 5-86.

Jackson, E. 2015. "What Online Apparel Shoppers Want," Accessed May 4, 2016, https://www.internetretailer.com/commentary/2015/04/23/what-online-apparel-shoppers-want

Kaplan, David. 1989. *Demonstratives. An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals* (originally 1971). In Themes from Kaplan, eds. Joseph Almog, John Perry and Howard Wettstein, 481-614. Oxford: Oxford University Press.

Kopalle, P. K., and Lehmann, D. R. 1995. "The Effects of Advertised and Observed Quality on Expectations about New Product Quality," *Journal of Marketing Research* (32:3), pp. 280-290.

Kwark, Y., Chen, J., and Raghunathan, S. 2014. "Online Product Reviews: Implications for Retailers and Competing Manufacturers," *Information Systems Research* (25:1), pp. 93-110.

Lal, R. and Sarvary, M. 1999. "When and How Is the Internet Likely to Decrease Price Competition?" *Marketing Science* (18:4), p. 485-503.

Lancaster, K.J. 1971. *Consumer Demand: A New Approach*, New York: Columbia University Press.

Lancaster. K.J. 1979. *Variety, Equity and Efficiency*, New York: Columbia University Press.

Lasersohn, Peter. 2005. "Context dependence, disagreement, and predicates of personal taste." *Linguistics and Philosophy* (28), pp. 643–686.

Li, X., and Hitt, L. M. 2008. "Self-Selection and Information Role of Online Product Reviews," *Information Systems Research* (19:4), pp. 456-474.

Li, X., and Hitt, L. M. 2010. "Price Effects in Online Product Reviews: An Analytical Model and Empirical Analysis," *MIS Quarterly* (34:4), pp. 809-831.

Liu, Y. 2006. "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of Marketing* (70:3), pp. 74-89.

Liu, B., Hu, M. and Cheng, J. 2005. "Opinion Observer: Analyzing and comparing opinions on the Web," in *Proceedings of the International Conference of the World Wide Web*, Chiba, Japan.

Liu, Y., Chen, P., and Hong, Y. 2014. "Value of Multi-Dimensional Rating Systems: An Information Transfer View," in *Proceedings of the 35th International Conference on Information Systems*, Auckland, New Zealand.

Long, J. S., and Ervin, L. H. 2000. "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model," *The American Statistician*, (54:3) pp. 217-224.

Ludwig, S., de Ruyter, K., Friedman, M., Bruggen, E.C., Wetzels, M. and Pfann, G. 2013. "More Than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates," *Journal of Marketing* (77:1), pp.87-103.

MacFarlane, John (2005). "Making Sense of Relative Truth."Proceedings of the Aristotelian Society, 105: pp. 321–39.

MacFarlane, J. (2007) "Relativism and Disagreement", *Philosophical Studies*, 132, 17-1MacKinnon, J. G., and White, H. 1985. "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics* (29:3) pp. 305-325.

Meyer, B. 1995. "Natural and Quasi-Experiments in Economics," *Journal of Business and Economic Statistics* (13:2) p. 151-161.

Petersen, J.A. and Kumar, V. 2010. "Can Product Returns Make You Money?," *Sloan Management Review* (51:3), p. 85-89.

Powell, J. L. 1984. "Least Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics* (25:3), pp. 303-325.

Qi, Z., Stoey, V. C., and Jabr, W. 2015. "Sentiment Analysis Meets Semantic Analysis: Constructing Insight Knowledge Bases," in *Proceedings of the 36th International Conference on Information Systems, Fort Worth, TX.*

Rosario, A.B., Sotgiu, F., De Valck, K., and Bijmolt, T.H.A. 2016. "The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors," *Journal of Marketing Research* (53), pp. 297-318

Sahoo, N., Dellarocas, C., and Srinivasan, S. 2015. "The Impact of Online Product Reviews on Product Returns," *Boston University School of Management Research Paper No. 2491276*. Available at SSRN: http://ssrn.com/abstract=2491276

Shadish, W., Cook, T., and Campbell, D. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston, MA: Houghton Mifflin Company.

Shulman, J.D., Cunha Jr., M. ad Saint Clair, J.K. 2015. "Consumer Uncertainty and Purchase Decision Reversals: Theory and Evidence," *Marketing Science* (34:4), p. 590-605.

Stephenson, T. 2007. "Judge Dependence, Epistemic Modals, and Predicates of Personal Taste," *Linguistics and Philosophy* (30), pp. 487-525.

Sun, M. 2012. "How Does the Variance of Product Ratings Matter?" *Management Science* (58:4), pp. 696-707.

The Economist. 2013. “Return to Santa,” Accessed May 4, 2016, http://www.economist.com/news/business/21591874-e-commerce-firms-have-hard-core-costly-impossible-please-customers-return-santa

Trueship.com. 2016. “One Size Doesn’t Fit All! How to Reduce Apparel Returns,” Accessed May 4, 2016, http://www.trueship.com/blog/2016/02/11/one-size-doesnt-fit-all-reducing-online-returns-in-fashion-ecommerce

Turney. P.D. 2002. “Thumbs up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” In *Proceedings of the 4oth Annual Meeting on Association for Computational Linguistics, July 7-12, Philadelphia, pp. 417-424.*

Ulrich, K. T. 2011. "Variety," Chapter from *Design: Creation of Artifacts in Society*. University of Pennsylvania.

Zhu, F., and Zhang X. M. 2010. “Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics,” *Journal of Marketing* (73:3), pp. 133-148.