Search Gaps*

Raluca M. Ursu[†] Qianyun Zhang[‡]

Current version: January 7, 2020
First version: July 23, 2019

PRELIMINARY AND INCOMPLETE

Abstract

Models of costly search generally describe consumers as inspecting products consecutively until they decide to stop searching, a decision which occurs only once before determining whether to purchase. In this paper, we use data on consumers' entire browsing history and show that the assumption of consecutive search with one stopping decision may not always hold. In particular, we find that consumers take frequent breaks in their search ("search gaps"), that is they obtain information on a number of products, stop, and only later restart their product search. After testing multiple hypotheses for the occurrence of search gaps, we find empirical evidence consistent with the existence of search fatigue: the more the consumer searches, the higher her search costs per option; taking a break reduces these costs and enables the consumer to restart her search at a later time. We then develop a model of sequential search that accounts for search gaps by proposing that a consumer faced with fatigue may decide not only whether to continue searching, but also when to search a product (with or without a delay). Finally, we test and validate the predictions of this model, further supporting the claim that fatigue can at least in part explain the occurrence of search gaps. By identifying when and which consumers stop searching because of fatigue rather than a low match value, search gaps can provide new targeting opportunities for managers and allow researchers to better measure consumer preference and search costs.

Keywords: consumer sequential search, search fatigue, search delay, online browsing behavior.

*We are thankful for comments from Tomomichi Amano, Alix Barasch, Bryan Bollinger, Daria Dzyabura, Eric Greenleaf, Sarah Komisarow, Eitan Muller, Anita Rao, Andrey Simonov, and from attendees of the 2019 Marketing Science/INFORMS conference. We have greatly benefited from the careful and meticulous help offered by Ishita Verma, Jian Zhang, and Xinyu Wei in augmenting and processing our data. We would also like to thank Siham El Kihal for helping us obtain the data. The usual disclaimer applies.

[†]Stern School of Business, New York University, E-mail: rursu@stern.nyu.edu.
[‡]Stern School of Business, New York University, E-mail: qianyunzhang@nyu.edu.

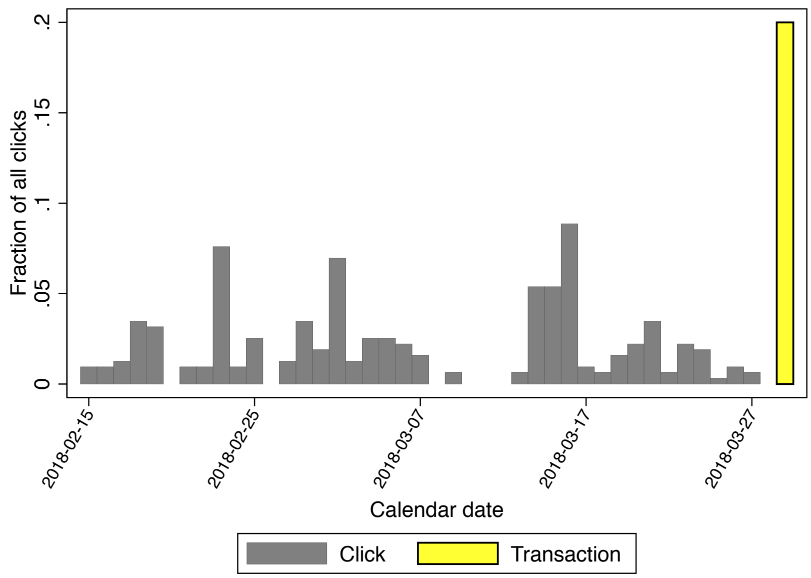
1 Introduction

Consumers typically face many product options before making a purchase decision. For example, consumers can choose among 1,000 television options under $500 on Amazon, more than 400 hotels in New York City on any given night, or almost 2,000 flights from Denver to Atlanta. Since evaluating all options can be prohibitively costly, consumers need to decide which products to become informed about and which ones to ignore.

A rich literature analyzes why consumers search some and not other products before choosing whether to purchase. More precisely, models of costly search generally describe consumers as inspecting products consecutively until they decide to stop searching, a decision which occurs only once before determining whether to purchase. For example, in the canonical sequential search model of Weitzman (1979), the consumer proceeds to searching options as long as the benefit from searching exceeds the cost. When this relation no longer holds, search ceases and the consumer determines whether to purchase. Similarly, in the simultaneous search model of Stigler (1961), the consumer stops searching and makes a purchase decision only after revealing information about the set of options with expected benefit exceeding the search cost.

This paper uses data on consumers' entire online browsing history and shows that the assumption of consecutive search with one stopping decision may not always hold. Our data come from GfK, Germany's largest market research company, and capture all web traffic (8 million clicks) of a panel of 4,726 consumers, including clicks in a focal category (fashion) and any other browsing activity that consumers perform, e.g. related to checking email, visiting social networking sites, or using search engines. With these data, we show that consumers take frequent breaks in their search activity ("search gaps"), that is they obtain information on a number of products, stop, and only later restart their product search. Precisely, 73% of consumers have at least one search gap and the average consumer has 6 search gaps. Figure 1 illustrates such search gaps in our data. Here we plot the daily search activity of a consumer in the focal category over our observation period. We also indicate when a transaction occurred. What the figure shows is that the consumer stops and restarts her search multiple times before making a purchase decision. Prior work ignores such search gaps, assuming a single stopping decision with products clicked consecutively, without any breaks in between.

Figure 1: Illustrating search gaps



Treating search as consecutive and assuming away the search gaps depicted in Figure 1 may be innocuous if the reason for such gaps were unrelated to consumer search decisions. This would be the case if for example observed search gaps resulted from the consumer's search activity being interrupted by an unrelated event, such as a work email or a planned offline activity.

However, there are several reasons to expect search gaps to be deterministic. For example, search gaps may occur when consumers expect prices or product features to change over time, and thus think they may benefit from restarting their search at a later time. Also, fatigue may increase search costs, encouraging consumers to stop searching and return after a break when these costs are lower. Or, consumers may delay their search when unsure about what other options to evaluate, and return having obtained more information from friends or other sources. Furthermore, consumers' limited budget of time or inability to afford a purchase may also explain their delayed search. Depending on the product category, one or several of these reasons might prevail and explain search gaps.

In this paper, our goal is to demonstrate that in the context of our data search gaps can be explained at least in part by fatigue: the more the consumer searches, the higher her search costs per option; taking a break reduces these costs and enables the consumer to restart her search at a later time. Towards this goal, we proceed in two steps. First, we develop a model of sequential search that account for search gaps in order to generate predictions about their occurrence. Our model extends the Weitzman (1979) framework in two directions: (i) it accounts for search gaps by allowing the consumer to decide not only whether to search or whether to stop, but also when to search an option: now or after taking a break; (ii) it allows for fatigue by modeling search costs as having two components: a baseline level and a component that depends on the number of products searched after the latest break. Second, we propose empirical tests for reasons other than fatigue that may explain the occurrence of search gaps. Using this two-step approach, we find the following.

First, a higher level of fatigue leads to more search gaps. More precisely, as predicted by our model, when the level of fatigue is higher, the value of continuing search after rather than before a break is also higher, increasing the likelihood of a search gap. Standard models of search cannot predict such an effect, since consumers are assumed to make a single stopping decision (and thus cannot have any search gaps). Empirically we show that consumers who are older, who search predominantly earlier in the day and on weekdays, and who visit websites that are slower to load, contain more photos, or are harder to read, have generally more search gaps. To the extent that these measures are good proxies for fatigue, our results confirms the model's prediction.

Second, search gaps are less likely when the consumer has uncovered a more valuable option through search. This is the case because searching after a gap involves discounting the value of continued search. The higher the value of the best option uncovered through search, the more costly it is to discount it by delaying search, implying fewer search gaps. Our data validate this prediction, which we test in two ways. First, we show that there are fewer gaps at the end of the consumer search process than at the beginning. Since the value of the best option observed thus far has to increase as the consumer continues searching, this result supports our prediction. Second, we show that consumers who start search with more valuable first options also have fewer search gaps, further strengthening the evidence for our model's prediction.

Finally, reasons other than fatigue cannot completely explain the occurrence of search gaps in our data. To arrive at this result, we first create a comprehensive list of hypotheses for why search gaps may occur, mostly influenced by prior work on choice deferral (Greenleaf and Lehmann, 1995; Dhar, 1997; Novemsky et al., 2007). The list contains reasons such as delaying search due to factors unrelated to the search process, due to an expectation of future changes in product features (e.g.

lower prices), due to uncertainty regarding what other options to search, or due to an inability to afford a purchase. Then, we empirically test these hypotheses and find that search gaps occur even absent of these reasons. Thus, our results strengthen the claim that search gaps are affected at least in part by fatigue.

These findings have important implications for marketing research and managerial practice. In terms of theory, we propose the first model of consumer search that accounts for search gaps. Observing search gaps allows researchers to account for an additional choice that consumers make while searching. To see this, consider the standard sequential search model à la Weitzman (1979) that does not account for search gaps. In the Weitzman (1979) model, consumers make two choices:

- *continue searching* if benefit(search) > benefit(stop)
- *stop and make a purchase decision* otherwise.

However, when a gap occurs, neither the benefit from searching exceeds the benefit from stopping, nor vice versa. More precisely, because the consumer chooses not to search now but to take a break, the benefit from searching cannot exceed that from stopping. And, because the consumer restarts her search at a later time and thus does not stop, the benefit from stopping cannot exceed that from searching. Rather, the occurrence of search gaps suggests that consumers choose to neither stop nor continue searching now, but instead choose to defer their search. This additional choice consumers make while searching is not captured by previous models, an omission which may lead to incorrect inferences about consumer preferences and search costs, as we demonstrate in Section 3.4. Moreover, we show that assuming fatigue does not influence search decisions can lead to an overestimation of the consideration set of the consumer (the set of products searched).

From a managerial perspective, our approach can help companies identify search fatigue and optimize pricing decisions. Managers have long sought ways to reduce consumers' burden and resulting fatigue from shopping. Reasons contributing the consumers' fatigue include the access to a plethora of information sources (expert reports, social media, customer ratings, opinion blogs, etc), the choice among an ever growing number of products and attributes, and the pressure to choose the best deal when receiving numerous promotions.[1] Succeeding in reducing fatigue may

[1] For more details, see `http://business.time.com/2012/11/14/consumer-fatigue-shopping-has-never-been-\easier-or-as-mentally-exhausting/` or `https://hbr.org/2009/09/death-by-information-overload`.

encourage consumers to search and purchase more products. Also, it may result in higher customer loyalty, as consumers reward companies that simplify their search process.[2] In this paper, we are the first to show that observing search gaps can provide managers with an indication of consumers' fatigue level, which is a necessary step to being able to measure and reduce it. Furthermore, search gaps reveal that consumers may stop searching because of a high fatigue level, rather than a low match value with a brand (high benefit from continued search). Identifying and targeting such consumers may be profitable for companies, since they are still active and thus more likely to restart searching and ultimately purchase (Schmittlein et al., 1987). In addition, this observation may have implications for firms' pricing decisions. More precisely, prior work on ordered search describes firms' optimal pricing decisions as a function of the order in which they are searched by consumers (Petrikaite, 2018; Chen and He, 2011; Zhou, 2011; Rhodes, 2011; Armstrong, Vickers, and Zhou, 2009; Armstrong and Zhou, 2011; Haan et al., 2017; Arbatskaya, 2007). In particular, Armstrong, Vickers, and Zhou (2009) show that the non-prominent firm can infer that the consumer searching it obtained a low match value at the prominent firm. In this case, the non-prominent firm will face a relatively more inelastic demand for its product, allowing it to charge a higher price than the prominent firm in equilibrium. This inference is not necessarily true when the consumer has the option to visit the non-prominent firm after a search gap. In particular, a consumer might visit the non-prominent firm even though her match at the prominent firm was high, due to her low search cost after the search gap. Thus, observing when the consumer searches an option (before or after a search gap) may inform firm pricing decisions in an ordered environment.

The rest of the paper is organized as follows. The next section discusses relevant prior work. In Section 3, we develop our model of sequential search that can explain search gaps and describe its predictions. Section 4 describes our data, while Section 5 tests the predictions of our model, as well as test other hypotheses for why search gaps arise. The last section concludes with possible future directions.

[2]More details can be found in the HBR article available at `https://hbr.org/2012/05/to-keep-your-customers-keep-it-simple`.

2 Literature Review

This paper relates primarily to three strands of the literature: (i) the theoretical consumer search literature, (ii) empirical work using individual search data to quantify consumer preferences and search costs, and (iii) prior work on choice deferral. We will now describe in detail how we related and contribute to the previous literature.

We contribute to theoretical work on consumer search in two main ways. First, we develop a new model of consumer search, adding to a rich literature that generally follows one of two frameworks: either the sequential search model of Weitzman (1979) or the simultaneous search method proposed by Stigler (1961). In both these frameworks, consumers are modeled as inspecting products consecutively until they decide to stop searching, a decision which occurs once before determining whether to purchase. For example, in the sequential search model of Weitzman (1989), the consumer proceeds to searching options as long as the benefit from searching exceeds the cost. When this relation no longer holds, search ceases and the consumer determines whether to purchase. Similarly, in the simultaneous search model of Stigler (1961) there is one stopping decision: after searching the set of options with expected benefit exceeding the search cost, the consumer stops and decides whether to buy one of the searched products. Thus, neither framework can be used to study search gaps, which is a contribution of our model. The only exception is the model developed by Morgan and Manning (1985). The authors demonstrate that under very general conditions, neither simultaneous nor sequential search is optimal, but rather a combination of the two is, where the consumer searches a set of options sequentially. Their model could give rise to search gaps, as consumers may choose sets of options to search at every occasion, taking a break between sets. However, their theory was only developed to study how consumers choose the number of options to search in every set, and not the identity of those options. Therefore, to the best of our knowledge, no prior theoretical work exists that can account for search gaps when consumers choose which products to search as well, which is one of our contributions.

Second, we contribute to the theoretical consumer search literature through our definition of search costs. Most prior work assumes search costs per product are independent of the number of products searched.[3] To the best of our knowledge, there are a few exceptions. Stiglitz (1987)

[3]For a review of theoretical work on consumer search, please see Baye et al. (2006) and Anderson and Renault (2018).

studies the effect of convex search costs on competition and the equilibrium number of firms in the market, linking this effect to the increasing scarcity of time and money that intensifies as the consumer continues searching. Levav et al. (2010) show experimentally that participants who need to customize a product (a suit or a car), are more likely to choose the default option when first presented with options that have many rather than few attributes. They argue that this result can be partially explained by convex costs of evaluating attributes, as demonstrated by literature in psychology and economics modeling self-control as a muscle that requires more effort on future rather than identical early stimulation (Ozdenoren, Salant, and Silverman, 2008; Vohs et al., 2008). Carlin and Ederer (2018) develop a model of search fatigue where the more products the consumer searched before the previous purchase, the higher her search costs when searching towards her current purchase decision. The authors then study the effect of search fatigue on firm pricing decisions in equilibrium. In contrast, in our paper the more the consumer searches before the current purchase, the higher her current search costs. Most closely related to our paper, Ursu and Dzyabura (2019) posit that search costs are increasing linearly in the number of alternatives searched and that they affect current search and purchase decisions. In our paper, search costs are also increasing linearly in the number of searches and affecting current decisions. However, the presence of increasing search costs is not sufficient for the occurrence of search gaps. More precisely, such search costs may explain why the consumer stops searching, but not why she restarts. For consumers to be willing to restart their search, search costs must also decrease during a gap (if the consumer's utility from the options available remained unchanged). To the best of our knowledge, no prior work on consumer search suggests this possibility. Instead, prior economics work on education finds that taking a break from academic classes in order to perform physical exercises, helps students recover from cognitive fatigue and perform better academically (Bednar and Rouse, 2019). We posit that a similar mechanism may drive search fatigue.

Our paper is also related to empirical work quantifying preference and search cost parameters using individual data on consumers' search activity (e.g. Hong and Shum, 2006; Moraga-Gonzalez and Wildenbeest, 2008; Kim et al., 2010, 2017; De los Santos et al. 2012; Seiler, 2013; Honka, 2014; Moraga-Gonzal ez et al., 2015; Chen and Yao, 2016; Honka and Chintagunta, 2016; De los Santos and Koulayev, 2017; Ursu, 2018). Most of this work assumes search costs per product are independent of the number of products searched. The exception is Koulayev (2014), who estimates higher search costs for products searched later, providing empirical support for the assumption of increasing search costs. Also, this rich empirical work, resting on the theoretical models of Weitzman (1979) and Stigler (1961), assumes consumers search options consecutively and can stop searching only once. Although some prior work recognizes the fact that consumers search in sessions (e.g. the search process is divided in deciles in Bronnenberg et al., 2016 and consumers learn across sessions in Wu et al., 2015), it does not explicitly model the decision of the consumer to stop and restart searching several times, and is thus not accounting for the presence of search gaps, which is our focus.

Finally, our paper relates to the literature on choice deferral. Work in consumer behavior shows that choice difficulty increases the probability of the consumer choosing none of the options and thus delaying her choice (Dhar, 1997; Novemsky et al., 2007). In the context of a search model, we view this finding as broadly suggesting that search gaps are more likely as search difficulty increases, a result which is in line with our model's predictions. More closely related is the work of Greenleaf and Lehmann (1995) that identifies several possible reasons for consumers to delay the decision to purchase a product. These include the absence of time to devote to the task, the expectation of future price decreases or other improvements in product characteristics, or the dislike for shopping. Although not described in the context of consumer search, these reasons could also influence search decisions. Thus, they form the basis of our hypothesis tests in Section 5. We contribute to this literature by developing a model of consumer search where consumers may stop and restart search, thereby formalizing the idea of delay in the context of search.

3 Model

3.1 Setup

Consider a consumer $i \in \{1, \ldots, N\}$ who seeks to purchase an alternative $j \in \{1, \ldots, J\}$ or choose the outside option of not purchasing (denoted by $j = 0$). The utility of the outside option is known (normalized to zero), but the consumer faces uncertainty about the J options. To resolve this uncertainty, the consumer can search for information before making a purchase decision, which involves paying a cost per search. The consumer's goal is to maximize her expected utility net of total search costs from the best option she will choose to purchase.

As is standard in other sequential search models such as Weitzman (1979), each search occasion, the consumer decides whether to continue searching, in which case she chooses a product to search, or whether to stop, in which case, she decides which product to purchase, if any. Let S denote the set of searched options, while $\bar{S}$ denotes the set of options still available to search. By searching an option j, the consumer reveals utility u_{ij}, characterized by a product specific continuous distribution function, $F_j(\cdot)$. Denote by y_i the best option observed by i among a set S of products searched, that is $y_i = \max_{j \in S \cup \{0\}} u_{ij}$. Consumer choices depend on the state variables $\bar{S}$ and y_i.

In this paper, we extend the Weitzman (1979) framework to account for search gaps. This involves making two modifications. First, we allow the consumer to decide not only what to search, but also when to search an option: now or after taking a break. To this end, we define a new state variable t, which tracks the number of options searched after the latest break, implying $|S| \geq t$. Having decided to continue searching, the consumer can search an option j after $t > 0$ other products, or after taking a break, which resets t to zero.

The second modification involves the nature of search costs. More precisely, in our model, we allow search costs per option to increase with the number of searched alternatives.[4] Following Ursu and Dzyabura (2019), we model the cost of searching an option as having two components. The first component, $c_{ij0} > 0$, gives the baseline cost of searching a product j regardless of the number of other products searched. This depends on characteristics of a product, such as its prominence, or on consumer characteristics, for example location or opportunity cost of time. The second component depends on t and represents the consumer's fatigue from searching. For simplicity, we assume the difference between subsequent searches is constant, that is it costs an additional $\alpha_i > 0$ for the consumer i to search j after t other products. More formally, searching j after having searched t other products costs

$$c_{ijt} = c_{ij0} + \alpha_i t. \tag{1}$$

This functional form implies that the cost of beginning search or of searching a product after taking a break equals c_{ij0}, while other searches involve paying a higher cost per product.

[4] An alternative model would to let search costs be a function of the elapsed time since search began. We leave this model to future research, and instead let search costs be a function of the number of options searched, consistent with most of the literature that considers increasing search costs (e.g. Stiglitz, 1987).

3.2 The Search Problem

For notational simplicity, we suppress the subscript i for the remainder of this section. Given state variables $(\bar{S}, t, y)$, at every search occasion when $t > 0$, the consumer solves

$$V(\bar{S}, t, y) = \max\{y, \max_{j \in \bar{S}} -c_{j0} - \alpha t + W_j(\bar{S}, t+1, y), \max_{j \in \bar{S}} \beta[-c_{j0} + W_j(\bar{S}, 1, y)]\}, \tag{2}$$

where $0 \leq \beta < 1$ is a discount factor, $V(\emptyset, t, y) = y$, and $W_j(\cdot)$ denotes the continuation value. The interpretation of the value function in equation (2) is as follows. Given a set of options available for search $\bar{S}$, a number t of options searched after the latest break, and a best option observed so far y, the consumer has three options. First, she can stop searching, in which case she gets y, representing the option of buying the product with the highest utility revealed among those searched or choosing the outside option of not purchasing. Second, she can continue searching by paying a relatively high search cost due to $t > 0$, but enjoy the continuation value now. Third, she can take a break and pay a lower search cost, but postpone receiving the continuation value, option which she discounts at the rate β. When the consumer decides to search j after taking a break, she receives no utility in the current period and t is reset to zero, resulting in a continuation value where $t = 1$ after she searches j.[5]

The continuation value $W_j(\cdot)$ is determined by the probability that the consumer will reveal a utility lower than y after searching j, in which case her problem is that of solving $V(\bar{S} \setminus \{j\}, t, y)$, and by the probability that she will reveal a utility higher than y, in which case she will solve a similar problem while holding a more valuable option in hand. Formally, the continuation value is defined by

$$W_j(\bar{S}, t, y) \quad = \quad V(\bar{S} \setminus \{j\}, t, y) \int_{-\infty}^{y} f_j(u) du + \int_{y}^{\infty} V(\bar{S} \setminus \{j\}, t, u) f_j(u) du. \tag{3}$$

If $\alpha = 0$, t is no longer a state variable in our model, so the value function simplifies to

[5]If the consumer decides to postpone searching j, then after the break she will actually want to search j, rather than switch to searching a different product. To see this, observe that equation (2) when $t = 0$ reduces to $V(\bar{S}, 0, y) = \max\{y, \max_{j \in \bar{S}} -c_{j0} + W_j(\bar{S}, 1, y), \max_{j \in \bar{S}} \beta[-c_{j0} + W_j(\bar{S}, 1, y)]\}$. Since the consumer chose to search j after the break, it means that $\max_{j \in \bar{S}} \beta[-c_{j0} + W_j(\bar{S}, 1, y)] \geq y$, which implies $\max_{j \in \bar{S}} -c_{j0} + W_j(\bar{S}, 1, y) \geq 0$ since $y \geq 0$ and $\beta \geq 0$. Thus, $\max_{j \in \bar{S}} -c_{j0} + W_j(\bar{S}, 1, y) > \max_{j \in \bar{S}} \beta[-c_{j0} + W_j(\bar{S}, 1, y)]$ and $\max_{j \in \bar{S}} -c_{j0} + W_j(\bar{S}, 1, y) > y$, proving our statement.

$V(\bar{S}, y) = \max\{y, \max_{j \in \bar{S}} -c_{j0} + W_j(\bar{S}, y)\}$ and the continuation value becomes $W_j(\bar{S}, y) = V(\bar{S} \setminus \{j\}, y) \int_{-\infty}^{y} f_j(u)du + \int_{y}^{\infty} V(\bar{S} \setminus \{j\}, u) f_j(u)du$.[6] In other words, when $\alpha = 0$, our framework reduces to the canonical sequential search model proposed by Weitzman (1979).

3.3 Proposed Solution

To solve the model, we make the following simplifying assumption about the functional form of the continuation value. Specifically, when deciding whether to continue searching, the consumer only takes into account the one step look ahead continuation value, $W_j(y)$, which is given by

$$W_j(y) \quad = \quad y \int_{-\infty}^{y} f_j(u)du + \int_{y}^{\infty} u f_j(u)du. \tag{4}$$

The one step look ahead continuation value has the following interpretation. When the consumer decides whether to continue searching, she only takes into account the value she gets from searching one more time, rather than from all possible future searches. That is, she reasons that if she searches j, then with some probability she will reveal a utility lower than the best so far, so her value will remain y, or she will reveal a higher utility, in which case the highest utility will be that of j. Our assumption means we are deviating from optimal search behavior in favor of tractability, similar to prior work (Gabaix et al., 2006; Hodgson and Lewis, 2018; Yang, Toubia, De Jong, 2015). We leave to future research to account for both search gaps in the consumer sequential search model and for optimal search rules.

Our assumption on the continuation value has two implications. First, fatigue affects the problem only through search costs, i.e. through the term αt. In fact, fatigue may have a bigger effect, influencing not only search costs, but also the continuation value (through the effect of t on $W_j(\cdot)$). In such a case, our results will provide a lower bound on the effect of fatigue. Second, the continuation value is no longer a function of $\bar{S}$ (except through the constraint that the product under consideration for search must not have been searched before). Thus, our model allows two interpretations of the set of available options to search. One, as is standard in Weitzman (1979), $\bar{S}$ is the complement of S (that is $S \cup \bar{S} = J$) and evolves deterministically as the consumer searches

[6] This simplification is correct if $-c_{j0} + W_j(\bar{S}, y)$ is positive, which, as we show in Section 3.3 needs to hold for the consumer to start searching.

more options. Another, is to think of $\bar{S}$ as any set of options available at a particular search occasion (and not previously searched), and this might change over time in a non-deterministic way. For example, in an online search setting, the products displayed on a list page (e.g. the sales page of a website) provide one set of options, while when the consumer access another list page, another set becomes available.

With these simplifications and interpretations, the problem presented in equation (2) becomes

$$V(\bar{S}, t, y) = \max\{y, \max_{j \in \bar{S}} -c_{j0} - \alpha t + W_j(y), \max_{j \in \bar{S}} \beta[-c_{j0} + W_j(y)]\}. \tag{5}$$

Our goal in this paper is to model consumer search behavior. Thus, for the initial search decision when $t = 0$ and $y = 0$, to ensure that the consumer will want to start searching, we will focus only on cases where $\exists j \in \bar{S}$ such that $-c_{j0} + W_j(0) > 0$. Otherwise, the consumer might prefer to postpone searching indefinitely. Then, $\max_{j \in \bar{S}} -c_{j0} + W_j(0) > 0$ since the statement holds for at least one $j \in \bar{S}$. Finally, note that $W_j(\cdot)$ is increasing in y (since $\partial W_j(y)/\partial y = F_j(y) \geq 0$). This implies that the value of searching now $\max_{j \in \bar{S}} -c_{j0} + W_j(y)$ is positive $\forall y$, and thus greater than the value of delaying search (which is discounted by β). This implies that the consumer will always prefer to continue searching than to take a break when $t = 0$ (conditional on not wanting to stop searching).

To solve the model, note that the product satisfying $j^* = \arg\max_{j \in \bar{S}} -c_{j0} + W_j(y)$ is the only product that the consumer would consider searching from $\bar{S}$ given y. This observation is important as it allows us to break up the problem in equation (5) into two steps. First, the consumer determines j^*. Second, she solves equation (5) for j^* and determines whether to stop searching, whether to search j^* at $t > 0$ or whether to search it after a break.

From equation (5), the search rules determining under what conditions the consumer stops, searches and takes a break are given by[7]

- Decide to search j^* after a break if

$$\begin{aligned} -c_{j^*0} + W_{j^*}(y) &\leq \frac{\alpha t}{1-\beta} \\ &\geq \frac{y}{\beta}. \end{aligned} \tag{6}$$

[7] We break ties as follows: the consumer prefers to search later if choosing between any of the three options, and prefers to search now rather than to stop.

- Decide to search j^* now if

$$\begin{aligned} -c_{j^*0} + W_{j^*}(y) &> \frac{\alpha t}{1-\beta} \\ &\geq y + \alpha t. \end{aligned} \tag{7}$$

- Decide to stop if

$$\begin{aligned} -c_{j^*0} + W_{j^*}(y) &< y + \alpha t \\ &< \frac{y}{\beta}. \end{aligned} \tag{8}$$

If the consumer stops, she chooses to buy the product with the highest observed utility among those searched, or chooses the outside option of not purchasing. Formally, the choice rule dictates that if the consumer chooses $\bar{j}$ (including the outside option), then her utility from this choice must exceed the utilities of all other options searched in S, that is

$$u_{\bar{j}} \geq \max_{k \in S \cup \{0\}} u_k. \tag{9}$$

3.4 Implications for Inference from Consumer Search Models

Preferences and Search Costs

As described in the introduction, our model accounts for an additional choice that consumers make while searching, imposing different inequalities on consumer preference and search costs than those found in the literature. To demonstrate this more formally, suppose a consumer searches product j after a gap. Existing models that assume away gaps would say that since the consumer searched j, then it must be that $-c_{j0} + W_j(y) \geq y$, for the relevant y at a given moment in the search process.[8] However, acknowledging the search gap, our model would reveal that the correct inequalities on the parameters of interest are twofold (following directly from the inequalities in 6): $-c_{j0}+W_j(y) \geq y/\beta$ and $-c_{j0} + W_j(y) \leq \alpha t/(1-\beta)$. There are two differences worth noting. First, our model imposes both an upper and a lower bound on the parameters of interest, rather than just a lower bound. Thus, it can more accurately pin down the values of $W_j(\cdot)$ and c_{j0}, representing the consumer's preferences and search costs. Second, the lower bound on $-c_{j0} + W_j(y)$ is higher than in models

[8] To see this, recall that the value function in a model without search gaps equals the one in Weitzman (1979): $V(\bar{S}, y) = \max\{y, \max_{j \in \bar{S}} -c_{j0} + W_j(\bar{S}, y)\}$. Having assumed the one step look ahead nature of the continuation value $W_j(\cdot)$, the statement follows.

that ignore search gaps (since $y/\beta > y$). This means that by ignoring search gaps, such models interpret the consumer's actions as being driven by lower preferences or higher search costs or both, than if they acknowledged the existence of search gaps. In other words, by ignoring search gaps, existing models may lead to incorrect inferences about consumer preferences and search costs.

Consideration Set Size

The consumer is less likely to search an option j when $\alpha > 0$ than if $\alpha = 0$, implying that the consideration set in the former case is smaller. To see this, consider the two cases separately. If $\alpha = 0$, then from the search rules above, the consumer will search an option j given y if $W_j(y) > y + c_{j0}$. In contrast, if $\alpha > 0$, she will search j given y only if $W_j(y) > y + c_{j0} + \alpha t$ or if $W_j(y) > \frac{y}{\beta} + c_{j0}$, meaning that she is less likely to search j than if $\alpha = 0$. Thus, our results highlight the importance of determining the value of fatigue α in order to correctly predict the size of the consideration set of the consumer. Assuming fatigue does not influence search decisions can lead to an overestimation of the size of consumers' consideration sets.

3.5 Testable Model Predictions

In this section, we develop two predictions from our model that distinguish it from existing models ignoring search gaps. In the next two sections, we introduce our data set and then use it to validate these predictions.

3.5.1 Effect of α on Search Gaps

The fatigue parameter $\alpha > 0$ denotes the change in baseline search costs due to searching an option without a break. It is straightforward to determine its effect on search gaps from equation 5. Proposition 1 describes our result.

Proposition 1. *A higher α will increase the likelihood of a search gap, ceteris paribus.*

Proof: A change in α only affects the decision to search product j now. In particular, it makes it less likely that the consumer searches when $t > 0$, and thus more likely that a search gap occurs, as can be seen directly from equation 5. $\square$

In words, we find that a higher fatigue level α increases the likelihood of a search gap.

3.5.2 Effect of y on Search Gaps

In our model, y denotes the best observed option after searching a set S, that is $y = \max_{j \in S \cup \{0\}} u_j$, where $y \geq 0$. To determine the effect of y, consider equation 5, which we reproduce here:

$$V(\bar{S}, t, y) = \max\{y, \max_{j \in \bar{S}} -c_{j0} - \alpha t + W_j(y), \max_{j \in \bar{S}} \beta[-c_{j0} + W_j(y)]\}.$$

Changing y affects all decisions of the consumer. Intuitively, a higher y makes it more likely that the consumer stops searching and picks the option j with the highest realized utility. Also, it increases $W_j(y)$ affecting both the decisions to continue searching now and after a break. The former effect is clear and well studied in the literature, saying that the more valuable the best option the consumer has in hand, the more likely she is to stop searching and to purchase it.

Here we focus on the effect of y on search gaps, that is on the decision of the consumer to search j now versus after a break, conditional on the consumer not stopping her search. To analyze this effect, we consider how changes in y affect the two search decisions of the consumer.[9] Proposition 2 below describes our result.

Proposition 2. *Conditional on search not ceasing, a higher y will decrease the likelihood of a search gap, ceteris paribus.*

Proof: A change in y changes the decision to search now by $F_j(y)$ and the decision to search after a break by $\beta F_j(y)$, given that $\partial W_j(y)/\partial y = F_j(y)$. Since $0 \leq \beta < 1$ and $F_j(y) \geq 0$, conditional on continuing search, there will be fewer search gaps as y increases. □

In words, what we find is that the consumer will prefer to search another option now rather than delay searching when the best option in hand is higher. This is the case because discounting a high value of continued search $W_j(\cdot)$ due to a high value of y is relatively more costly than when y is low, leading to fewer search gaps. This result is surprising and cannot be intuited or derived from existing search models that ignore search gaps. In addition, as we show in Section 5, we find empirical evidence for this effect, suggesting that search gaps can at least in part be explained by the mechanism we propose in our model, i.e. by fatigue.

[9]More precisely, in determining the effect of y on the consumer's behavior, we assume $\bar{S}$, α, and t remain fixed and consider only the local effect, that is the decision to search j now versus after a break as a function of y.

We test the two predictions derived from our model in Section 5 and find evidence supporting them. Before providing this evidence, in the next section we introduce our data and describe the extent of search gaps we observe.

4 Data

4.1 Data Sources and Additions

Our primary source of data comes from GfK, Germany's largest market research company. Among its initiatives, GfK recruits and maintains an online panel of representative consumers for whom online browsing data are collected via a browser extension installed on the panelists' device (PC, smartphone, tablet), recording all their online traffic. In this paper, we have access to the PC browsing history of online panel members in the Netherlands for the period February 15 to May 1, 2018. The data are provided at the level of an exact URL address clicked by a user, and contain information about the user (demographics, such as age and gender), the time of the click, and the website visited. In addition, clicks separated by a period of inactivity of less than 30 minutes are grouped into "sessions". Also, GfK classifies clicks into activities, such as email, social networking, fashion, search engine use, banking, or gaming. Finally, GfK codes the transaction funnel, identifying a website visit, product view, basket addition, checkout, and an order confirmation.

For our analysis, we focused on data for all session visits with at least one click to a fashion website, resulting in 7,877,551 observations with 437,659 fashion clicks. We chose to focus on clicks in the fashion category for two reasons. First, this category is frequently visited by consumers, allowing us to observe enough search activity. Second, we were able to obtain product information from the URLs provided in the GfK data, given that these are stable over time. In contrast, choosing a category such as travel allows us to observe enough search activity, but not to scrape any product information since this changes dynamically and may be personalized, while a category of durable goods only contains searches from a limited set of consumers, which would restrict our analysis given our relatively short observation window.

We augmented the GfK data in a number of ways. First, we scraped product information from the top 50 fashion websites, accounting for more than 46% of fashion clicks, a large percentage given a total of 1,160 websites in our data. This data collection stage occurred within one month of the last observation day in our sample to prevent changes in the web pages accessed. Product information we obtained includes price (current and any promotions), page title, brand name, product name, and when available, product color, reviews, star rating, number of photos, product description, shipping information, speed score of the website (page loading speed), as well as word counts, sentiment on the page, and reading ease.[10] Second, we categorized visits into product versus list pages. Of the 277,784 list and product pages clicked, we have price and product information for 125,922 of them (45%). Third, we identified the product purchased as the last product searched before engaging in transaction related clicks (e.g. adding to cart, checking out, confirming an order). Using this information, we defined a "spell" as all the sessions searched by a consumer before a purchase (or before the end of our observation period if no purchase occurred).[11] Finally, we used URLs, page titles, and the information scraped (e.g. the description of the product) to identify the product category that the consumer searched. We provide more details about our data collection, classification, and cleaning steps in Appendix 7.1.

4.2 Data Descriptives

We observe 3,168 product purchases made by 4,726 consumers in 5,850 spells and 41,665 sessions over nine distinct product categories. Of all spells, 76% contain no purchased product, while of the remaining, 90% contain four of less products purchased, with approximately half purchasing exactly one product. Consumers are 64% female, with an average (median) age of 48 (49), but a large standard deviation (16). Click duration is on average half a minute, slightly higher for fashion clicks and product fashion clicks (0.50 versus 0.53 and 0.57 minutes, respectively).

Table 1 summarizes session and spell characteristics. We find that activity in each session is extensive: on average, consumers make 190 clicks, on 30 websites, and spend more than one hour online. In contrast, fashion search in a session is more modest: the average consumer makes 11 fashion clicks, spends about 5 minutes searching, and visits one product category. Although most clicks are neither on list nor product pages, such pages represent 40% and 23% of all clicks, respectively, since consumers who do click on these pages search several of them. Also, at the

[10]We obtain website speed score information from Google `https://developers.google.com/speed/pagespeed/insights/` and other website features such as word counts, sentiment on the page and reading ease from urlprofiler `https://urlprofiler.com/`.

[11]We borrowed the intuition of De los Santos et al. (2012) and assigned all search before a transaction to a spell.

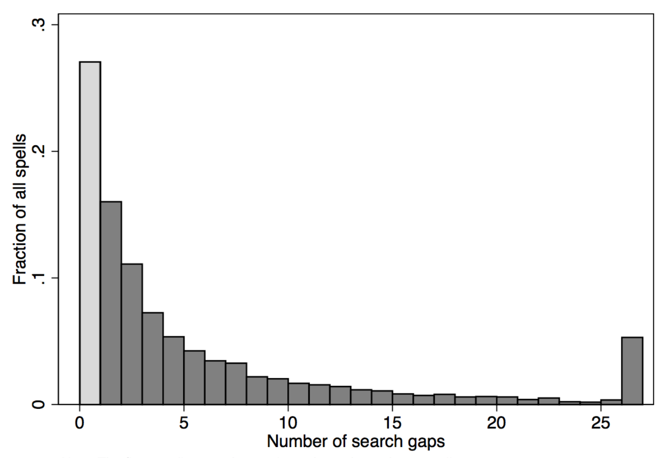
session website level, for consumers who do search at least one product page, they spend 33% of the total time (on average 2 minutes) navigating the homepage or list pages before choosing the first product to search.

Table 1: Session and spell characteristics

	Mean	Median	SD
Session: All clicks			
Clicks	189.07	116.00	250.85
Websites	29.38	20.00	28.44
Duration (minutes)	95.58	64.25	106.11
Session: Fashion clicks			
Clicks	10.50	3.00	26.30
Websites	1.73	1.00	1.45
Duration (minutes)	5.60	1.03	12.22
Product page clicks (percent)	0.15	0.00	0.27
List page clicks (percent)	0.29	0.00	0.36
Product categories	1.11	1.00	1.30
Spell			
Number of search gaps	6.12	2.00	11.45
Length (days)	24.20	16.01	24.91
Search gap (days)	3.94	1.44	6.76

On average, a spell contains 7 sessions, and thus 6 search gaps. Search gaps are very prevalent in the data, as Figure 2 below illustrates. Specifically, 73% of spells contain at least two sessions, and therefore at least one search gap. For this reason, our paper focuses on studying why search gaps occur and how they can be understood from the lens of a search model.[12]

Figure 2: Histogram of search gaps



[12]We also observe when the consumer takes a break within a session, with 19% of all sessions containing such gaps. Nevertheless, our focus remains analyzing search gaps that happen across sessions, since they are more prevalent and can be observed even in data sets less granular than ours, making our implications more relevant to future work.

The most popular activities in our data are email, social networking, and fashion accounting for approximately 25% of all clicks. The most popular websites visited are google.com, live.com, and facebook.com. Table 4 lists the top fashion websites in terms of their transactions. Zalando is the most popular fashion website in our data, and among online retailers in the Netherlands.[13] More precisely, Zalando has more than 40% of transactions in our data (15% of all fashion clicks), followed by H&M with less than 10%. The most commonly purchased product categories are "shirts, tops, & blouses" and "pants & jeans". The category we labeled as "jackets & vests" is the most expensive, with an average transaction price of 60€, while children's, the cheapest category, has an average price of less than 20€.

Table 2: Ordered fashion websites and product categories

Top fashion websites by transactions	Product categories by transactions
1. zalando.nl	1. shirts, tops, & blouses
2. hm.com	2. pants & jeans
3. esprit.nl	3. shoes
4. aboutyou.nl	4. underwear
5. c-and-a.com	5. dresses & skirts
6. your-look-for-less.nl	6. children's
7. adidas.nl	7. sweaters
8. debijenkorf/nl	8. jackets & vests
9. vente-exclusive.com	9. accessories
10. missetam.nl	

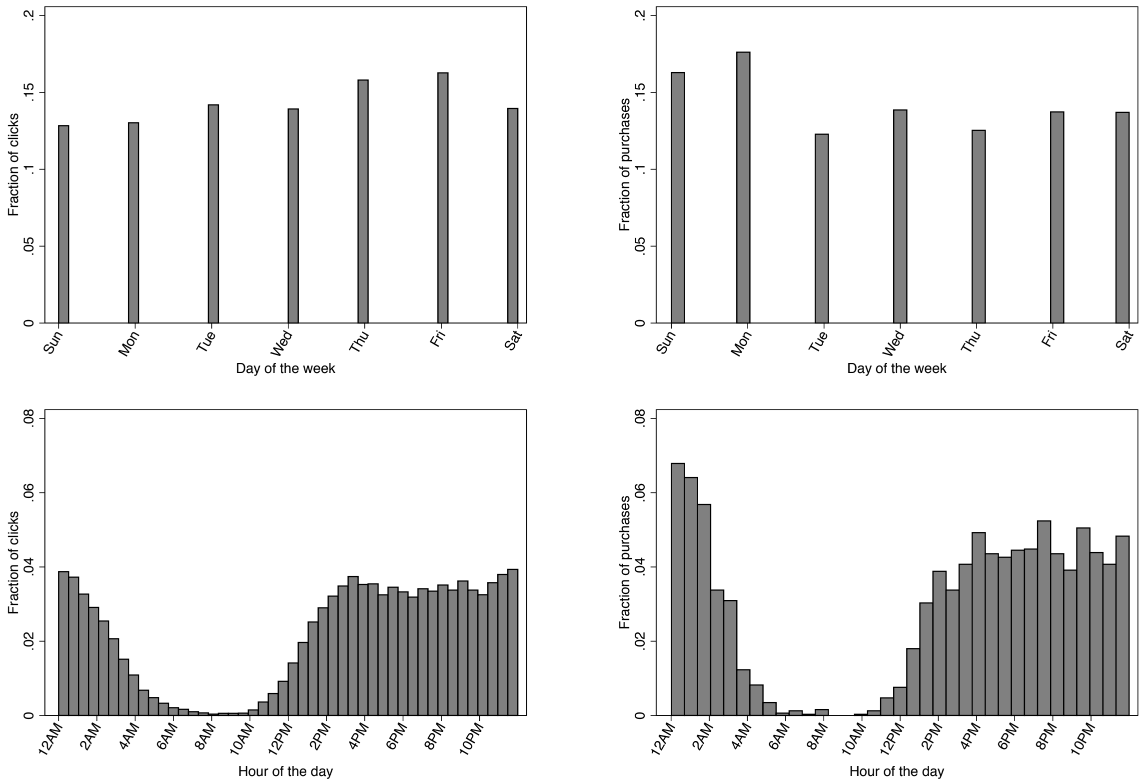
Consumers rarely revisit products they previously clicked, with only 28.7% of all fashion clicks that are revisits. This pattern is consistent with that reported in prior work, for example Bronnenberg et al. (2016) find that about a third of all searches are revisits. One consequence of the low frequency of revisits is that consumers discover the product they buy relatively late in their search spell, also consistent with Bronnenberg et al. (2016). In particular, we find that in spells with one product purchased, this product is first searched at the end of the spell: as the 92% of all clicks in the spell. In addition, since we observe the actual URLs accessed by a consumer, we can further tell whether they click on very similar but not identical products, for example checking a different color or size of the same product. We find that 25% of all clicks are on such a similar product.

Finally, we illustrate when search and purchase decisions are most common. In Figure 3, we plot the percent of clicks and purchases by the day of the week and by the time of the day. The first

[13]For details, see `https://ecommercenews.eu/top-10-online-stores-in-the-netherlands/`.

two panels show that consumers search and purchase decisions are fairly stable across the week, with a slightly lower concentration in the middle of the week. In contrast, search and purchase decisions vary considerably across the day, with hours in the early morning seeing the least activity.

Figure 3: Histogram of searches and transactions



(a) Searches (b) Transactions

5 Why Search Gaps Occur

In this section, we aim to understand what reasons are more likely to explain the occurrence of search gaps in our data. We start by testing the predictions of our model introduced in Section 3. Then, we examine a host of other possible explanations, including the possibility of search gaps not being deterministic.

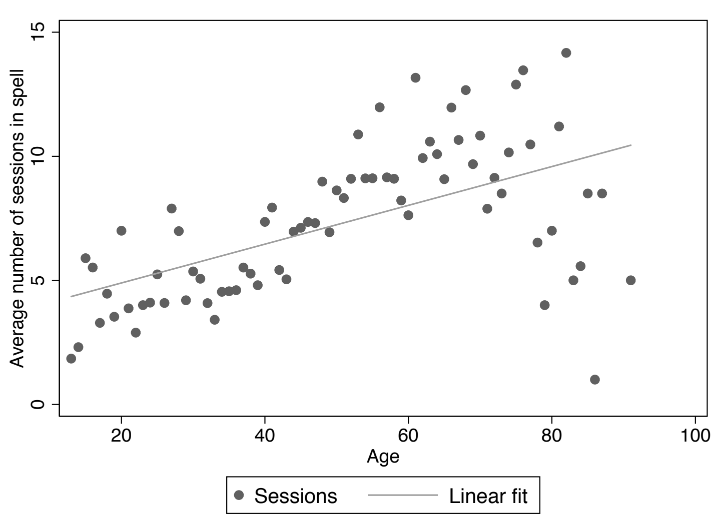
5.1 Empirical Evidence for Search Fatigue Affecting Search Gaps

5.1.1 Test of Proposition 1

Proposition 1 states that a higher fatigue level α increases the likelihood of a search gap. Towards confirming this prediction, we present three pieces of evidence. First, since we do not directly observe the fatigue level of a consumer, we consider a number of proxies and check whether they are related to the number of gaps a consumer has. There is an abundance of medical evidence supporting the idea that mental processing abilities are affected by age, with observed declines in conceptual reasoning, memory, processing speed, and attention to stimuli in older individuals (Harada et al., 2013). These changes in mental abilities can affect decision-making processes in marketing relevant contexts (Peters, 2010; Carpenter and Yoon, 2011). For example, older consumers have been shown to make better decisions when presented with fewer options (Abaluck and Gruber, 2011; Tanius et al., 2009). Also, research shows that older consumers are more likely to use heuristics, to search for a shorter amount of time, and to build smaller consideration sets in order to reduce cognitive effort (Lambert-Pandraud et al, 2005; Kim et al., 2005).

Motivated by this evidence, we consider age as a possible indicator of a consumer's proneness to fatigue and check whether older consumers have more search gaps, as predicted by Proposition 1. For this analysis, we restrict attention to consumers who perform only one spell in the data, to prevent double counting (4022 spells). Figure 4 shows that consumers who are older generally have

Figure 4: Search gaps by age



spells with more sessions and thus more search gaps.[14] This finding is consistent with the idea that age affects fatigue more and thus has an effect on search gaps.

In Table 3, we consider a larger set of factors that may affect the occurrence of search gaps. For this analysis, we define our dependent variable as the percent of search gaps in a spell, that is the number of search gaps observed in a spell as a fraction of all possible gaps (the number of products searched in the spell minus one). We restrict attention to spells with at least two sessions and at least two products searched (and thus at least one search gap). Once again, we find that older consumers generally have more search gaps in a spell. Interestingly, gender also plays a role, with male consumers having more search gaps. This finding is consistent with survey results showing that men can generally become bored of shopping much faster than women.[15] In addition, we find

Table 3

	Dependent variable:	
	Percentage of search gaps in a spell	
	(1)	(2)
Age	0.0022***	0.0015***
	(0.0003)	(0.0005)
Gender: male	0.0477***	0.0373**
	(0.0093)	(0.0168)
Weekend (percent)	−0.0263**	0.0122
	(0.0114)	(0.0201)
Evening (percent)	−0.0113	−0.0105
	(0.0110)	(0.0195)
Speed score	−0.0022***	−0.0015***
	(0.0003)	(0.0005)
Number of images	0.0003***	−0.0001
	(0.0001)	(0.0001)
Readability score	0.0015*	0.0007
	(0.0008)	(0.0012)
All spells	Yes	Subset
Observations	3,143	652
R^2	0.0697	0.0406

Standard errors in parentheses
Notes: Effects of consumer, time, and website variables on the percent of search gaps, both across all spells and in the subset of spells with at least 80% of clicks performed in a single product category.
*p<0.1; **p<0.05; ***p<0.01

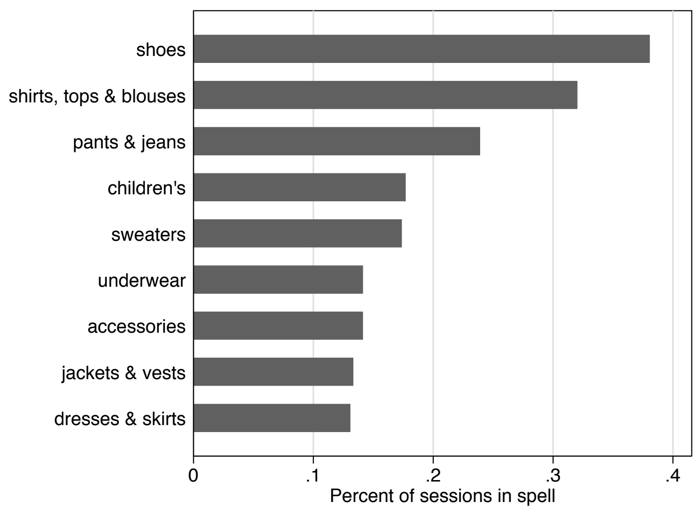
[14] This effect is persistent for most consumers, except those 80 years or older.

[15] Article available at `https://www.telegraph.co.uk/news/uknews/10161610/Average-male-gets-bored-on-shopping-trip-after-just-26-minutes.html`.

that consumers who predominantly search on the weekend or in the evening (6pm to midnight), when there are likely fewer constraints on their time, have fewer search gaps. Finally, consumers who visit websites that are slower to load (lower speed score), have more images, and are harder to read (higher readability score), have more search gaps. If these measures are suitable proxies for consumer fatigue, our results are consistent with Proposition 1, showing that a higher fatigue level leads to more search gaps. In contrast, no such relationship between fatigue and search gaps should exist if search costs were constant across alternatives.

Second, we consider whether search gaps differ by product category. Intuitively, if Proposition 1 is correct, consumers searching in categories that are harder to search should have more search gaps. In Figure 5, we plot the percent of sessions in a spell in which each category is searched.

Figure 5: Search gaps by product category



If a category appears in more sessions in a spell it means it is searched with more gaps. We find that the category with the most search gaps is "shoes", searched in 38% of sessions in a spell, while a category such as"accessories" is searched in only 14% of sessions and thus has fewer gaps. To understand whether this pattern is related to search fatigue, we then check whether categories with more search gaps are also harder to search. To proxy for the difficulty of search and thus fatigue, we consider the number of websites selling products in each category (because a higher number may make it more difficult for the consumer to find the best alternative) and the average speed score for loading a page on a website in each category (since every search action takes longer). We consider two dependent measures for the number of gaps by category. First, we use the percent of sessions in a spell, as defined above. Second, we focus on spells where consumers searched predominantly in one category (at least 80% of all clicks performed in one category) and report the average number of sessions in those spells. Our results in Table 4 show a positive correlation between search gaps and the number of websites in a category and a negative correlation with the average speed score. In other words, categories that are harder to search generally involve consumers spreading their search across more sessions, and thus having more search gaps, as predicted by our model.

Table 4: Correlation matrix for product categories

	Search gaps measures	
	Percent of sessions in spell	Number of sessions
Fatigue measures		
Number of websites	0.69	0.47
Websites' average speed score	-0.39	-0.71

Our third piece of evidence is presented in Table 5, where we summarize the ratio of fashion related searches in the entire spell (clicks or duration) to the length of the first (or longest) session. We find that the median consumer could have finished all her fashion search in the first session. Similarly, all consumers could have finished all their fashion search in the longest session in the spell. This observation suggests that consumers' stopping decision was less affected by the benefit from continued search (which was relatively high since they returned to search), and more by the cost of searching. This higher cost is consistent with the presence of search fatigue and links it to the occurrence of search gaps.

Table 5: Ratio of fashion related searches to length of:

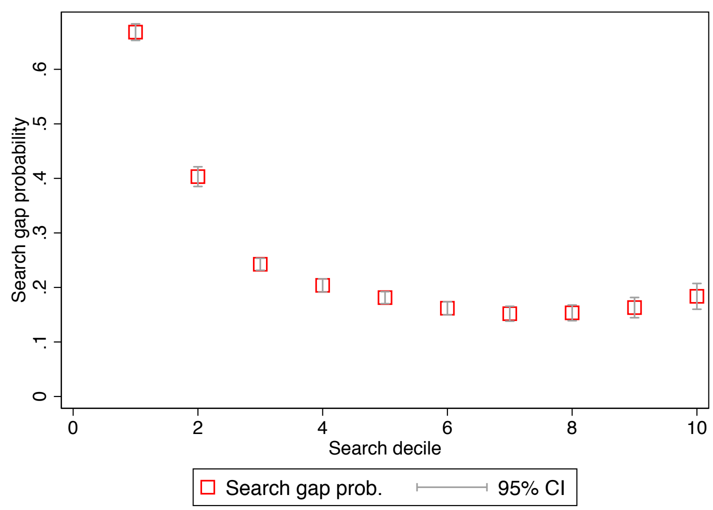
	Mean	Median	Pctl(75)
Clicks			
First session	1.76	0.38	1.01
Longest session	0.41	0.24	0.54
Duration			
First session	3.64	0.46	1.43
Longest session	0.36	0.19	0.47

5.1.2 Test of Proposition 2

Proposition 2 states that a higher level of y, the best option observed so far, leads to fewer gaps. This result is conditional on the consumer not wanting to stop, since a higher y also increases the probability of stopping search and making a purchase decision. To test this prediction, we first need to find a suitable proxy for y. We provide two such proxies.

First, we consider how the probability of a search gap evolves as search progresses towards the end of a spell. Our assumption is that as the consumer continues searching, the value of y has to increase since it is defined as the best option observed so far (assuming perfect recall). In this case, Proposition 2 says that we should observe fewer gaps towards the end a search spell.

Figure 6: Search gap probability by progress in spell



In Figure 6, we divide search spells by decile, using the method found in Bronnenberg et al. (2016). More specifically, a decile is defined by $d(t, N) = ceil(\frac{10(t-r(0,1))}{N-1})$, where N gives the total number of searches in a spell, t denotes the specific search, and $r(0, 1)$ is a draw from a uniform distribution $U(0, 1)$. We then consider the probability of a search gap by decile across spells. For example, a spell with 30 total searches and a search gap only after the first click, would have a search gap in decile 1, but none later, while one with 20 total searches and a gap after the first and before the last click, would have search gaps in deciles 1 and 10. Averaging these across spells, we obtain our dependent measure in Figure 6. This figure shows that search gaps are less frequent at the end of the search spell than at the beginning.[16] We also confirm, using a rank ordered logit model, that search gaps are less likely to occur at the end of the spell (coefficient$=-1.7851$, standard error$=0.2031$, LL$=-12,256$, and number of observations$=4,947$). These results are consistent with our prediction that as y increases, we should expect fewer search gaps.

Second, consumers differ in the value of the option searched first. In particular, on the first search some consumers may reveal a more valuable option, while others one less so. Since this value also influences y, it could serve as another proxy. Using this intuition, we then check whether the value of the first search influences the percent of search gaps in a spell (as defined above). We restrict attention to spells with at least two sessions and at least two products searched (and thus at least one search gap). Our results can be found in Table 6. As expected, we find that clicking

Table 6

	Dependent variable:					
	Percentage of search gaps in a spell					
	(1)	(2)	(3)	(4)	(5)	(6)
Popular website (first click)	-0.0721^{***}	-0.0696^{***}	-0.0642^{***}	-0.0633^{***}	-0.0580^{***}	-0.0573^{**}
	(0.0102)	(0.0100)	(0.0110)	(0.0109)	(0.0224)	(0.0225)
Purchased website (first click)		-0.1216^{***}	-0.1089^{***}	-0.1087^{***}	-0.0339	-0.0297
		(0.0239)	(0.0296)	(0.0296)	(0.0618)	(0.0625)
First List price (first list, stand.)			0.0018		-0.0075	
			(0.0046)		(0.0126)	
Min List price (first list, stand.)				0.0031		0.0198^{*}
				(0.0058)		(0.0106)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
All spells	Yes	Yes	Yes	Yes	Subset	Subset
Observations	3,044	3,044	2,492	2,511	447	450
R^2	0.0871	0.1129	0.1275	0.1257	0.0939	0.0988

Standard errors in parentheses
Notes: The value of the option searched first and its effects on the percent of search gaps, both across all spells and in the subset of spells with at least 80% of clicks performed in a single product category. Controls include: consumers' age and gender, average speed score of the websites, the speed score of the first click, ad initiated search spell indicator, average number of images on the websites and the average readability of the website.
*p<0.1; **p<0.05; ***p<0.01

first on a more popular website (defined as having more transactions than the 95th percentile of all fashion websites), or on the website that the consumer ends up purchasing from is correlated with fewer search gaps. Also, if the first list page clicked contains more expensive products (that is

[16] Our results are robust to conditioning on spells with a transaction or including not only product clicks, but also clicks on the list page of a website. In addition, the same patterns as in Figure 6 hold if we instead divide search spells into tertiles or quintiles. Analysis available upon request.

a higher minimum price), the spell contains more search gaps. In other words, we again find that a higher y, as proxied by a more valuable first click, is correlated with fewer search gaps, a result that is consistent with Proposition 2 in our model.

In this section, we tested our model predictions from Section 3 and provided evidence for search gaps being affected by search fatigue. In what follows, we provide additional tests to show that other reasons for the occurrence of search gaps are less common in our data application, emphasizing again the role of fatigue.

5.2 Tests of Alternative Hypotheses for the Occurrence of Search Gaps

Our alternative hypotheses for the occurrence of search gaps are based on prior work on choice deferral. Specifically, we adapt to a search context and directly test several of the hypotheses uncovered by Greenleaf and Lehmann (1995) for choice deferral: delay due to an expectation of future changes in prices or product features, due to uncertainty regarding what other options to consider, due to a limited budget of time, or due to an inability to afford a purchase.[17] We also test the hypothesis that search gaps occur for reasons unrelated to the consumer search process, making search gaps not deterministic. Finally, given our data application, we add the possibility of consumers switching between sub-categories during a search gap. In total, we identify six possible alternative reasons for the occurrence of search gaps. Since we acknowledge that multiple reasons may affect the occurrence of search gaps, our goal is not to rule out all of these hypotheses (although we provide evidence against several of them). Rather, we aim to show that search gaps occur even absent of these reasons, allowing us to rule in the effect of fatigue based on the evidence presented in the previous section.

1. Search-Unrelated Factors

The consumer may stop searching in the focal category and later restart because her search activity is interrupted by an event that is external to her decision making in that category and thus unrelated to her search. For example, the consumer may stop searching to attend to a work email

[17] The remaining hypothesis described in Greenleaf and Lehmann (1995), related to consumers' perceived risk of the purchase, their desire to consult others before a purchase, their uncertainty related to their need for the product, and the possible availability of a substitute product at home, are either not applicable in our empirical setting or cannot be tested with our current data.

or participate in a planned offline activity. In such cases, search gaps would not be deterministic. The fact that consumers' product search is generally only one among several online daily activities (for example, only approximately 6% of our observations are on fashion clicks) could support this hypothesis.

To test this hypothesis, we exploit the fact that we observe not only search in a focal category, but also all web behavior of consumers. One feature of these data is particularly relevant. Specifically, we observe when a notification (e.g. announcing a new email), which is arguably exogenous to the consumer search activity for fashion products, interrupts search and how consumers react to this event. In our data, notifications account for $107,400$ observations. We find that 91.6% of consumers who get a notification after searching in the fashion category, return to searching fashion in the same session, with 95.2% of them returning within the next five clicks. Since notifications typically direct consumers' attention to a new email they received, most consumers' first activity after a notification is to check their email. However, most consumers quickly return to searching in the fashion category in the same session. Thus, we find that search gaps across sessions cannot be completely explained by the interruption of search activity by a factor that is unrelated to search.

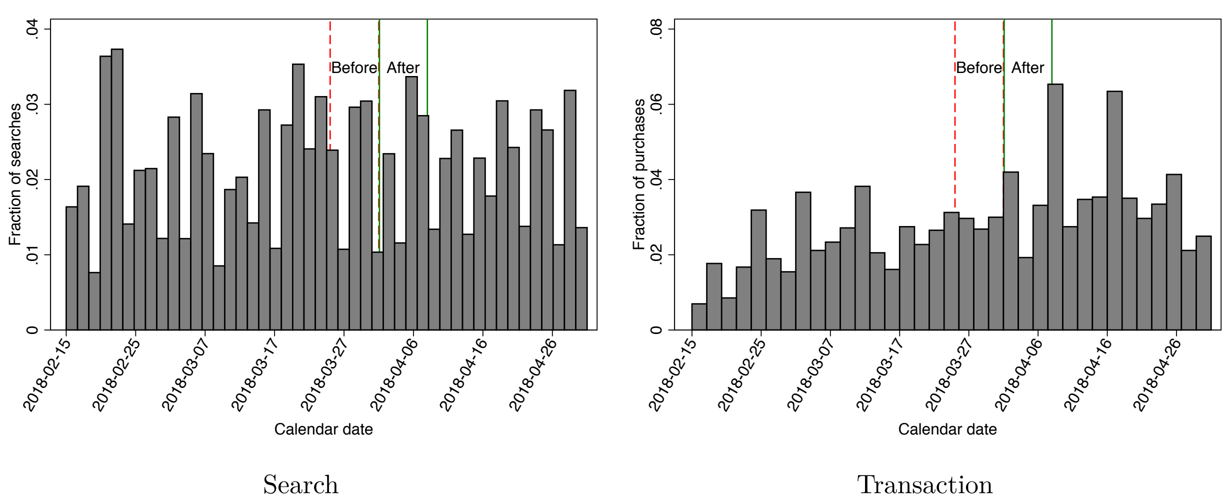
2. Delay Search Expecting a Change in Product Features

Consumers may delay their search because they expect prices to decrease or other product features to improve. This reason is particularly pertinent in a category such as travel, where airfare and hotel prices change frequently and dynamically in response to changes in demand and the available supply of options. Thus, in such categories, we would expect search gaps to occur in part because consumers expect such changes in product features.

However, in the fashion product category, price and product feature changes are less frequent. Price changes or sales occur most often around holidays, and product features change mostly every season. In contrast, the average (median) search gap in our data is only 4 (1) days long, making it unlikely that most search gaps occur because consumers expect prices or other features to change. Nevertheless, we proceed to testing this hypothesis more formally.

Since our data collection on product information occurred within one month of the last observation date in our sample, we do not directly observe any price or product changes. More precisely, although we can tell whether a product is on sale at the time of our data collection, we do not know when the sale began and thus whether consumers observed the sale. Thus, we cannot directly test the hypothesis that consumers await price or product changes in delaying their search. However, we attempt to test both possibilities indirectly. First, if search gaps occur in part because consumers delay their search expecting a price reduction, then we should see consumers searching more right before a holiday than at other times, when such price promotions are more frequent. We observe a major holiday within our observation period: Easter occurred on April 1, 2018. We then look at whether consumers make more searches (i.e. begin more sessions) or more transactions the week before this holiday than the week after Easter. Figure 7 shows our result. The bars indicate the fraction of searches and transactions on each day in our sample, while the interrupted and continuous vertical lines demarcate the week before and after Easter, respectively. As can be seen readily from the figure, there does not seem to be evidence for consumers searching or purchasing more during the week before Easter than after. In addition to this visual evidence, we also do a t-test comparing the fraction of transactions one week before versus one week after Easter, and find no significant difference ($t = -0.7849$, difference $= -.0241$). Also, although there are slightly more searches one week before rather than one week after Easter, the difference is small ($t = 3.5235$, difference $= 0.0120$).

Figure 7: Histogram of fashion searches and transactions by calendar date



Second, it is possible that consumers delay their search expecting a change in product features. In the fashion category, features change mostly every season and there are two major seasons yearly: Fall/Winter and Spring/Summer. The Spring/Summer search starts in January and runs until around June, and the Fall/Winter season goes from July to December.[18] Thus, since our observation period (February 15 to May 1) falls within a single season, it is unlikely that search gaps can be explained for the reason stated.

So far, we have shown evidence of fatigue affecting search (Section 5.1.1) and of search gaps occurring even when consumers expect no changes in product features. These results imply that search costs reset to a lower level when consumers take a break from searching, since otherwise they would not restart searching. Taken together, these pieces of evidence validate our model.

3. Delay Search When Unsure What Else to Search

Consumers may delay their search because they are unsure what else to search. For example, the consumer may be unsure what other websites sell the products she is interested in. However, in our data, 44.8% of consumer search spells have at least one session where the website searched in one session coincides with the website searched in the following session. Thus, search gaps occur even when the consumer continues searching the same website after a gap. Also, the fact that many consumers continue searching the same website after a gap as before, further supports the idea that fatigue played a role in their decision to defer search.

4. Switch Product Sub-Category

Consumers in our data search for products in the fashion category. However, we have identified nine sub-categories of products, such as shoes or accessories. It is possible that consumers search one sub-category in a session and then switch to a different sub-category after a search gap. In this case, the desire to search for a different type of product might explain the occurrence of search gaps. However, in our data 40.18% of spells contain no switches between product categories when a search gap occurs. Thus, search gaps occur even when no switch between sub-categories happens.

5. Limited Time Budget

It is possible that the consumer stops searching and restarts later because she only has a predetermined budget of time available to allocate to searching in the current session.[19] However, we find

[18]For more details, see `https://www.leaf.tv/articles/when-do-fashion-seasons-start/`.

[19]Note that this reason differs from fatigue because of the predetermined nature of a budget of time.

that consumers' online activity rarely ends when her search in the focal category ends. Specifically, only 10% of sessions end with a fashion click. This suggests that the consumer had more time available to allocate to online activities, but chose not to spend more time searching in the fashion category. Furthermore, the two most popular activities after the last fashion click are email and social networking, accounting for more than 20% of clicks, even when restricting to clicks in the evening (6pm to midnight) or on the weekend. Although we cannot rule out the possibility that the consumer has a predetermined time limit on fashion searches specifically, the fact that she chose to spend time on activities that could be considered substitutes to fashion search (or are at least not work related), casts further doubt on this hypothesis. Finally, this hypothesis cannot explain why search gaps are less frequent at the end of the search spell, than at the beginning.[20] We conclude that a limited budget of time is an unlikely explanation for the occurrence of search gaps.

6. Inability to Afford a Purchase

Consumers may delay their search due to an inability to afford a purchase. In this case, we should observe consumers delay their search until they can secure additional funds. However, in our data, search gaps are particularly short, occurring on average less than four days apart, with a median of one day. This evidence suggests that most search gaps are unlikely to occur because consumers cannot afford the items searched. Also, products in the fashion category are relatively inexpensive. For example the average transaction price in the most expensive product category is 60€. Both of these pieces of evidence suggest that most search gaps in our data occur for a different reason that consumers' inability to purchase items searched.

In sum, our findings suggest that in the context of our data, fatigue can at least in part explain the occurrence of search gaps, even when these alternative reasons are not present.

6 Conclusion and Future Directions

In this paper, we investigate possible reasons for the occurrence of search gaps, that is delays in the consumer search process characterized by the choice to search a number of products, to stop, and then to restart the product search. Using a data set on consumers' entire browsing history,

[20] We thank Daria Dzyabura for this observation.

we find empirical evidence consistent with the existence of search fatigue: the more the consumer searches, the higher her search costs per option; taking a break reduces these costs and enables the consumer to restart her search at a later time. We then develop a model of sequential search that allows for search gaps by proposing that a consumer faced with increasing search costs may decide not only whether to continue searching, but also when to search a product (with or without a delay). Finally, we test and validate the predictions of this model, as well as test other possible explanations, further supporting the claim that fatigue can at least in part explain the occurrence of search gaps in our data.

There are several potentially useful extensions of our approach. First, for different product categories, search gaps may occur for other reasons than fatigue. The empirical tests we develop in this paper can provide a starting point for identifying the mechanism behind search gaps in such settings. Second, observing search gaps could allow empiricists to better quantify consumer preferences and search costs by accounting for the additional choice that consumers make while searching, a choice formalized by our model. Third, future work could consider the important aspects of consumer learning and forgetting and their role in affecting search gaps. Another possible extension would be to look at what explains the length of search gaps, in addition to their occurrence. We leave these and other related topics to future research.

References

[1] Abaluck, J., J. Gruber. 2011. Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program. *American Economic Review* **101**(4) 1180-1210.

[2] Anderson, S., R. Renault. 2018. Firm Pricing with Consumer Search. *Handbook of Game Theory and Industrial Organization* eds. Corchon L., Marini M. **2** Edward Elgar, 177-224.

[3] Arbatskaya, M. 2007. Ordered search. *The RAND Journal of Economics*, *38*(1), 119-126.

[4] Armstrong, M., J. Vickers, J. Zhou. 2009. Prominence and consumer search. *The RAND Journal of Economics*, *40*(2), 209-233.

[5] Armstrong, M., J. Zhou. 2011. Paying for prominence. *The Economic Journal*, *121*(556), 368-395.

[6] Baye, M., J. Morgan, P. Scholten. 2006. Information, Search, and Price Dispersion. *Handbook of Economics and Information Systems* **1** 323-377.

[7] Bednar, S., K. Rouse. 2019. Does Taking a Break Make Students (un)Productive? Working paper.

[8] Bronnenberg, B., J. Kim, C. Mela. 2016. Zooming In on Choice: How Do Consumers Search for Cameras Online? *Marketing Science* **35**(5) 693-712.

[9] Carlin, B., F. Ederer. 2018. Search fatigue. *Review of Industrial Organization*. Forthcoming.

[10] Carpenter, S., C. Yoon. 2011. Aging and Consumer Decision Making. *Annals of the New York Academy of Sciences* **1235**.

[11] Chen, Y., C. He. 2011. Paid placement: Advertising and search on the internet. *The Economic Journal*, *121*(556), 309-328.

[12] Chen, Y., S. Yao. 2016. Sequential Search with Refinement: Model and Application with Click-Stream Data. *Management Science* **34**(4) 606-623.

[13] De los Santos, B., A. Hortaçsu, M. R. Wildenbeest. 2012. Testing Models of Consumer Search using Data on Web Browsing and Purchasing Behavior. *American Economic Review* **102**(6) 2955-2980.

[14] De los Santos, B., S. Koulayev. 2017. Optimizing Click-through in Online Rankings for Partially Anonymous Consumers. *Marketing Science* **36**(4) 542-564.

[15] Dhar, R. 1997. Context and Task Effects on Choice Deferral. *Marketing Letters* **8**(1) 119-130.

[16] Gabaix, X., D. Laibson, G. Moloche, S. Weinberg. 2006. Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model. *American Economic Review* **96**(4) 1043-1068.

[17] Greenleaf, E., D. Lehmann. 1995. Reasons for Substantial Delay in Consumer Decision Making. *Journal of Consumer Research* **22**(2) 186-199.

[18] Haan, M., J. Moraga-Gonzalez, V. Petrikaite. 2018. A model of directed consumer search. *International Journal of Industrial Organization*, forthcoming.

[19] Harada, C., M. Natelson Love, K. Triebel. 2013. Normal Cognitive Aging. *Clinics in geriatric medicine* **29**(4) 737-752.

[20] Hodgson, C., G. Lewis. 2018. You Can Lead a Horse to Water: Spatial Learning and Path Dependence in Consumer Search. Working paper.

[21] Hong, H., M. Shum. 2006. Using Price Distributions to Estimate Search Cost. *RAND J. Econom.* **37**(2) 257-275.

[22] Honka, E. 2014. Quantifying Search and Switching Costs in the U.S. Auto Insurance Industry. *RAND J. Econom.* **45**(4) 847-884.

[23] Honka, E., P. Chintagunta. 2017. Simultaneous or Sequential? Search Strategies in the U.S. Auto Insurance Industry. *Marketing Science* **36**(1) 21-42.

[24] Kim S., D. Goldstein, L. Hasher, R. Zacks. 2005. Framing Effects in Younger and Older Adults. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences* **60** 215-218.

[25] Kim, J. B., P. Albuquerque, B. J. Bronnenberg. 2010. Online Demand under Limited Consumer Search. *Marketing Science* **29**(6) 1001-1023.

[26] Kim, J. B., P. Albuquerque, B. J. Bronnenberg. 2017. The Probit Choice Model Under Sequential Search with an Application to Online Retailing. *Management Science* **63**(11) 3531-3997.

[27] Koulayev, S. 2014. Search for Differentiated Products: Identification and Estimation. *RAND J. Econom.* **45**(3) 553-575.

[28] Lambert-Pandraud R., G. Laurent, E. Lapersonne. 2005. Repeat purchasing of new automobiles by older consumers: empirical evidence and interpretations. *Journal of Marketing* **69** 97-103.

[29] Levav, J., M. Heitmann, A. Herrmann, S. Iyengar. 2010. Order in product customization decisions: Evidence from field experiments. *Journal of Political Economy*, *118*(2), 274-299.

[30] Moraga-Gonzalez, J., M. Wildenbeest. 2008. Maximum Likelihood Estimation of Search Cost. *European Economic Review* **52** 820-848.

[31] Moraga-Gonzalez, J.L., Z. Sandor, M. Wildenbeest. 2015. Consumer Search and Prices in the Automobile Market. Working paper.

[32] Morgan, P., R. Manning. 1985. Optimal Search. *Econometrica* **53**(4), 923-944.

[33] Novemsky, N., R. Dhar, N. Schwarz, I. Simonson. 2007. Preference Fluency in Choice. *Journal of Marketing Research* **44**(3) 347-356.

[34] Ozdenoren, E., S. Salant, D. Silverman. 2008. Willpower and the Optimal Control of Visceral Urges. Working paper.

[35] Peters, E. 2010. Age Related Changes in Decision Making. *In: Drolet A, Schwarz N, Yoon C, editors. The Aging Consumer: Perspectives from Psychology and Economics. Taylor and Francis; London* 75-101.

[36] Petrikaite, V. 2018. Consumer Obfuscation by a Multiproduct Firm. *The RAND Journal of Economics*, *49*(1), 206-223.

[37] Rhodes, A. 2011. Can Prominence Matter Even in an Almost Frictionless Market? *The Economic Journal*, *121*(556), 297-308.

[38] Schmittlein, D., D. Morrison, R. Colombo. 1987. Counting Your Customers: Who Are They and What Will They Do Next? *Management Science* **33**(1) 1-24.

[39] Seiler, S. 2013. The Impact of Search Costs on Consumer Behavior: A Dynamic Approach. *Quantitative Marketing and Economics* **11**(2) 155-203.

[40] Stigler, G.J. 1961. The Economics of Information. *Journal of Political Economy* **65**(3) 213-225.

[41] Tanius B., S. Wood, Y. Hanoch, T. Rice. 2009. Aging and Choice: Applications to Medicare Part D. *Judgement and Decision Making* **4** 92-101.

[42] Ursu, R. 2018. The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions. *Marketing Science* **37**(4) 530-552.

[43] Ursu, R., D. Dzyabura. 2019. Retailers' Product Location Problem with Consumer Search. *Quantitative Marketing and Economics.* Forthcoming.

[44] Vohs, K., R. Baumeister, B. Schmeichel, J. Twenge, N. Nelson, D. Tice. 2008. Making Choices Impairs Subsequent Self-Control: A Limited-Resource Account of Decision Making, Self- Regulation, and Active Initiative. *J. Personality and Soc. Psychology* **94** 883-98.

[45] Weitzman, M. 1979. Optimal Search for the Best Alternative. *Econometrica* **47**(3) 641-654.

[46] Wu, C., H. Che, T. Chan, X. Lu. 2015. The Economic Value of Online Reviews. *Marketing Science* **34**(5) 739-754.

[47] Yang, L., O. Toubia, M. De Jong. 2015. A Bounded Rationality Model of Information Search and Choice in Preference Measurement. *Journal of Marketing Research* **52**(2) 166-183.

[48] Zhou, J. 2011. Ordered search in differentiated markets. *International Journal of Industrial Organization*, *29*(2), 253-262.

7 Appendix

7.1 Construction of the Final Data Sample

The raw data contain information on the user (including demographics, such as age and gender), session, and time of the click, as well as the website name and entire URL address of the website visited. Also, GfK coded the transaction funnel, identifying a website visit, a product view, a basket addition, a checkout, or an order confirmation.

Data Augmentation: Scraping

Using the full URLs provided, we scraped the top 50 fashion websites, ranked by the number of clicks. These websites account for over 46% of fashion clicks in our data. This data collection stage occurred within one month of the last observation day in our sample to prevent changes in the web pages accessed. The information we gathered by scraping contains:

1. Price
2. Price promotion (if any)
3. Page title
4. Brand name
5. Product name
6. If available: product color, reviews, star rating, number of photos, product description, shipping information, speed score of the website (page loading speed), as well as word counts, sentiment on the page, and reading ease.

Data Classification: List and Product Pages

In this paper, we consider two types of clicks: clicks on list versus product pages. An example of a list page is "https://www.adidas.com/us/women-originals-shoes", where consumers can see a list of shoes along with a photo, the product name and its price. If a consumer clicks on a product in this list, she navigates to that product's page. An example of such a product page could be “https://www.adidas.com/us/adidas-sleek-shoes/EE4723.html” for a consumer who clicked on the product “adidas sleek shoes” on the list page. The product page contains more detailed information about the product, such as a product description, additional photos, reviews, and so on. Of the 437,659 fashion clicks in our data, 277,784 of them are on list or product pages. The rest we labeled as “others” to represent clicks to the homepage of a website, account pages, or any transaction related pages, such as the cart page.

To categorized clicks into either product, list, or other pages, we performed the following steps. First, during data scraping, we were able to identify list and product pages by examining whether there was any product information available on the page and if so, how many products were available on the page. Second, we used the following rules:

- ‘Other’ pages:
    1. Pages labeled as ‘Add to Basket’, ‘Start Checkout’ and ‘Order Confirmation’ by GfK
    2. The homepage of a website, such as ‘www.zara.com’
- ‘Product’ page:
    1. Pages labeled as ‘Product View’ by GfK
    2. Pages from which we can scrape a single product’s information
    3. URLs that contains product SKU or product IDs (rules differ for each website)
    4. URLs with specific keywords such as ‘product-view’ or ‘shop-by-item’
- ‘List’ page:
    1. Pages from which we can scrape information on multiple products
    2. Pages that display search results, for example, ‘https://www.adidas.com/us/search?q=redshoes’
    3. URLs with specific keywords that indicate their function as a list page, such as ‘shop-per-categorie’, ‘page=’, or ‘category=’
    4. URLs with specific keywords that indicate the sorting or filtering function available on the page, such as ‘price_max=’, ‘productsoort’, ‘pagenumber’ or ‘filter=’

We further categorized URLs labeled as ‘others’ by manually checking them and hired an RA to also independently manually check our categorization.

Data Classification: Product Categories

To identify the product category searched, we used URLs, page titles, and the information scraped (e.g. the description of the product) to search for keywords identifying nine broad categories (as defined on the most popular website in our data, Zalando): accessories, children's, dresses & skirts, jackets & vests, pants & jeans, shirts & tops and blouses, shoes, sweater and underwear. The keywords used to identify the product categories include, but are not limited to:

- **accessories**: 'accesso', 'sjaal', 'lippen', 'earr', 'necklace', 'jewelry', 'bracelet', 'bag', 'eastpak', 'hals', 'banden'
  - with exception of: 'brand', 'bracelet', 'braad', 'brax', 'dirk', 'brace', 'overnachtingen', 'aangebrachte'
- **children's**: 'jongens', 'kinder', 'meisjes', 'baby', 'tiener', 'kids', 'boys' 'girls'
- **dresses & skirts**: 'roecke', 'jurken', 'dress', 'jumpsuit', 'jurkje', 'jurk', 'rok'
- **jackets & vests**: 'trench', 'jack', 'fleece', 'blazer', 'mantel', 'coat', 'parka', 'tussenjas', 'winterjas', 'jas'
- **pants & jeans**: 'hosen', 'broek', 'jogger', 'tights', 'shorts', 'sweatpant', 'pants', 'pantalon', 'leggin', 'trouser', 'tregging', 'jegging'
  - with exception of: 'brand', 'bracelet', 'braad', 'brax', 'dirk', 'brace', 'overnachtingen', 'aangebrachte'
- **shirts, tops & blouses**: 'top', 'hemden', 'langarm', 'kurzarm', 'blusen', 'shirt', 'singlet', 'blouse', 'blouson', 'polo-', '-polo', 'polos', 'longsleeve', 'overhemd', 'onderhemd'
  - with exception of: 'topseller', 'topic', 'topbox', 'topgear', 'topdeals', 'topper', 'topman', 'sniztop', 'marc-c-polo', 'topcom', 'topbloemen', 'laptop'
- **shoes**: 'schuhe', 'stiefel', 'schoen', 'shoe', 'sneaker', 'sandal', 'birkenstock', 'fitflop', 'teva', 'footwear', 'e-walk', 'ecco', 'gabor', 'instappe', 'pumps'
  - with exception of domain names that contain the word 'shoe'

- **sweater**: ‘parka’, ‘hoodie’, ‘poncho’, ‘westen’, ‘trui’, ‘capuchon’, ‘pullover’, ‘tuniek’, ‘vest’, ‘cardigan’, ‘sweater’, ‘jumper’
- **underwear**: ‘thong’, ‘nightwear’, ‘bra’, ‘lingerie’, ‘sleep’, ‘swim’, ‘badpak’, ‘ondergoed’, ‘underwear’, ‘panties’, ‘sock’, ‘sok’, ‘bustier’, ‘push-up’, ‘boxer’, ‘badmode’, ‘bikini’, ‘tanga’, ‘tankini’

Data Classification: Activities

GfK classified clicks into activities, such as ‘Fashion’, ‘Social Networking’, or ‘Web Search’. We used this classification as well as a number of other rules that we detail below to further identify the type of online activity the consumer is engaged in. This resulted in 10 categories defined using the following rules:

1. Fashion:
    - GfK’s classification as ‘Fashion’
2. Search engine:
    - GfK’s classification as ‘Web Search’
    - URLs that contain the keyword ‘search’ when the website visited is Google, Yahoo or Bing.
    - URLs where the website is ‘ask.com’
3. Email
    - GfK’s classification as ’Communication’
    - URLs that contain keywords ‘mail.google’, ‘outlook’, and ‘webmail’.
    - URLs that contain the keyword ‘mail’ when the website visited is Google, Yahoo or Bing
    - URLs that contain the keyword ‘messenger’ when the website visited is Yahoo
4. Social Networking
    - GfK’s classification as ’Social Networking’

- The website visited is one of the 5 major social media platforms: facebook, pinterest, twitter, instagram, linkedin

5. Banking

- GfK's classification as 'Money Management'
- The website visited is or contains 'rabobank.nl', 'abnamro.nl', 'bank', 'achmea' or 'vanlanschot'

6. Cashback

- The website visited is one of: 'geldrace.nl', 'geldkoffer.info', 'geldwolf.info', 'zinngeld.nl', 'mailbeurs.nl', 'extraeuro.nl', 'centmail.nl', 'cashhier.nl', 'spaar4cash.nl', 'snelverdienen.nl', 'ipay.nl', 'spaaractief.nl', 'nucash.nl', 'myflavours.nl', 'directverdiend.nl', 'dieselmail.nl', 'spaar4cash.nl', 'dutcheuro.nl', 'extraeuro.nl', 'cashparadijs.nl', 'sneleuro.nl', 'myclics.nl', 'spaar-voor-euries.nl', 'jiggy.nl', 'qlics.nl', 'quidco'

7. Surveys

- The website visited is one of: 'gfk.com', 'ssisurveys.com', 'focusvision.com', 'opinionbar.com', 'globaltestmarket.com

8. Media

- GfK's classification as 'Media Broadcasting' or 'Media On-Demand'
- URLs that contain the keyword 'tvgids'

9. Google exclude

- URLs from 'google.com' that are not classified as search engine or email related (this includes Google Drive, Maps, etc)

10. Gaming

- GfK's classification as 'Gaming'
- URLs containing the keywords: 'casino', 'game', 'unibet', 'nederlandseloterij.nl'

Data Cleaning: Removing Non-search Activity

The raw data we obtained from GfK contains 9,531,448 observations. The final data we used in our analysis has 7,877,551 observations. To obtain the final data set, we removed observations in the following cases:

- Consumers use a web browser to open local files on their computers, rather than browse the Internet
- A new tab is opened but no webpage is visited on that tab
- Consumers open web browsers' extensions
- Any URL that does not contain 'ttp'
- Duplicates at the session-time level, where the same URL is clicked more than once at the same time or two different URLs are clicked at the same time. In both cases, we only kept a record of the first click.
- Spells where sessions overlap in time (one instance)
- Spells with a transaction but no clicks on product pages observed (in these rare cases, websites likely offer the option of adding a product to the cart directly from the list page)
- Spells with a transaction and observed product clicks but no product bought
- Spells that end within the first week of our observation (before February 23rd, 2018), since it is likely that such observations are left truncated.