DISCUSSION PAPER SERIES

IZA DP No. 8643

Consumer Search Costs and Preferences on the Internet

Grégory Jolivet
Hélène Turon

November 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Consumer Search Costs and Preferences on the Internet

Grégory Jolivet
University of Bristol

Hélène Turon
University of Bristol
and IZA

Discussion Paper No. 8643
November 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 8643
November 2014

ABSTRACT

Consumer Search Costs and Preferences on the Internet*

We analyse consumers' search and purchase decisions on an Internet platform. Using a rich dataset on all adverts posted and transactions made on a major French Internet platform (PriceMinister), we show evidence of substantial price dispersion among adverts for the same product. We also show that consumers do not necessarily choose the cheapest advert available and sometimes even choose an advert that is dominated in price and non-price characteristics (such as seller's reputation) by another available advert. To explain the transactions observed on the platform, we derive and estimate a structural model of sequential directed search where consumers observe all advert prices but have to pay a search cost to see the other advert characteristics. We allow for flexible heterogeneity in consumers' preferences and search costs. After deriving tractable identification conditions for our model, we estimate sets of parameters that can rationalize each transaction. Our model can predict a wide range of consumer search strategies and fits almost all transactions observed in our sample. We find empirical evidence of heterogenous, sometimes positive and substantially large search costs and marginal willingness to pay for advert hedonic characteristics.

JEL Classification: C13, D12, D81, D83, L13

Keywords: consumer search, revealed preferences, individual heterogeneity, price dispersion, internet

Corresponding author:

Hélène Turon
Department of Economics
University of Bristol
8 Woodland Road
Bristol BS8 1TN
United Kingdom
E-mail: helene.turon-lacarrieu@bris.ac.uk

* We are deeply grateful to PriceMinister.com for giving us access to their data. We thank Ian Crawford, Jose Moraga- Gonzalez and David Pacini for giving us very relevant feedback on this work. We also thank audiences at the 2014 Search and Matching conference in Edinburgh, the 2014 EARIE in Milan and at various seminars for their questions and comments.

1 Introduction

The advent of e-commerce, in particular Internet platforms, was initially presumed to increase competition and thus decrease prices and price dispersion, since it allowed the gathering of information on many potential suppliers at little physical and time cost for the consumer. However, casual observation of trading websites as well as the emergence of rich datasets documenting the variance of prices among adverts and transactions convey a compelling message: price dispersion remains and can be substantial.[1] Two potential explanations have been investigated by the recent literature. First, even when controlling for a very specific product, there is still room for heterogeneity through the condition of the item and/or the characteristics of the seller (reputation, size, etc.). If consumers do care for these characteristics as well as for the product itself, then differentiation persists and can result in price dispersion. Second, the presence of search frictions in the process of aggregating and comparing pieces of information offered by each advert displayed on the screen will further reduce the degree of competition and the scope for the "law of one price" to prevail.

In this paper, we aim to give new insights on the role of consumer preferences and search costs on the Internet by conducting a structural analysis of consumer search and purchase behaviour using administrative data from one of France's largest e-commerce websites (priceminister.com). We show empirical evidence on price dispersion but also on consumer purchase behaviour, which we observe to be sometimes at odds with a hedonic perfect-information model. We then consider a model of consumer search where buyers sequentially direct their search along one dimension of the desired item that is instantly and costlessly available. In our application, this dimension is the price, which is prominent on the website display of adverts. Our theoretical framework allows for a wide range of sampling patterns (such as search by increasing or non-monotonous order of price). An innovative feature of our analysis is that we borrow from the revealed preferences literature to set-identify and estimate our model whilst allowing for heterogeneity in consumer's marginal willingness to pay for hedonic characteristics and search costs. Our estimation results will thus not hinge on distributional assumptions made on these sources of consumer heterogeneity.

Our rich dataset, which comprises administrative data from Price Minister, allows us to observe all transations and all adverts posted on the site. For each transaction, we can gather information on all the adverts that were available at that particular date for the very same product, e.g. a given CD, identified by a barcode. These are the adverts that the consumer saw on his computer screen when searching for this specific CD on this website. These adverts may vary in price but also in other characteristics such as the condition of the item, the seller's reputation etc. We observe substantial price dispersion among adverts and among transactions for the same product. We also find that the consumer often does not buy the cheapest available advert and sometimes chooses an advert that is clearly dominated (in price and in non-price characteristics) by another available advert. These stylized facts motivate our structural analysis which focuses on two objects: consumer preferences for advert price and non-price characteristics and search costs.

We set up a sequential model of directed search (on prices). Recall of previously sampled adverts is permitted, and consumers decide optimally in what order to sample items on offer, when to stop sampling

[1] See e.g. Baye et al. (2004).

and which sampled advert to buy. Consistently with the design of the PriceMinister website, we assume that prices are instantly and costlessly visible, but that consumers must pay a search cost to "sample" i.e. to examine an advert's hedonic characteristics and thus compute the utility they will get from this advert. The search process is thus directed along the price dimension. In an application on web browsing of online stores De Los Santos et al. (2012) argue that a non-sequential search model is a better representation of consumer behaviour. However, their ground to reject the sequential approach is that, in their setting, it leads to consumers always buying from the last store visited –which is counterfactual in their data. In a sequential *directed* search framework, however, this is not necessarily the case, as our results will confirm.

An interesting feature of our model is that it can describe a wide range of sampling patterns. In particular consumers do not necessarily search in ascending price order and/or do not necessarily buy the last item sampled. Some consumers will sample very few adverts, others will almost exhaust all the offers. These different patterns will arise from heterogeneity in search costs and in consumers' marginal willingness to pay (MWP) for the advert's hedonic characteristics. Hence, a consumer's choice set is formed endogenously, depending on his (individual-specific) preferences and search cost. Whilst the optimal stopping rule is often incorporated in consumer search models, this analysis of the optimal search order as a direct consequence of the consumer's preferences within a structural estimation of search costs is the main innovation of our paper.

Two other key features of our setting are that we allow consumers to value non-price attributes of the adverts and that both this taste for characteristics and search costs are allowed to be heterogeneous across consumers. Hong and Shum (2006) estimate a search cost distribution for consumers buying academic bestsellers online, but rule out a consumer's valuation of non-price characteristics of different adverts. They estimate both a sequential and a non-sequential model and find a ten-fold difference between their search costs estimates. Analyzing mutual funds' market shares, Hortaçsu and Syverson (2004) allow consumers to have a (homogeneous) taste for non-portfolio attributes and heterogeneous search costs. In these papers, however, search order is not driven by consumers' preferences. Hortaçsu and Syverson (2004) relate the different sampling probabilities attached to mutual funds to their visibility, which itself is not modelled and results more from mutual funds' marketing efforts than from consumers preferences driving the sampling order.

Our model is directly related to the seminal article by Weitzman (1979) who derived the optimal sequence and stopping rule in a sequential directed search model without learning. Each item offers expected gains over the consumer's outside option and the optimal strategy in a nutshell consists in sampling the item with the highest expected gains over the current outside option –until no such gains remain in the set of unseen items.[2] To take this theoretical framework to the data, we first provide two analytical results. First, we show that the optimal search and purchase rule derived by Weitzman (1979) is equivalent to a set of inequalities on utilities and reservation utilities. Second, we show that the reservation utilities defined above can be written

[2]The setting considered by Weitzman (1979), as well as the vast majority of the consumer search literature, rules out learning. We will also assume that consumers do not update their beliefs after each draw. Allowing for learning in the search process is a challenging task that has been tackled by two recent papers, albeit in a different search setting than ours. De Los Santos et al. (2013) set up a parametric search model with Bayesian learning about the distribution of utilities among offers (for MP3 players) but where search is not directed. Koulayev (2014) also allows for learning in a model where the order of search (for hotels) is imposed and thus exogenous to the consumer's preferences, and with no directed search within a given webpage of offers.

in closed form using a function readily available from the data. This leads to a tractable characterization of the set of parameters consistent with each transaction.

We then follow an estimation approach in the spirit of the revealed preference literature (see Blow et al. (2008), Cherchye et al. (2009), Cosaert and Demuynck (2014)) and use the conditions derived from our theoretical analysis to test whether each parameter value is consistent with each transaction. In particular we do not include a behavioural error term relating either to the consumer's sampling or purchasing choice. Thus, contrary to the existing structural literature on consumer search, our estimation strategy does not impose any restriction on the shape of consumer heterogeneity with respect to search costs or to the MWP for advert hedonic characteristics. We do however need to specify a functional form for the individual utility, without which we would not be able to compute beliefs regarding the joint distribution of prices and hedonic characteristics. As in many revealed preference applications, our approach will only produce bounds on the joint distribution of these two parameters. As we will see, these bounds will still be informative to assess the importance of search costs and consumer preferences for online transactions in our data. We are, to the best of our knowledge, the first to use this empirical approach in the consumer search literature.

Our model fits the data very well. Our benchmark specification can explain 94% of the CD transactions observed on the website in a specific quarter in 2007, which is our benchmark estimation sample. As mentioned above, many transactions are such that the advert sold is dominated by an alternative advert in price and hedonic characteristics. A hedonic perfect information model cannot explain these transactions, whereas our model is able to rationalize 76% of these transactions with reasonable values of the MWP and search costs.

We find that most consumers do care for retailer characteristics –the median of the marginal willingness to pay for a marginal increase in seller reputation is between 1 and 2€. Positive search costs are needed to explain a large fraction (26%) of the transactions observed. We also find substantial consumer heterogeneity in these two dimensions and that search frictions play a larger role when the number of adverts available per transaction increases.

As for search patterns, we find that consumers who face strictly positive search costs buy the first advert that they sample 63% of the time, but it can also be the case that the sold advert is sampled once most of the other adverts have been drawn. Besides, we find that for 7% of transactions with positive search costs, the consumer has carried on sampling after finding the advert that he would eventually buy. This fraction increases with the number of available adverts per transaction.

Since our main estimation targets are demand-side structural objects, namely consumer preferences and search costs, we retain a partial equilibrium analysis and focus on developing a flexible estimation approach whilst allowing for heterogeneity and elaborate search strategies. Naturally, our results trigger questions related to the price-setting behaviour of sellers in view of these search costs and heterogeneous preferences on the consumers' side. Papers incorporating search costs into an equilibrium approach include Zhou (2011), who presents an (exogenously) ordered search model in which firms visited late in the search process enjoy some monopoly power since consumers visiting them do so when they have a low valuation

of the products offered by firms already visited.[3] Janssen and Moraga-Gonzales (2004) also analyze firm behaviour when placed in oligopolistic competition and faced with consumers searching non-sequentially. Moraga-Gonzales and Petrikaite (2013) derive an equilibrium sequential search model where consumers can direct their search towards merging firms depending on their expectations over price. A recent paper by Dinerstein et al. (2014) uses rich data on eBay to estimate an equilibrium non-sequential search model with homogenous consumers and to simulate the effects of changes in the platform's design.

Our paper is organised as follows. We detail our theoretical framework in the Section 2. Section 3 describes our dataset and shows new empirical evidence on price dispersion and search and purchase behaviours. Our empirical strategy is detailed in Section 4. We present in Section 5 our estimation results on consumers' search strategies and on the joint distribution of search costs and preferences across observed transactions. Section 6 concludes.

2 Theory

2.1 The environment

Consider a buyer who wants to purchase one unit of a specific product (a given CD or video game) on an Internet platform. Let $J \geq 1$ be the number of adverts for this product that are currently posted on the platform. Each advert $j \in \{1, J\}$ consists of a price p_j and a vector of characteristics x_j. In our application, x will contain the seller's reputation index, its size, its status (professional or not) or the condition of the good being sold. We assume that the consumer's outside option, i.e. not buying anything, is very low so that one advert is always bought (we will be using data on transactions).[4]

Preferences. Consumers have heterogenous preferences for the set of characteristics x. To capture this, we introduce a parameter γ which has the same dimension as x and is heterogenous in the population of consumers. We assume a very simple form for the consumers' utility function: a consumer with preferences γ buying an advert with price and characteristics (p, x) will derive a utility of:

$$u(p, x, \gamma) = \gamma x - p. \tag{1}$$

For notational convenience, we will sometimes write the utility offered by advert j as u_j.

Search frictions. With search frictions, the consumer may not observe all of the J adverts. We assume that consumers search sequentially, with possibility of recall. This assumption needs to be discussed in light of the literature on the optimality of sequential vs. non-sequential search (see e.g. Morgan and Manning, 1985) as well as of recent empirical papers (De Los Santos et al., 2012). Under the sequential search assumption, consumers decide to draw adverts one at a time, whereas in a non-sequential search environment, they would decide on an optimal number of draws ex-ante. We believe the former to be more realistic to model search

[3] In contrast with our analysis, Zhou (2011) focuses on consumers sampling adverts by increasing order of price.

[4] This last assumption is not problematic for our partial-equilibrium approach where we will go after primitive parameters on the demand side of the market. One should just keep in mind that our results will be over the population of individuals who actually buy a product during our observation period. The selection effects arising from consumers just visiting the website and not buying would be more of an issue in an equilibrium analysis as they would affect sellers' expected profits.

on an Internet platform, which is a different context from the one studied by De Los Santos et al. (2012).[5] We also assume that drawing an advert incurs a search cost $s \geq 0$ which is constant across draws but can be heterogenous in the population of consumers. Drawing an advert means collecting all the relevant information, (p, x) and thus knowing the level of utility offered by an advert. We will denote as $H(\gamma, s)$ the cumulative distribution function (cdf thereafter) of taste parameters γ and search costs s in the population of buyers. This distribution will be the main target of our empirical analysis.

Beliefs. The consumer believes that the J adverts presented to him are independent draws from a joint distribution of prices and characteristics (P, X), denoted F.[6] We assume that consumers' beliefs stay the same during the search process, i.e. we rule out learning. This will allow us to derive a simple optimal search strategy for consumers.[7] We also need to consider the marginal distribution $F(X|P)$, which is what consumers believe to be the cdf of X for a given price P.

Let F^0 denote the cdf of prices and characteristics in the population of all adverts actually posted on the platform (for a given product category) in a given time window. This distribution could follow from sellers' pricing strategies. For instance, sellers could differ with respect to their characteristics x and, given their value of x and consumers' preferences and search strategy, set prices that maximize their expected profit. The resulting distribution, say $F^0(P|X)$, combined with the distribution of seller characteristics $F^0(X)$ would then lead to the observed distribution of prices and characteristics $F^0(P, X)$. In this paper, since we restrict our analysis to a partial equilibrium, we will take $F^0(P, X)$, which we can directly observe in the data, as given.

The last assumption we need before presenting consumers' search strategies pertains to the consumers' beliefs. A natural way to anchor consumers' beliefs would be to assume that $F(P, X) = F^0(P, X)$. In a full-equilibrium setting, this means that consumers' beliefs are consistent with sellers' pricing strategies. From an empirical perspective, this assumption allows us to estimate consumers' beliefs as the observed distribution of prices and characteristics in the population of adverts. However, as we will discuss in detail in Section 4.2, one may impose further restrictions on the beliefs in order to limit the level of sophistication in consumers' predictions.

2.2 Consumers' search and purchase decision

We now describe how consumers search for and buy adverts in the environment we have just outlined. First, consider a case where there are no search frictions, $s = 0$. In this perfect-information model, the consumer chooses an advert in $\{1, J\}$ that offers a utility equal to $\max_{j \in [1,J]} \{u(p_j, x_j, \gamma)\}$. If more than one advert offers this level of utility, the consumer randomly chooses one of them. Since we have assumed that the consumer's reservation utility is very low, the consumer will definitely choose one advert. If all observed transactions could be explained by this model, we would not be able to claim evidence of consumer search frictions. As

[5] De Los Santos et al. (2012) show that the behaviour of consumers looking for books on different websites is not consistent with a sequential random search model, as consumers sometimes buy from a previously visited website. In this paper, we will consider a *directed* search model so sequential search will be consistent with consumers retracing their steps.

[6] We use capital letters for random variables and small letters for their realizations.

[7] Deriving and estimating optimal search strategies with learning is a very challenging task that has recently received attention in economics (see e.g. Koulayev, 2014).

we will see in the empirical analysis, this is not the case. In the following, we allow search costs to take any value and present our preferred model.

A directed search model. We assume that the consumer can see all the available advert prices instantly but has to incur a utility cost s to observe a given advert's characteristics x. This follows from the design of the website used in our empirical application. When consumers are looking at adverts for a given product, they first see all adverts ranked by increasing order of price. The price is shown in a larger font than other characteristics (such as seller reputation etc.). We thus think that it is realitic to assume that collecting information on prices is costless for consumers but that they must pay a utility cost to gather additional information on adverts as these details are less visible and not ranked by default. Consumers can then use advert prices to direct their search.

The search cost can then be thought of in a number of ways, such as a cost of looking at the set of characteristics x or the cost of processing the information given by the new advert examined in the context of the choice optimisation under way. As mentioned above, we allow s to be heterogeneous across consumers, but restrict it to be constant across draws within one consumer's search process, i.e. it is not increasingly (or decreasingly) costly to look at additional adverts as more adverts have already been examined.

Note that if $s = 0$, we have the perfect information model but this is also the case if $\gamma = 0$. Indeed, if a consumer only cares about prices and if information about advert prices is available at no (search) cost, this consumer will look at all the advert prices and buy the cheapest advert, as if there were no search frictions. Hence our directed search model embeds the perfect information case. In the rest of this section, we will thus focus on the case where $s > 0$ and $\gamma \neq 0$.

The optimal search and purchase strategy. We now present the optimal search and purchase strategy used by consumers in this directed search model. To this end, we will use a result from an influential article by Weitzman (1979). In the next section, we will then show how this result can be used to identify consumers' preference and search cost parameters.

Let a consumer's preferences and search cost be given by γ and s respectively. This consumer has to search among J adverts. At any given point of his search we denote as $\tilde{u}$ the best utility drawn so far. If search has not yet begun, $\tilde{u}$ is so low that it makes any draw worthwhile. If this consumer now has to choose between sampling an advert at price p or stopping. Based on his beliefs regarding the distribution of x at this price level, $F(X|P)$, he will choose to sample this advert if the expected utility gain over $\tilde{u}$ given price p is greater than s. Formally, this reads:

$$s \quad < \quad \int_{u(p,x,\gamma)>\tilde{u}} [u(p,x,\gamma) - \tilde{u}]\, dF(x|p). \tag{2}$$

We can now define an important quantity that will drive the consumer's search strategy. We can see with (2) that the expected benefit of drawing an advert decreases as $\tilde{u}$ increases. Hence there exists a threshold level above which it will not be worth drawing an advert at price p. Of course, this threshold will depend on the consumer's characteristics, (s, γ). This determines a threshold "reservation utility" for each price p, preference parameter γ and search cost s, denoted $r(p, s, \gamma)$, and defined as the solution of the following

equation:[8]

$$s \quad = \quad \int_{u(p,x,\gamma)>r(p,s,\gamma)} [u(p,x,\gamma) - r(p,s,\gamma)]\, dF(x|p). \tag{3}$$

$r(p,s,\gamma)$ is the utility level that makes the consumer indifferent between drawing an advert with price p (thus enjoying the attached expected gain and incurring the search cost s) and not drawing it. In other words, it is the minimum level of reservation utility that will make the sampling of p unattractive. It is apparent from (3) that this reservation utility depends on the price p, on the parameters γ and s but also on consumers' beliefs F. We will sometimes denote the reservation utility offered by advert j simply as r_j, instead of $r\left(p_j,s,\gamma\right)$. Note that all consumers share the same beliefs F but are heterogeneous in terms of their personal characteristics (s,γ). The sequence $(r_j)_{j=1..J}$ will thus be individual-specific. Of particular interest is the fact that our model rationalises the search order and that this order may well vary across individuals. This will be illustrated with our data in Section 5.1.

We can now give the optimal sequential search and purchase strategy, as derived by Weitzman (1979). A consumer with personal characteristics (s,γ) and beliefs F about the joint distribution of (P,X) should compute all the reservation utilities of the J adverts presented to him and sort them in decreasing order of r_j. He should then start by drawing the advert with the highest r_j and proceed as follows:

- Let $\tilde{u}$ either be the highest utility offered by the adverts sampled so far or the (very low) value of the outside option if no advert has yet been sampled.

- If $\tilde{u}$ is strictly lower than the highest r among adverts not yet sampled then sample another advert (one with the highest r among the adverts not sampled).

- If $\tilde{u}$ is larger than the highest r among adverts not yet sampled, stop sampling and purchase the best advert drawn so far (one that offers a utility of $\tilde{u}$).

Ties are assumed to be resolved in the following way. If several adverts have the same reservation utility r, consumers sample them in a random order. If several adverts that have been drawn offer the same maximum level of utility, the consumer chooses one randomly. When indifferent between stopping his search and sampling another advert, the consumer stops searching.

This strategy illustrates an interesting feature of sequential directed search models: consumers may go back to adverts previously drawn even though they have not exhausted all offers. This will happen when the utility u offered by a drawn advert, say advert i, is larger than that of all adverts previously drawn but lower than the reservation utilities of adverts not yet drawn. The consumer will then draw these other adverts and, if the maximum utility they offer is lower than that of advert i, he will eventually go back and buy advert i. Hence, the search patterns highlighted by recent empirical papers (for instance De Los Santos et al., 2012) may not be at odds with a sequential search model, provided one allows for directed search (instead of random sampling).

[8] Equation (3) defines one and only one reservation utility as the search cost s is positive or zero and the right-hand side of (3) is a continuous and strictly decreasing function of r which takes values between 0 and $+\infty$.

2.3 Identification of preferences and search costs: a revealed preference approach

In the spirit of the revealed preference literature (see Blow et al., 2008), we now undertake to estimate sets of parameters that are consistent with the choices observed in the data, with no further assumptions on consumer behaviour, particularly with respect to optimisation errors. Consumers are allowed to be heterogeneous with respect to their marginal willingness to pay γ for the hedonic characteristics and with respect to their individual search cost s, but, given these, our model does not include any error term that would rationalise observed choices not quite consistent with the theoretical framework outlined in this section. We will now describe how these sets of parameter values are identified.

Consider a transaction where advert i is sold. We may also refer to this transaction as transaction i.[9] From now on, for each transaction, all quantities (p, x, u, r) will be indexed by i if they refer to the advert bought, and by $j \in J$ if they refer to an advert which was on the screen but was not bought.

We now derive necessary and sufficient conditions for a pair (γ, s) to be consistent with the fact that i was bought while advert j was also available but not chosen. These conditions will characterize a set S_{ij}. We can then define the set S_i of parameters consistent with transaction i as the intersection of all the sets S_{ij} for all available adverts for this transaction.[10]

We now characterize the set S_i. First, we can assess whether a transaction can be explained by a perfect information model:

$$(s = 0, \gamma) \in S_i \quad \Leftrightarrow \quad \gamma x_i - p_i \geq \max_j \{\gamma x_j - p_j\}. \tag{4}$$

In words, if there are no search costs the consumer must choose an advert that yields the maximum utility. The sets of values of γ consistent with this may be empty. Likewise, we can check whether a transaction requires non-zero preferences for the non-price advert characteristics x. An individual preference such that $\gamma = 0$ will be revealed by a behaviour satisfying the following condition:

$$(s, \gamma = 0) \in S_i \quad \Leftrightarrow \quad p_i \leq \min_j \{p_j\}. \tag{5}$$

This means that consumers who only care about prices shoud buy the cheapest advert (since information on prices is available for free).

We now turn to the more challenging cases where consumers face positive search costs and have non-zero preferences for x. We can characterize each S_{ij}. A pair (γ, s), where $\gamma \neq 0$ and $s > 0$, is *not* consistent with i being bought instead of j, that is $(\gamma, s) \notin S_{ij}$, if and only if at least one of two statements is true:

$$\begin{cases} u_i \geq r_i \quad \text{and} \quad r_j > r_i \quad \text{and} \quad u_j \geq r_i, \\ \text{or} \\ u_i < r_i \quad \text{and} \quad r_j > u_i \quad \text{and} \quad u_j > u_i. \end{cases} \tag{6}$$

[9] In a slight abuse of language, we refer to the distribution of parameters across transactions i as the distribution of parameters across purchasing consumers. This is only valid if all consumers buy exactly once, and in the absence of information of consumers' identities in the data, we are not in a position to confirm this.

[10] If the data allowed us to identify consumers, we would be able to narrow the set even further by considering the intersection of all the S_i's pertaining to purchases made by a consumer.

Proof. Start with the case $u_i \geq r_i$. If $r_j < r_i$ then j is not drawn (because $r_j < r_i \leq u_i$) so we cannot reject (γ, s). If $r_j = r_i$ there is a positive probability that i is drawn first, in which case j will not be drawn and we cannot reject (γ, s). If however $r_j > r_i$ then j is drawn before i and i will be drawn only if $u_j < r_i$ (in which case i will also be bought as $u_i > r_i$). We can thus reject (γ, s) when $r_j > r_i$ and $u_j \geq r_i$. This means that i is not drawn, as j is drawn first and its utility u_j is higher than r_i.

Now turning to the case $u_i < r_i$. If $r_j \leq u_i$ then j is not drawn and we cannot reject (γ, s). If $r_j > u_i$ then j will be drawn (either before or after i, depending on r_j vs. r_i). For i to be bought we must then have $u_i \geq u_j$ otherwise we have to reject (γ, s) as j would be drawn (before or after i) and would offer more utility than i.■

Our characterization of S_i thus follows from simple inequalities. For each transaction i and each parameter value (s, γ), we just need to compute the instantaneous and reservation utilities and check whether (6) holds for any advert j. If this is not the case, this parameter value rationalizes the transaction. We can thus follow an empirical approach similar to that used in the empirical revealed preference literature and test for each parameter value whether each transaction is consistent with the model.

2.4 The case of a scalar hedonic index

So far, we have considered a general case where the non-price characteristics of adverts consisted of a vector x. From now on, we will assume that x, and thus γ, is a scalar. Moreover, we will assume that x is valued positively by all consumers so that $\gamma \geq 0$. In this section, we show how we can use this assumption and the results from the previous section to get a more elegant and far more tractable characterization of the sets of identified parameters. Considering a scalar hedonic index will also greatly facilitate the exposition of the results as the sets of interest will now be of dimension 2 (one for s and one for γ).

The interpretation of this assumption is that, for all consumers, the different non-price advert characteristics can be aggregated into a scalar index x. This means that the advert characteristics (such as seller reputation, seller size, etc.) can be projected onto a scalar index and that this projection is the same for all consumers. In other words, consumers all have the same marginal rate of substitution between two non-price advert characteristics. Importantly, we still allow for heterogeneity in the marginal willingness to pay for the hedonic index x as we make no assumption on the distribution of γ. We do however constrain γ to be positive (or zero) but this is not restrictive as, in our data, we can easily find an advert characteristic for which the marginal willingness to pay is unlikely to be strictly negative (for instance seller reputation or product condition). We will show in detail in section 4.1 how we construct a structural projection of advert characteristics onto the scalar hedonic index x.

Transactions with 'better' alternatives With a scalar hedonic index, we can now give some intuition on the sources of information used to identify search costs. We first define a type of transactions that will play an important role in the identification of search costs. Consider a transaction i where an unsold advert j is such that $p_j \leq p_i$ and $x_j \geq x_i$ with at least one of these inequalities being slack. Since $\gamma \geq 0$, advert j is then 'better' than advert i in that $u_j \geq u_i$ for all consumers. This transaction cannot be explained without

a strictly positive search cost, unless $p_i = p_j$ and $\gamma = 0$. Transactions with 'better' alternatives will thus provide information on positive search costs.

Now consider a transaction where there are no 'better' alternatives to the advert sold i. Each set S_{ij} is then non empty and contains $s = 0$ as each comparison between the sold advert i and an alternative j can be explained with a set of marginal wilingnesses to pay γ. However, the intersection S_i of all S_{ij}'s may not contain $s = 0$. This will be the case if two alternatives adverts j and k are such that j is slightly more expensive than i but offers a much larger x, which suggests a low willingness to pay for x, and k offers a slightly lower x but is much cheaper than i, which can only be explained by a high willingness to pay for x. Suppose we have: $0 < \frac{p_j - p_i}{x_j - x_i} < \frac{p_i - p_k}{x_i - x_k}$. Then, without search frictions, γ would have to be smaller than $\frac{p_j - p_i}{x_j - x_i}$ and larger than $\frac{p_i - p_k}{x_i - x_k}$, which is impossible. Hence, the absence of 'better' alternatives may not always imply that the transaction is consistent with a perfect-information model.

A useful function for directed search. We now show how the scalar index assumption can be used to improve on the characterisation the identified sets. In order to obtain simple analytical translations of the inequalities in (6), we introduce the following function:

$$\psi_p(x) \quad = \quad E(X - x | X > x, P = p) \quad = \quad \int_x^{+\infty} (x' - x) dF(x'|p). \tag{7}$$

$\psi_p(x)$ reflects the expected gain over x (the scalar hedonic index) when an item of price p is sampled. An important feature of this function, which will be very useful for identification and estimation, is that it does not depend on (γ, s). The function ψ_p is differentiable and strictly decreasing in x on the support of x given p. We can thus define its inverse ψ_p^{-1}.

Note also that ψ_p is closely linked to consumers' beliefs regarding the distribution of the hedonic index at a given price $F(\cdot|p)$. In particular, if $F(\cdot|p')$ stochastically dominates $F(\cdot|p)$ when $p' > p$, i.e. if consumers believe that a higher price means a better hedonic characteristic x, then $\psi_{p'}(x) \geq \psi_p(x)$.

Now, let a consumer with taste γ have a reservation utility of $\tilde{u}$ reflecting either the utility of not buying anything if the consumer has yet to sample his first item or the best utility found so far if the consumer has already sampled some item(s). At price p, this would be achieved with an equivalent hedonic index of $\tilde{x} = \frac{\tilde{u} + p}{\gamma}$. The quantity $\gamma \psi_p \left(\frac{\tilde{u} + p}{\gamma} \right)$ thus measures the expected utility gain for this consumer of drawing an advert at price p. The reservation threshold $r(p, s, \gamma)$, defined by (3), driving the directed search process for a consumer with preferences summarised by (s, γ) can then be written as:

$$\gamma \cdot \psi_p \left[\frac{r(p, s, \gamma) + p}{\gamma} \right] = s \quad \Leftrightarrow \quad r(p, s, \gamma) = \gamma \cdot \psi_p^{-1} \left(\frac{s}{\gamma} \right) - p \tag{8}$$

where the function $\psi_p(\cdot)$ does not depend on the parameters s and γ. Note that the expression for $r(p, s, \gamma)$ then mirrors the specification for utility, $u(p, x, \gamma) = \gamma x - p$. Also, note that (8) shows that the sampling order may not be a monotone function of the price as the sign of r'_p will depend on s, γ and ψ'_p. This modelling of the sampling order and subsequent estimation is, to the best of our knowledge, a new contribution to the consumer search literature. We will illustrate this in detail in Section 5.1.

A tractable characterization of the identified sets. We can now use the expressions for the utility (1) and the reservation utilities (8) to rewrite the inequalities (6) characterizing the identified sets. If s or γ equals 0, we can still use conditions (4) or (5). If $s > 0$ and $\gamma > 0$, the conditions for (s, γ) *not* to be consistent with the observed transaction are the following:

$$(s,\gamma) \notin S_{ij} \quad \Leftrightarrow$$

$$\begin{cases} \frac{s}{\gamma} \geq \psi_{p_i}(x_i) \quad \text{and} \quad \gamma\left[\psi_{p_j}^{-1}\left(\frac{s}{\gamma}\right) - \psi_{p_i}^{-1}\left(\frac{s}{\gamma}\right)\right] > p_j - p_i \quad \text{and} \quad \gamma\left[x_j - \psi_{p_i}^{-1}\left(\frac{s}{\gamma}\right)\right] \geq p_j - p_i, \\ \text{or} \\ \frac{s}{\gamma} < \psi_{p_i}(x_i) \quad \text{and} \quad \gamma\left[\psi_{p_j}^{-1}\left(\frac{s}{\gamma}\right) - x_i\right] > p_j - p_i \quad \text{and} \quad \gamma(x_j - x_i) > p_j - p_i. \end{cases} \tag{9}$$

The main advantage of (9) compared to (6) is that the conditions are now simple plug-in functions of the parameter values (s, γ). Once we have an estimate of the ψ_p^{-1} functions for each price (and this can be done without looking at consumers' choices), finding the set of parameters consistent with a given transaction can easily be done by a simple grid-search method, using (9) as a pass/rejet criterion.

3 Data and descriptive statistics

3.1 The PriceMinister website

We use data from PriceMinister, a French company organizing on-line trading of new and second-hand products between buyers and professional or non-professional sellers. We will focus on the company's French website www.priceminister.com. PriceMinister is one of the largest e-commerce websites in France with 11 million registered users in 2010 (the site opened in 2001) and over 120 millions products for sale in 2010.[11] Whilst many different items can be bought from the website (books, television sets, shoes, computers), we will focus on CDs and, in a robustness check, on DVDs. The 'cultural' goods (books, CDs, video games and DVDs) represented the vast majority of transactions during our observation period.

The website is a platform where sellers, professional (registered businesses) or non professional (private individuals), can post adverts for goods which can be used or (for professional sellers since 2003) new.[12] When a potential buyer searches for a specific item, the website returns a page of available adverts. These include the price (adverts are sorted by increasing prices by default), the condition of the item: new or used ('as new', 'very good', 'good'), the seller's status (professional or not), reputation and size.[13]

In this paper, we will focus on the consumer's search behaviour once he reaches a page of adverts for a specific product. We do not model how the consumer behaved before he reached this page. We have data on transactions so we know that, for each of those, the consumer must have reached the page of adverts for this product before he made his purchasing decision. Since we impose the standard assumption that the cost

[11] PriceMinister was ranked first among e-commerce websites in terms of ratings in a survey conducted by Mediamétrie in March 2010. The other main e-commerce websites in France are Amazon, eBay and Fnac.

[12] PriceMinister does not charge a sign-on fee, and posting an advert is free of charge. However for each completed transaction, sellers have to pay a variable fee to PriceMinister. The fee scale is posted on the PriceMinister.com website.

[13] The advert also shows the seller's name, country and the different shipping options. In this version of the paper, we do not include sellers' country in the characteristics vector x because it is France for the overwhelming majority of sellers. We will discuss shipping options and costs later on in Section 3.2.

of sampling an advert does not depend on the number of past draws, we can identify the consumer's search cost and preferences for advert characteristics from this last stage.

A seller's reputation is the average of feedbacks received since the creation of the seller's account. To understand the feedback mechanism, we must explain how transactions take place on the website. When a buyer purchases a given product from a given seller, the buyer's payment is made to PriceMinister in the first instance. At this point the seller is informed that a buyer has chosen her product and ships the item to the buyer. Once the buyer has received the product, he is prompted to go on the website and give his feedback on the transaction. PriceMinister then closes the transaction and pays the seller.[14] The buyer's feedback consists of a grade, or rating, which by default is equal to 5. The buyer can change it to any integer between 1 (very disappointed) and 5 (very satisfied).[15] The seller's reputation as posted on the website is the rounded average (to the nearest first decimal) of the feedbacks received for all completed transactions. A seller's size at a given date is then the number of transactions that she has completed so far.

We should mention that PriceMinister differs from other e-commerce websites that are studied in the economic literature with respect to several features that are important for our analysis. First, PriceMinister itself does not sell any products: it is a platform (unlike, e.g., Amazon). Hence consumers may not direct their search towards a seller that also operates the platform. Secondly, prices are posted by sellers, there are no auctions (unlike eBay).[16]

3.2 The dataset

We have two administrative datasets obtained from PriceMinister: one with all the transactions between 2001 and 2008, and one with all the adverts posted between 2001 and 2008. For each transaction or advert, we observe the price, product and seller ID (not the buyer's), and all the characteristics mentioned above (product condition, seller's status, reputation and size). We can thus construct a dataset where, for each transaction, we observe all the adverts that were on the screen for the same specific product when the consumer made his choice.[17] Note that products are precisely identified on the website (for instance by their barcode). Although we do not make use of this for now, we should bear in mind that the buyer can also see information that is not included in the data, for example a line of text accompanying each advert that sellers have the option to post.

In this paper, we will focus on transactions of CDs that took place in the third quarter of 2007. Unless otherwise mentioned, all the following descriptive statistics and estimation results will be produced for this selected sample. At the end of the paper, we will also produce results for other time periods and for DVDs as robustness checks.

For the price variable, we use the advert price net of shipping costs. This is a way of making all prices

[14] When a buyer files a complaint, PriceMinister investigates and puts the payment on hold. If the buyer does not contact PriceMinister within 6 weeks, he is sent a reminder e-mail. If he does not respond, PriceMinister closes the transaction and pays the seller.

[15] The fact that buyers must give feedback in order to validate the transaction ensures a high feedback rate (above 90% for transactions with individual sellers).

[16] In recent years, buyers may be offered the option to negociate the price but this option was introduced at the end of our observation period and, at the time, rarely used.

[17] The construction of the dataset with all live adverts at each transaction date hinges on some assumptions which we present and discuss in Appendix A.

comparable. On PriceMinister, sellers cannot differentiate themselves with respect to shipping costs. In any transaction, the choice of shipping mode (essentially, standard or registered mail) is up to the buyer, subject to a fixed shipping cost scale imposed by PriceMinister. Specifically, the buyer chooses a particular shipping option at the time of purchase and the corresponding fee on the shipping cost scale is added to the bill and transferred to the seller by PriceMinister. It is then up to the seller to minimize its costs, subject of course to complying with the buyer's specific choice of shipping mode. In theory, sellers could still differentiate themselves by offering a specific type of shipment that may not be offered by other sellers. Unfortunately this information is not available in our dataset. We presume that, given the menu of shipment choices made available by default within the PM system, there is little incentive for an individual seller to offer yet another choice, especially for the category of products we study (CDs).

3.3 Descriptive statistics

As mentioned above, the estimation sample is taken from the third quarter of 2007. The reason why we restrict ourselves to a short time window is that the site has known a rapid growth rate and our estimation strategy assumes that beliefs regarding the joint distribution of characteristics and prices are constant.

For the same reason, we use data on sales of CDs with a catalog price ranging from 10 to €25 (hence leaving out EPs, CD singles or collector CDs). We discard transactions for which there was only one advert posted on the screen, as well as transactions for which one of the posted adverts had a price outside the range €1-20 (7% of adverts have a price outside this interval). This leaves us with 77,753 transactions, involving 23,538 sellers, 25,818 products and 145,823 adverts.[18]

The distribution of the number of adverts by transaction can be seen in Table 1. We note that the majority of transactions were made while there were few (less than 5) adverts available but there are also many transactions for which the consumer had to choose from a large number of adverts. Recall that the search cost and preference parameters are allowed to be heterogenous across transactions so our estimation results will not be driven by the number of transactions with few adverts. Moreover, we will break our estimation results down by the number of adverts per transaction so that we can assess whether search frictions increase as there are more adverts on the screen.

[18] Any statistics based on the population of adverts will be produced on a sample where each advert is counted only once, at the first time when it appears in a transaction (whether it is sold or not).

Table 1: Distribution of the number of posted adverts per transaction

# adverts	Frequency	Percentage	Cumulated Percentage
2	19,248	24.76	24.76
3	13,234	17.02	41.78
4	9,597	12.34	54.12
5	6,975	8.97	63.09
6	5,415	6.96	70.05
7	4,106	5.28	75.33
8	3,244	4.17	79.51
9	2,645	3.40	82.91
$[10, 19]$	10,545	13.56	96.47
≥ 20	2,744	3.53	100.00
Any	77,753	100.00	100.00

Even though each advert or transaction refers to a specific CD (as defined by its barcode), we observe substantial price dispersion, in both the populations of adverts and of transactions. Table 2 shows that the average number of advert prices per product is above 5 (see last row). Comparing the first and the third columns, we also note that, for a given product, there are almost as many advert prices as there are adverts. Not only are advert prices different for a given product but they are also spread over a large support. The last column of Table 2 shows that on average the highest advert price is more than twice as large as the lowest one. This ratio increases substantially if we focus on products for which there are many adverts (for example, the ratio is above 3 if there are more than 6 adverts).

Table 2: Number of adverts, advert prices and advert price dispersion per product (sold at least twice)

# adverts per product	Frequency	Average number of advert prices	Mean $p_{\max}/p_{\min}$
2	7,015	1.95	1.49
3	4,760	2.89	1.92
4	3,267	3.80	2.30
5	2,384	4.69	2.71
6	1,724	5.57	3.03
7	1,270	6.47	3.35
8	947	7.29	3.56
9	784	8.09	3.77
10	2,934	11.4	5.02
20	733	21.0	7.22
Any	25,818	5.07	2.69

Price dispersion is also substantial if we consider transactions. Table 3 shows that a given product, in the quarter of our observation sample, can be sold on average at more than 3 different prices (3.24 in the last row). Products with more than 8 transactions were on average sold at least at 5 different prices. Looking at the last column, we see that on average the highest transaction price is 76% higher than the lowest one for the same product. This relative difference rises above 100% for products sold more than 5 times.

Table 3: Number of transactions, transaction prices and transaction price dispersion per product

# transactions per product	Frequency	Average number of transaction prices	Mean $p_{\max}/p_{\min}$
2	5,401	1.79	1.38
3	2,971	2.47	1.62
4	1,863	3.12	1.82
5	1,118	3.66	2.00
6	790	4.27	2.16
7	554	4.73	2.26
8	415	5.28	2.37
9	300	5.83	2.21
$[10, 19]$	891	7.60	2.67
≥ 20	244	14.40	2.86
Any	14,547	3.24	1.76

Table 4 shows how often the cheapest advert on the screen is chosen by consumers. We see on the last row that for 48.5% of all transactions, the advert that was sold was not the cheapest. Also, when consumers do not buy the cheapest advert, they choose an advert which is on average 56% more expensive than the cheapest advert. The other rows show that, as the number of adverts increases, the cheapest advert is less likely to be the chosen one and, when it is not chosen, is relatively cheaper than the advert actually bought by the consumer.

Table 4: Cheapest and sold advert prices per transaction

			If $p_{\text{sold}} > p_{\min}$	
# adverts	Freq.	% $\{p_{\text{sold}} = p_{\min}\}$	mean $\frac{p_{\text{sold}}}{p_{\min}}$	rank p_{sold}
2	19,248	67.96	1.35	2.00
3	13,234	55.26	1.41	2.29
4	9,597	47.75	1.47	2.54
5	6,975	42.09	1.52	2.74
6	5,415	39.48	1.53	2.92
7	4,106	37.46	1.56	3.06
8	3,244	36.65	1.61	3.29
9	2,645	35.54	1.62	3.37
$[10, 19]$	10,545	31.76	1.80	3.84
≥ 20	2,744	23.87	2.05	5.02
Any	77,753	48.51	1.56	2.94

We have shown that there is a substantial dispersion in prices (for the exact same product) and that consumers do not necessarily choose the cheapest advert. These stylized facts may have three different explanations: seller/advert characteristics, consumer preferences for advert characteristics and/or search costs. In our model, these sources are captured by x, γ and s respectively and each of these objects can be heterogenous. We continue this descriptive analysis by looking at advert/seller characteristics: reputation, size, seller status and product condition.

The distribution of product quality across adverts and across transactions is shown in Table 5. From now on, the condition "fair", which is attached to only a few adverts, will be merged with the next best one, "good". This table shows that trade on the PriceMinister website mostly involves second-hand items, and

that shares of trade roughly reflect shares of adverts in terms of the condition of the item sold. We also note that although the second-hand product condition is self-reported, sellers are not tempted to systematically post the best condition. In fact, for 40% of the adverts, the product is not advertised as "new" or "as new".

Table 5: Distribution of product condition among adverts or transactions

	Good	Very Good	As New	New
Adverts	10.20	29.41	42.27	18.12
Transactions	7.65	21.81	38.76	31.78

We now look at the distribution of seller status among transactions and adverts. As Table 6 shows, the ratio between individual and professional sellers is roughly 2:1 among transactions or adverts.

Table 6: Number of seller status among adverts or transactions

	Adverts		Transactions	
	Frequency	Percentage	Frequency	Percentage
Professional	44,309	30.39	28,436	36.57
Individual	101,514	69.61	49,317	63.43
Any	145,823	100.00	77,753	100.00

Table 7 shows the distribution of seller size in the populations of adverts and transactions. As expected, professional sellers complete far more transactions than individual sellers but we still observe some individual sellers with hundreds, even thousands, of transactions. Looking at the low quantiles, we also note that some buyers trade with very small sellers.

Table 7: Distribution of seller size among adverts or transactions

	$Q_{5\%}$	$Q_{25\%}$	$Q_{50\%}$	$Q_{75\%}$	$Q_{95\%}$	Average
Adverts						
Professional	70	777	3,192	12,066	102,738	23,045
Individual	1	27	117	435	2,107	459
Transactions						
Professional	94	1,018	4,435	92,615	120,551	35,610
Individual	1	29	127	481	2,124	482

In Table 8 we show the distribution of seller reputation among adverts and transactions. Recall that each transaction is completed once feedback, an integer between 1 and 5, is given by the buyer. The seller's rounded average feedback is then shown on all his live adverts. We note that in most cases, the seller's reputation is higher than 4.5. Only 2.72% (respectively 2.77%) of adverts (respectively transactions) are

posted by sellers with a reputation strictly below 4.3. There is substantial dispersion between 4.5 and 5.

Table 8: Distribution of seller reputation among adverts or transactions

	≤ 4	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5
adverts	1.4	0.8	0.6	1.1	2.3	10.4	7.1	12.3	28.2	27.9	7.9
transactions	1.4	0.7	0.6	1.1	2.4	15.8	6.2	11.6	26.8	26.6	6.7

We have just shown that, in addition to price dispersion, there is also dispersion in advert/seller characteristics. If consumers care for these characteristics, this source of differentiation could explain some of the price dispersion, in the context of a perfect information model. We now need to show some evidence that search frictions may also be at play on this Internet platform, so that heterogeneity in advert characteristics and in consumer preferences cannot fully explain the dispersion in prices.

We showed above that in around 49% of transactions, the cheapest advert was not the one chosen by the consumer (see Table 4). The consumer's choice may thus be driven by other advert characteristics and/or hindered by search frictions. To motivate a more structural analysis of this issue, we compute the following statistic on the population of transactions. For a given transaction, we still denote as i the advert that was bought and j any other advert available at the time of purchase. Let us now define what we will refer to as a transaction with an 'unambiguously better' advert. We say that an advert j is 'unambiguously better' than i if j is at least 'as good as' i in terms of price, seller reputation and product condition and strictly better in at least one of these dimensions and if the sellers of adverts i and j have the same status (professional or individual) and are in the same size category[19]. This definition does not depend on any assumptions on consumers' preferences for seller size or status and only assumes that consumers' utility weakly increases with lower prices or better reputation or product condition.[20]

Counting these transactions with 'unambiguously better' alternatives will give a very conservative lower bound on the number of transactions that cannot be explained by heterogenous preferences for observed advert characteristics only. The proportion of transactions with 'unambiguously better' adverts is shown in Table 9. We see that this proportion is higher than 7% on average and that it increases with the number of adverts on the screen. This motivates the introduction of consumer search costs.[21]

[19] Where we define four categories: $[0, 50[$, $[50, 500[$, $[500, 5000[$ and ≥ 5000.

[20] Note that if advert characteristics can be summarized by a scalar indicator x valued positively by consumers then this definition of transactions with 'unambiguously better' adverts coincides with the definition of 'better' adverts presented in Section 2.4.

[21] Another explanation for the statistics in Table 9 would be the presence of unobserved advert characteristics.

Table 9: Transactions where an 'unambiguously better' advert was available

# adverts	Freq.	% 'unambigously better'
2	19,248	4.28
3	13,234	5.83
4	9,597	6.88
5	6,975	8.00
6	5,415	8.57
7	4,106	8.13
8	3,244	9.86
9	2,645	8.96
$[10, 19]$	10,545	10.21
≥ 20	2,744	14.07
Any	77,753	7.24

4 Empirical strategy

In this section, we describe the three stages of our empirical approach. In the first one, we project the vector of advert characteristics onto a scalar variable x, referred to as the hedonic index. In the second, we estimate the joint distribution of (P, X) in our dataset, from which we compute consumers' beliefs and, in particular, the conditional distribution $F(X|P)$ and the function ψ. In the third, we estimate the joint distribution of the consumer preference parameter γ and search cost s using a grid search approach and the outputs from the first two stages.

4.1 Aggregation of advert characteristics

As seen is Section 2.4, handling the consumer problem with only two dimensions of choice x and p greatly improves the tractability of the search problem. With a scalar x, the characterization (9) of the identified sets consists of a simple comparison of instantaneous and reservation utilities, the latter being easily obtained, through expression (8), using the structural function ψ. Also, in practice, since we only have set identification of the parameters of interest, it will be difficult to find the identified sets if we proceed with more than 2 dimensions. On balance, we think that the gains in clarity and tractability of working with a scalar hedonic index outweigh its only drawback, which is the lack of heterogeneity in the marginal rate of substitution between two non-price characteristics (recall that we allow for heterogeneity in the MWP for the scalar hedonic index).

Consider an advert characterized by its price p and a vector of $K \geq 1$ characteristics $\left\{x^{(k)}\right\}_{k=1,K}$, where $x^{(k)}$ relates to the seller's reputation, size, professional status and the state of the item for sale (new/as new/used ..). We define our aggregate hedonic index as a linear projection of these characteristics:[22]

$$x = \sum_{k \geq 1} \beta^{(k)} x^{(k)}, \tag{10}$$

This is in effect reducing the amount of preference heterogeneity allowed across consumers since the vector

[22] We also impose $\beta^{(1)} = 1$ for normalization (the related characteristic will be one that is unambigously valued positively by consumers, say reputation).

parameter $\beta = \left(\beta^{(1)}, ..., \beta^{(K)}\right)$ is homogenous among consumers. Preference heterogeneity is now reduced to one dimension, embodied by the scalar parameter γ. We thus assume that consumers share the same marginal rates of substitution between two advert non-price characteristics but we let the marginal willingness to pay γ for the hedonic variable x (and thus for any non-price characteristic) be heterogeneous across consumers.

Estimation. As we set $\beta^{(1)} = 1$ we need to choose $x^{(1)}$ to be a characteristic that cannot negatively affect utility. We choose seller reputation. This ensures that $\gamma \geq 0$. We also consider the following advert characteristics: five seller size dummies (≤ 50, $[51, 100]$, $[101, 500]$, $[501, 5000]$, > 5000), four product condition dummies ('good, 'very good, 'as new', 'new') and a professional seller dummy. For reputation we use a variable that is equal to 10 times the seller's reputation so that γ can be interpreted as the marginal willingness to pay for a 0.1 increase in reputation. This is because most of the variation in reputation is between 4 and 5 and the relevant changes are those measured in decimals.[23]

For the estimation of the hedonic index, we only consider the transactions where there were only $J = 2$ adverts available at the time of purchase. We assume that for these transactions, the search cost is equal to 0. Our reasoning is the following: the consumer observes immediately (at no cost) that there are only two adverts on the page, and we assume that, given this reduced potential search horizon, the consumer examines them both. In other words, when there are only two adverts on the screen, the consumer systematically behaves according to the perfect-information model (see Section 2).[24] As Table 1 in Section 3.3 showed, 19,248 transactions involved only two adverts. In this case, when advert i is chosen over advert j, we must have:

$$\gamma \sum_{k \geq 1} \beta^{(k)} x_i^{(k)} - p_i \geq \gamma \sum_{k \geq 1} \beta^{(k)} x_j^{(k)} - p_j. \tag{11}$$

Now, since γ is allowed to be heterogeneous across consumers, we cannot use variations across transactions to estimate the $\beta^{(k)}$ directly from this inequality. Also, if $p_i \leq p_j$, the transaction can be explained with $\gamma = 0$ so we will not get any information on β. We thus use transactions with two adverts and for which the advert sold is strictly more expensive than the alternative. There are 5,984 such transactions. In these cases $p_i > p_j$ and, given $\gamma > 0$ and (11), the following inequality should hold:

$$\sum_{k \geq 1} \beta^{(k)} x_i^{(k)} > \sum_{k \geq 1} \beta^{(k)} x_j^{(k)}. \tag{12}$$

Our estimate of β will be such that the number of violations of (12) is minimal. It should thus belong to the following set:

$$B = \underset{\beta}{\text{argmax}}\ C_1(\beta), \quad \text{where} \quad C_1(\beta) = \sum_{J=2, p_i > p_j} \mathbf{1}\left\{ \sum_{k \geq 1} \beta^{(k)} x_i^{(k)} > \sum_{k \geq 1} \beta^{(k)} x_j^{(k)} \right\} \tag{13}$$

The set B is not a singleton in general so maximizing the criterion $C_1(\beta)$ does not achieve point identification of β. Also, in our data, for any value of β, there are still transactions (with $J = 2$ and $p_i > p_j$)

[23] The very few sellers with reputation levels below 4 (40 with our re-scaling) or with no reputation yet (no completed transactions), have their reputation set at 40.

[24] Importantly, if $J \geq 3$ this does not mean that consumers have two 'free' draws. In the directed search model with several adverts, any draw is costly. We just assume here that if $J = 2$, the cost is 0.

for which (12) is violated. This means that, for these transactions, the difference $\beta(x_j - x_i)$ if positive. We will use these transactions to select a value of β within the set B and minimize the mean squared "error":

$$C_2(\beta) = \sum_{J=2, p_i > p_j} \mathbf{1}\{\beta(x_j - x_i) \geq 0\} [\beta(x_j - x_i)]^2 . \quad (14)$$

The resulting value of β will then be such that it maximizes the number of transactions verifying condition (12) and, for transactions which do not verify this condition, minimizes the average squared error. A direct way to find this value of β is to maximize $C_1(\beta) \cdot \exp(-C_2(\beta))$.

Whilst maximizing $C_1(\beta)$ is fully consistent with the structure of our model, minimizing $C_2(\beta)$ does not come from our model and is thus arbitrary. It can be seen as a way to calibrate the parameter β within the identified set B. We will check the robustness of our results to other selection rules within the set B. For instance, each transaction in C_2 will be weighted by the relative price difference. In a robustness section, 5.3, we will see that, in our data, using C_2 or another criterion has little effect on the results.

The estimated scalar hedonic index. The estimated coefficients on advert characteristics are reported in Table 10. We note that the item condition dummies are ranked intuitively. With this value of β, we can explain 4,387 of the 5,984 transactions (73.3%) with only two adverts and for which the most expensive advert was sold.[25]

Table 10: Estimation of the β parameters

Reputation ($\beta^{(1)}$, normalized)	1.000
Size $\in [50, 100]$	0.724
Size $\in [101, 500]$	1.002
Size $\in [501, 5000]$	1.724
Size > 5000	3.716
as new	-10.994
very good	-12.281
good	-18.994
Professional	0.008

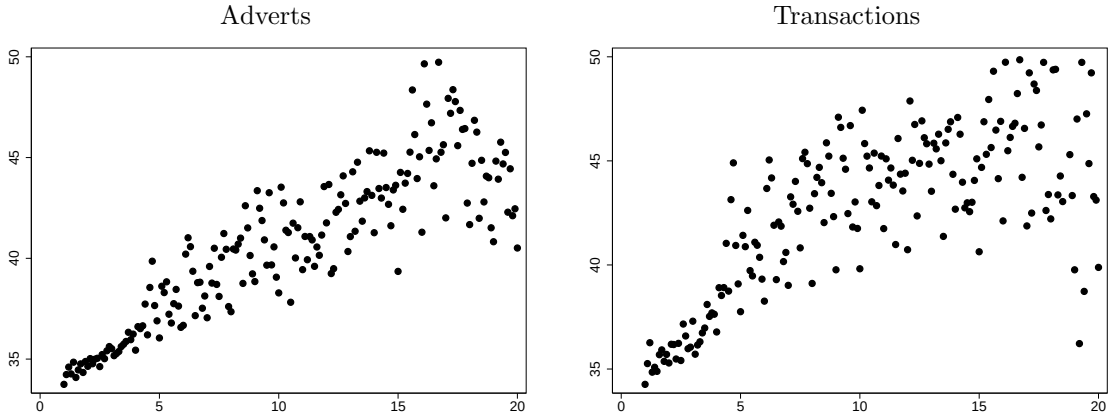
The resulting aggregate hedonic index x ranges from 21.0 to 52.7, and its distribution in the population of adverts has an average of 38.1, a standard error of 6.3 and a median at 37.7. All the benchmark estimation results presented and discussed below rest on this parameterisation of the single index x.

A casual observation worth reporting at this point is that this hedonic index x tends to increase on average with price as can be seen in Figure 1, both for the population of adverts and transactions. The average slope of these curves suggests that the mean index, $E(X|p)$, increases by 0.5 by unit of price (euro). As the coefficient on reputation ($\times 10$) in x equals 1, this means that each increase in price by €1 can be expected to reflect, on average, an increase in 0.05 of the reputation indicator (recall that reputation is essentially between 4.3 and 5) or any change in other advert characteristic bringing the same change in hedonic value x to the consumer. This gives an intuitive justification as to why the item sold is not always the cheapest, a fact reported above. More expensive items tend to have a higher x and since some of the buyers care about

[25] The fit for all transactions with 2 adverts is then much higher as if $p_i \leq p_j$, the model is accepted with $\gamma = 0$.

x they are prepared to pay the increase in price to enjoy the corresponding expected increase in the hedonic index.

Figure 1: Mean x by price (rounded to the 1st decimal)



Note: the mean x is on the vertical axis, price is on the horizontal axis.

4.2 Estimation of consumer beliefs

As shown in Section 2.4, consumers' search and purchase stragegy depends on a function $\psi_p(x)$ which represents the expected hedonic gain of drawing an advert at price p. This function follows from consumers' beliefs about the joint distribution of p and x. In this section, we show how we estimate this distribution and then discuss the two specifications we will use to estimate the ψ function.

Non-parametric estimation of the joint density of (p, x). We consider that we are in an equilibrium, where consumers' beliefs about the joint distribution of p and x coincide with the actual distribution of these two variables in the population of adverts (which follows from the sellers' profit maximization program). This distribution is observed in the data and can thus be estimated non-parametrically. In what follows, we assume that a consumer's beliefs about x depend on prices only up to the first decimal. This is done for practical reasons and we think it is not unrealistic to assume that buyers may expect the same characteristics between two adverts with prices at 10.64€and 10.67€. In the following estimation of beliefs, the data is grouped by price rounded to the nearest first decimal.

The raw joint distribution calculated as an empirical cumulative distribution function from our advert sample yields a somewhat jagged result mostly because of the small size of some (p, x) cells in the data. We thus estimate the joint $F(P, X)$ with a product Gaussian kernel. The probability density function of (p, x) in the population of adverts is computed as:

$$f(p,x) = \frac{1}{N_a h_p h_x} \sum_{k=1}^{N_a} \phi\left(\frac{p - p_k}{h_p}\right) \cdot \phi\left(\frac{x - x_k}{h_x}\right)$$

where $\phi(.)$ is the standard normal density, N_a is our advert sample size and $h_p = 0.363$ and $h_x = 0.632$ are the chosen bandwiths in the p and x dimensions respectively.[26]

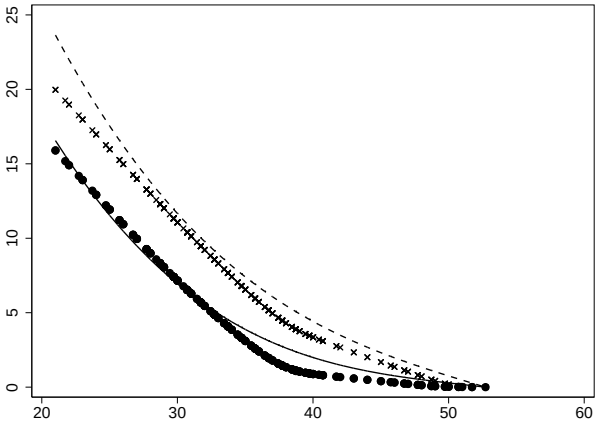
[26] From Silverman (1998)'s rule of thumb: $h = 1.06\sigma N_a^{-1/5}$, where σ is the sample standard deviation.

Two specifications of the ψ function. Once we have this non-parametric estimate of the joint density $f(p, x)$, we can compute the ψ function. Plugging our kernel estimate of the conditional $F(x|p)$ into equation (7), we get our first estimate of ψ. We will call this the "kernel" specification as it imposes no parametric assumption on the beliefs (beyond the assumptions used to produce the kernel density).

As shown in Figure 2, this first specification yields a $\psi_p(x)$ function that, given p, is decreasing and relatively smooth with respect to x but, given x, is not always monotone with respect to p. In words, the expected hedonic gain does not systematically increase with price, even though the trend is clearly increasing. This is because for a few price values (around 10, 15 or after 17€), the average x decrases in the population of adverts in our sample.

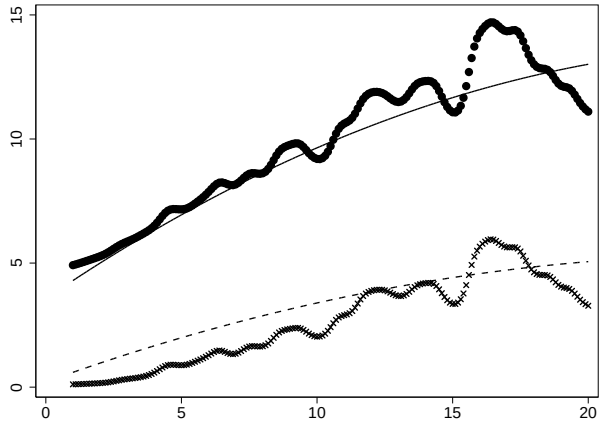
Figure 2: $\psi_p(x)$ functions using the kernel or parametric specification

Figure 2a: $\psi_p(x)$ vs. x, with p fixed



$\psi_p(x)$ on y-axis, x on x-axis
$p = 5$: kernel (dot) - parametric (solid)
$p = 15$: kernel (cross) - parametric (dashed)

Figure 2b: $\psi_p(x)$ vs. p, with x fixed



$\psi_p(x)$ on y-axis, p on x-axis
$x = 30$: kernel (dot) - parametric (solid)
$p = 40$: kernel (cross) - parametric (dashed)

A reasonable alternative would be to limit the sophistication in consumers' beliefs and impose more smoothness and force ψ to increase with respect to price. This will be our second, and benchmark, specification. We thus construct what we will be calling "parametric" beliefs by fitting a polynomial of order 3 in x and p to the values of $\psi_p(x)$ obtained with the kernel specification. In order to ensure that fitted values are consistent with ψ being the mathematical object defined in (7), we constrain these to be positive and decreasing in x. Besides, we constrain the smooth beliefs to be increasing in p as we find it intuitively appealing to constrain consumers to believe that expected gains in "quality" increase with price, at all quality levels. This is already the case for the overwhelming majority of prices. This is equivalent to assuming stochastic dominance of $F(p', X)$ over $F(p, X)$ for all $p' > p$. As is shown in Figure 2 parametric beliefs are close to the values obtained with the kernel estimation.

There is a trade-off between estimating beliefs close to the joint distribution observed in the data and using beliefs that are consistent with smooth variations in expectations over price but are further from the data. We will use the parametric specification as our benchmark but for completeness we will also present

estimation results using the kernel specification.

4.3 Grid search of preference and search cost parameters

The last stage of our estimation procedure is relatively straightfoward. We use the scalar hedonic index estimated in the first stage (in Section 4.1), and the ψ function estimated in Section 4.2 to compute the utility u and reservation utility r for each advert. We then browse a two-dimensional grid,[27] checking at each point in the grid whether the parameter values (s, γ) are consistent with each transaction i using condition (4), (5) or (9). The first condition is used when $s = 0$, the second when $\gamma = 0$ and the last one if $s \cdot \gamma > 0$.

Note that this last step requires the outcomes, x and ψ, from the first two stages but does not depend on the methodology adopted in these first two steps. We can thus conduct robustness checks where we change the scalar hedonic index and/or the specification of the ψ function and still use the grid search procedure described in this section.

5 Results

5.1 Search strategies

Before presenting our results on the preference and search cost parameters in the next section, we look at consumer search patterns. Using the ψ function estimated in Section 4.2 with the parametric specification and equation (8), we can compute reservation utilities r for any price p and any value of s and γ. We can thus study how the sampling order depends on consumers' preferences (γ) and search cost (s).

In Figure 3 we plot price against sampling order for different values of s and γ. As detailed in the above theory section, the optimal search strategy is to sample items in descending order of reservation utilities. Note that the following results do not require information on transactions as they pertain to sampling behaviour rather than to purchases. Our theoretical framework delivers rich predictions in terms of sampling order, which we now give several illustrations of. For each (s, γ), we compute $r(p, s, \gamma)$ for all prices on a grid from 1 to €20 (with a step of 0.1). The price with the highest (resp. 2$^{\text{nd}}$ highest, resp. lowest) reservation value r is then sampled 1$^{\text{st}}$ (resp. 2$^{\text{nd}}$, resp. last). Figure 3 displays the sampling rank on the horizontal axis versus the price in the vertical axis. A monotonic sampling order will then be represented by an increasing (respectively decreasing) line when the consumer samples items in increasing (resp. decreasing) order of price. Non-monotonic sampling behaviour arises in some cases whereby the consumer starts sampling a mid-range price level and subsequently samples items higher and lower than the initial price in an alternating pattern predicted by the series of reservation utilites for the consumer's individual search cost and marginal willingness to pay for the hedonic index. In this case, the figure shows sampled prices getting further and further away from the initial price level as the sampling order goes up.

[27] Values on this grid are the following. For s: 0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5(0.5)6 and 7(1)10. For γ: 0(0.1)5, 6(0.5)10, 20. The notation 0(0.1)5 means any value between 0 and 5, with a step of 0.1.

Figure 3: Price vs. search order, the effect of preferences and search cost

Figure 3a: $\gamma = 0.1$ and $s = 0.1$ (dash) or $s = 1$ (solid)

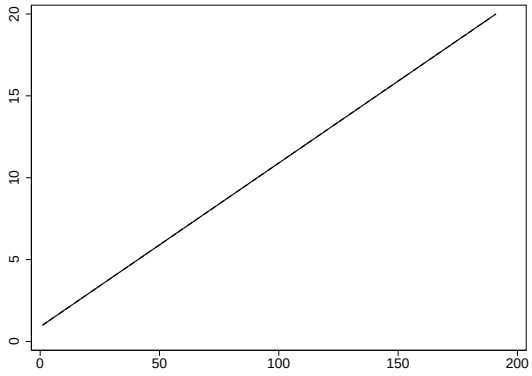


Figure 3b: $s = 2$ and $\gamma = 0.1$ (solid) or $\gamma = 1$ (dash)

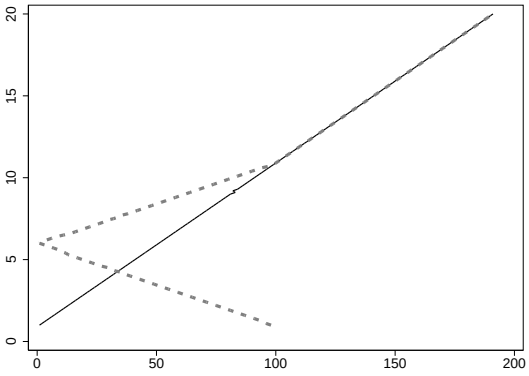


Figure 3c: $s = 2$ and $\gamma = 1$ (solid) or $\gamma = 4$ (dash)

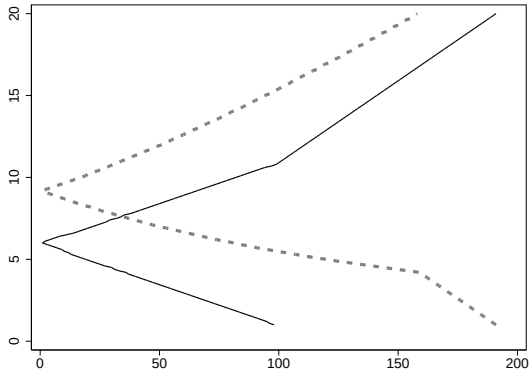
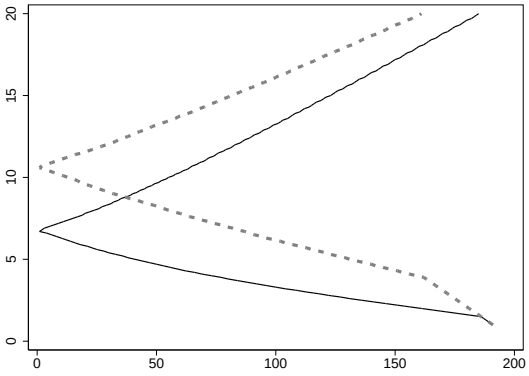


Figure 3d: $\gamma = 2$ and $s = 1$ (solid) or $s = 4$ (dash)



Note: price on y-axis, search order - rank of $r(p)$ - on x-axis.

In Figure 3a, we show the sampling order when the MWP to pay for x is very low (€0.1) and the search cost equals 0.1 or €1. This shows that when consumers barely care about the hedonic index x they sample adverts by increasing order of price. This does not depend on search costs (the two lines in Figure 3a are superimposed) because the advert characteristic that is important to consumers, price, is observable at no cost. This is important for our results: the observed consumer behaviour needs to be explained by the *joint* presence of search costs and a consumer's taste for non-price characteristics.

Figure 3b shows that, in the presence of search costs, the search pattern markedly changes when preferences for x are stronger. The sampling order is no longer increasing in price i.e. consumers do not sample the cheapest advert first. If $s = 2$ and $\gamma = 1$, we see that they would first look at adverts with a price around €6 then at prices of €5 or €7. The most expensive adverts, above €11, are still sampled last.

Increasing further the MWP for x yields another search pattern. In Figure 3c, we see that, for a given search cost, say $s = 2$, as the taste for the hedonic index γ increases from 1 to 4, the first price sampled by the consumer rises from €6 to €9. Another difference is that the most expensive adverts are no longer sampled last. This is intuitive as the more consumers care about hedonic advert characteristics, the more likely they

are to look at expensive adverts before cheap adverts. The last case we consider, in Figure 3d, shows that, keeping preferences fixed, the search order also depends on the level of search cost. In this example, the graph shifts upwards when s goes from 1 to 4, meaning that the alternative pattern of sampling above and below the initial price remains, but starting at a higher point.

To summarize, this section illustrates the flexibility of our search model with respect to sampling patterns. Depending on parameter values, consumers may not sample adverts by increasing order of price. The sampling order and thus the consumer's choice set will depend on his preferences and search cost. As far as we know, ours is the first empirical analysis of this type of consumer behaviour in the context of a directed search model.

5.2 Preferences and search cost

We now present the results from the last stage of our estimation procedure and show the estimated sets of search cost and preference parameters. All these results make use of the parametric specification for the estimation of the beliefs as outlined in section 4.2. Robustness checks using alternative specifications will be shown in Section 5.3.

Model fit. For each transaction, we will consider that our model fits the observed choice if S_i is not empty i.e. there is at least one value of (s, γ) such that we cannot reject the model with conditions (4), (5) or (9) for all the (i, j) comparisons relevant to this transaction. Table 11 shows the fit of our model, with a breakdown by the number of adverts per transaction. Note that we do not include transactions with only 2 adverts as these transactions were used to produce the scalar hedonic index (in Section 4.1) and were assumed to trigger a slightly different consumer behaviour (without search costs). We also break transactions down into two categories, depending on whether an alternative 'better' advert was available but not bought. 'Better' adverts were defined in Section 2.4: advert j (not sold) is 'better' than advert i (sold) if it is both cheaper and offering a higher hedonic index, i.e. $p_j \leq p_i$, $x_j \geq x_i$ with at least one of these two inequalities being slack. As discussed in Section 2.4, transactions with 'better' adverts play an important role in the identification of positive search costs. The last row of Table 11 shows that 14,640 out of 58,505 (24.7%) of all transactions had at least one 'better' advert.

The main result from Table 11 is that our model explains almost all transactions (94%), whether the number of adverts was small (96% if 3 adverts) or large (88% if strictly more than 15 adverts). The fit is even higher (99.5%) among transactions with no 'better' advert. The fit is still high (76.3%) if we look at transactions with at least one 'better' advert and remains relatively stable when the number of adverts per transaction increases. Our search model thus does a very good job at explaining the transactions on this Internet platform, even when considering transactions with 'better' adverts, that is transactions that would be poorly explained by a perfect information model.

Table 11: Pass rate (%) among transactions with J adverts

# adverts	All transactions		Transactions with no 'better' advert		Transactions with ≥ 1 'better' advert	
	Freq.	% Pass	Freq.	% Pass	Freq.	% Pass
3	13,234	96.34	11,117	99.71	2,117	78.65
4	9,597	95.57	7,561	99.74	2,036	80.11
5	6,975	94.71	5,306	99.74	1,669	78.73
6	5,415	93.41	3,954	99.67	1,461	76.45
7	4,106	93.64	3,052	99.38	1,054	77.04
8	3,244	92.05	2,299	99.35	945	74.29
9	2,645	93.42	1,853	99.19	792	79.92
10	2,174	92.09	1,527	99.21	647	75.27
11	1,707	91.62	1,191	99.08	516	74.42
12	1,472	90.56	992	98.79	480	73.54
13	1,192	89.85	792	99.12	400	71.50
14	1,063	89.93	718	98.75	345	71.59
15	827	90.45	554	99.10	273	72.89
> 15	4,854	87.62	2,949	98.78	1,905	70.34
Any	58,505	93.69	43,865	99.50	14,640	76.30

Freq: number of transactions in each category.

Pass: proportion (in %) of transactions that can be explained - $S_i \neq \{\emptyset\}$.

Strictly positive search costs. We now assess whether the fit of the model is due to its search component or whether a perfect-information model would fit the data just as well. In other words, do we need strictly positive search costs to explain thes transactions? For each transaction i that is explained by the model, we can define $\underline{s}_i$ as the lowest value of s that allows our model to be consistent with the observed transaction i. We can then look at the average of $\mathbf{1}\{\underline{s}_i > 0\}$ –an indicator that a strictly positive search cost is needed to account for transaction i– among all adverts explained by the model. Results are in Table 12.

We see that strictly positive search costs are needed for 26% of all transactions explained by our model. This proportion increases steeply with the number of adverts per transactions (14% if 3 adverts, 44% if strictly more than 15 adverts). This is intuitive as search costs may not be a major hurdle when there are a few adverts on the screen but they are far more likely to affect the consumer's decision when the number of adverts is large.

For transactions with no 'better' adverts, positive search costs are rarely needed (9%) unless the number of adverts is large. This is not surprising as most transactions with no 'better' advert can be explained by a perfect-information model and the appropriate taste for x. This is not systematic though, as we discussed in Section 2.4, and our results suggest that there is a non-negligible proportion of transactions for which the MWP's implied by comparing the sold advert with two different alternatives, i.e. the ranges of γ compatible with the observed choice, are not consistent, so that the transaction can only be explained with a positive search cost, since we need the theoretical framework to predict that the consumer did not sample all adverts.

As we expect from the discussion in Section 2.4, the overwhelming majority of explained transactions with a 'better' advert require a strictly positive search cost (92%). Not all of them do because some of the 'better' transactions are such that $p_i = p_j$ and $x_i < x_j$: these can be explained without search costs by

relying on a zero taste for the hedonic index ($\gamma = 0$) and our rule for ties. This comparison can be explained with $\gamma = s = 0$. In any other case, a transaction with a 'better' advert cannot be explained unless $s > 0$.

Table 12: Proportion (%) of explained transactions requiring strictly positive search costs

	All transactions		Transactions with no 'better' advert		Transactions with ≥ 1 'better' advert	
# adverts	Freq.	% $\underline{s} > 0$	Freq.	% $\underline{s} > 0$	Freq.	% $\underline{s} > 0$
3	12,750	14.03	11,085	2.72	1,665	89.31
4	9,172	20.72	7,541	5.58	1,631	90.68
5	6,606	25.42	5,292	9.18	1,314	90.79
6	5,058	28.25	3,941	9.97	1,117	92.75
7	3,845	29.49	3,033	12.07	812	94.58
8	2,986	32.48	2,284	13.57	702	94.02
9	2,471	34.36	1,838	14.53	633	91.94
10	2,002	33.82	1,515	14.98	487	92.40
11	1,564	36.70	1,180	18.56	384	92.45
12	1,333	37.58	980	18.27	353	91.22
13	1,071	36.79	785	16.31	286	93.01
14	956	37.03	709	17.35	247	93.52
15	748	38.24	549	17.30	199	95.98
> 15	4,253	43.90	2,913	20.67	1,340	94.40
Any	54,815	26.28	43,645	9.44	11,170	92.08

Freq: number of transactions in each category.

$\underline{s} > 0$*: proportion (in %) of explained transactions that need a strictly positive search cost.*

We view the results from Table 12 as particularly important as they provide empirical evidence of strictly positive search costs on an Internet platform and yet are based on a very flexible specification of consumers' preferences (we allow for full heterogeneity in γ and impose no parametric assumption in its distribution) and of their search behaviour, where the sampling order is shaped by preferences and search costs.

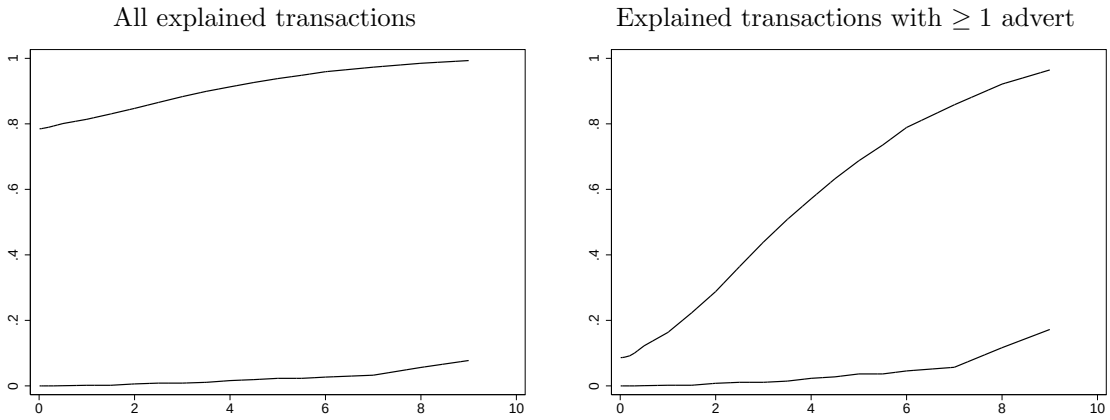
Bounds on the distribution of search cost and preference parameters. We now give empirical evidence on the distribution of s and γ in the population of transactions that can be explained by our model (94% of our sample). We will denote the set of transactions i that can be fitted by our model as I_f and the number of these transactions as N_f. We have already defined $\underline{s_i}$ as the lowest search cost that makes the model consistent with a given transaction i. We can also define $\overline{s_i}$ as the highest search cost that fits this transaction. Similar bounds can be defined for the preference parameter γ and be denoted as $\underline{\gamma_i}$ and $\overline{\gamma_i}$. The lower bounds exist as S_i is not empty (we only consider explained transactions) and both $s, \gamma \geq 0$. The upper bounds may be infinite in theory and, in this section, are limited by the highest value on our grid (10 for s and 20 for γ). These chosen highest values are arbitrary but are intuitively consistent with our idea of an upper bound for these two quantities: a search cost of €10 euros to sample the next item and a willingness to pay of €20 for an increase in seller's reputation of 0.1 both seem very large, even for the most extreme individual buyer.

For each value of s, we define the lower and upper bounds of the cdf of search costs as follows:

$$\underline{H}(s) = \frac{1}{N_f} \cdot \sum_{i \in I_f} \mathbf{1}\left\{\overline{s}_i < s\right\} \quad \text{and} \quad \overline{H}(s) = \frac{1}{N_f} \cdot \sum_{i \in I_f} \mathbf{1}\left\{\underline{s}_i < s\right\}. \tag{15}$$

We can similarly define bounds for the cdf of γ. Also, if we take the averages in (15) among transactions with at least one 'better' advert, we get bounds on the cdf of s and γ for these specific transactions. It is important to note that these bounds are not sharp: all cdf's consistent with the data must be within these bounds but the converse is not true. To get sharp bounds for the cdf of (s, γ), we would need convex identified sets S_i. Even if we impose more structure on consumers' beliefs, this may not be the case.[28] We thus choose to keep a flexible specification of the model and produce non-necessarily sharp bounds, which are still informative.

Figure 4: Bounds on the cdf of minimum search cost $\underline{s}$

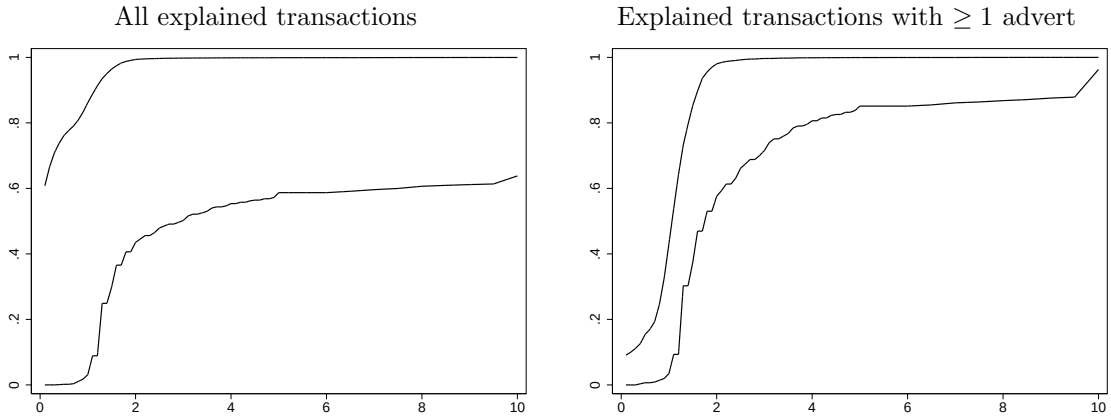


The bounds on the cdf of search costs are shown in Figure 4, in the population of all explained transactions and of explained transactions with 'better' adverts. We first note that in both cases, the lower bound is close to (but different from) 0. This is not systematic though as we can see that the lower bound is not flat. The second remark is that if we look at all explained transactions, the bounds we find are wide. This was expected as we know from Table 12 that 73.7% of transactions can be explained with search costs equal to 0. However, if we focus on transactions with 'better' adverts, we get tighter bounds and search costs can be relatively high, with a median at €4 for $\underline{s}$. Hence, the average search cost over the whole population can be very low but, for those who have a positive search cost, this cost can be quite high.

Turning now to the bounds on the cdf of γ displayed in Figure 5, we see that these bounds are relatively tight. The median of the MWP for x is between 0 and €3 and, if we focus on transactions with 'better' adverts, between 1 and €2. Hence, in addition to finding positive and relatively large search costs, we also show that consumers reveal a substantial willingness to pay for non-price characteristics.

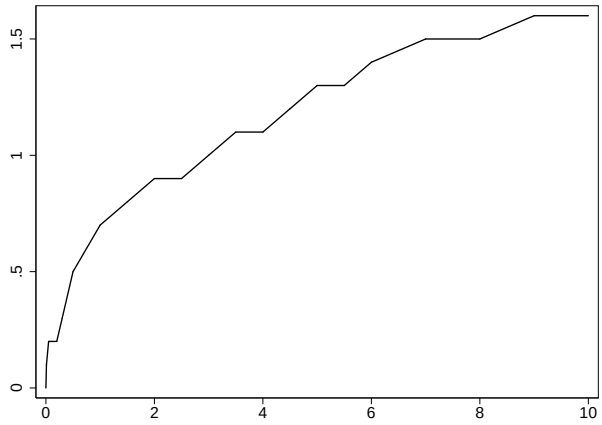
[28] We have explored how additional structure on the beliefs can shape the identified sets S_i. More details on this issue are available upon request.

Figure 5: Bounds on the cdf of minimum preference parameter $\underline{\gamma}$



We can also have a first look at the joint distribution of s and γ. We have defined $\underline{s}_i$ as the lowest search cost that makes our model fit transaction i. For this transaction i and this value of s, we can now define $\underline{\gamma}_i(\underline{s})$ as the lowest MWP such that the model fits transaction i when $s = \underline{s}_i$. We now check whether $\underline{\gamma}(\underline{s})$ increases with $\underline{s}$ i.e. whether consumers facing higher search costs have stronger preferences for the hedonic component x. To this end, we show in Figure 6 how the median of $\underline{\gamma}(\underline{s})$ varies with $\underline{s}$. The pattern is clearly increasing. This suggests a positive correlation between the presence of seach costs and a positive willingness to pay for x among consumers.

Figure 6: First look at the joint distribution of preferences and search cost, $\underline{\gamma}(s)$ vs s.



$\underline{\gamma}(s)$ = lowest γ such that the transaction is explained with search cost s.

Sampling strategies. Now that we have identified the sets of parameters (s, γ), we can delve further into the sampling patterns studied in Section 5.1 and produce more statistics on the sampling order. Since (s, γ) are only set-identified, we need to select for each explained transaction i a value of (s, γ) in S_i. We choose to focus on the lowest search cost needed to explain the transaction, $s = \underline{s}_i$, and its corresponding lowest MWP, $\gamma = \underline{\gamma}(\underline{s})$. Of course this is only an illustrative example, but it will help to connect our results on the

distribution of (s, γ) to the considerations regarding sampling behaviour in the previous section. For these values, we can compute for each transaction the utilities and reservation utilities of each advert. We can then see how often the sold advert i has the highest r, in which case it will be sampled first (or sampled first with positive probability if other adverts have the same r). We also look at the highest sampling order of the advert sold i, which will be given by the highest observed rank of r_i –in other words the number of adverts sampled before the chosen one. Lastly, we can check whether the consumer keeps sampling adverts once he has drawn the advert that he will eventually buy. According to the model shown in Section 2, this will happen if there is an alternative advert j such that $u_j \leq u_i < r_j < r_i$. For these statistics to make sense, we focus on transactions that can be explained by the model and for which search costs are strictly positive, $\underline{s} > 0$ (if $s = 0$ can account for the observed choice then sampling order is not a meaningful concept any more).

Table 13: Sampling strategies, explained transactions with $s = \underline{s} > 0$

# Adverts	Freq.	% Samp. 1st	Max	% Keep samp.
3	1,789	82.39	2	0.34
4	1,900	74.16	3	2.90
5	1,679	69.03	4	4.82
6	1,429	66.06	5	6.79
7	1,134	62.26	6	8.29
8	970	60.41	7	8.97
9	849	58.54	8	8.72
10	677	56.28	9	9.60
11	574	51.22	10	11.32
12	501	51.70	11	8.78
13	394	53.81	9	11.68
14	354	54.24	10	12.99
15	286	52.45	11	10.14
> 15	1,867	46.33	28	9.27
Any	14 403	63.38	28	6.68

% Samp. 1st: % of transactions where the advert bought was sampled first.
Max: maximum sampling order of advert bought.
% keep samp.: % of transactions where the advert bought was not sampled last.

Results are in Table 13. We see that the sold advert has the highest reservation value (and is thus sampled first) 63% of the time. This figure decreases sharply, from 82% to less than 50%, with the number of adverts. We also note that adverts sold can also have a high sampling order, as the maximum rank of r_i is often close to the actual number of adverts on the screen. This means that in some cases, the sold advert was sampled after most of the other adverts. The last column of Table 13 shows that consumers may carry on sampling adverts after they have drawn the one that they will eventually buy. This is consistent with the search patterns found by De Los Santos et al. (2012) albeit in a different context (search for books across e-commerce websites). As discussed in Section 2 and now shown with this set of results, such a consumer behaviour is consistent with a sequential directed search model.

5.3 Robustness checks

Alternative specifications. In this paper, we have chosen to estimate sets of parameters and not to impose restrictions on the distribution of heterogeneity. Our results thus do not hinge on parametric distributional assumptions. However, to produce the scalar hedonic index β and the consumers' beliefs ψ in Sections 4.1 and 4.2, we had to make two choices. First, since β is only set-identified, we had to impose a criterion, (14), to select a value in the identified set. Secondly, we have used a parametric specification of the beliefs ψ rather than the one estimated by kernel. In this section, we show that our results are robust to changes with respect to these two specification choices.

We explore two alternatives for the estimation of β. In the first one, we still consider the set B obtained by maximizing the criterion C_1 (see (13)), but instead of minimizing the average error C_2 (see (14)) we select a value in this set that minimizes the weighted average error, where the weights are given by the relative price difference:

$$\sum_{J=2,p_i>p_j} \mathbf{1}\left\{\beta\left(x_j - x_i\right) \geq 0\right\}\left[\beta\left(x_j - x_i\right)\right]^2 \cdot w_i, \text{ where } w_i = \frac{p_i - p_j}{\sum\limits_{J=2,p_i>p_j} \mathbf{1}\left\{\beta\left(x_j - x_i\right) \geq 0\right\}\left(p_i - p_j\right)} \tag{16}$$

The motivation for these weights is that if p_i is much larger than p_j, then the consumer must have a large MWP for x to compensate for this price difference. Violations of (12) will be more problematic for these transactions, i.e. "further" away from a behaviour consistent with our model, so we assign them a larger weight when minimizing the average error. The resulting estimate of β is very close to the one we obtained with the benchmark specification.

The other specification we consider for β is based on a logit regression. This time we do not use the identified set of values that maximizes C_1. Instead, we follow a discrete choice approach. We consider transactions with $J = 2$ adverts and $p_i > p_j$ and we estimate the probability that one advert is sold as a logistic function of the difference between the non-price characteristics of the two adverts. This yields an estimate of β which is still close to the benchmark value found in Section 4.1 so, even if it does not maximize C_1, it still fits the transactions with two adverts and $p_i > p_j$ very well.

As the results in Table 14 show, changing the specification for β does not affect the results, whether we look at the fit, at the role of search costs or at the distribution of s and γ.

The second robustness check pertains to the specification of consumer beliefs, through the ψ function. In our benchmark results, we have used a parametric approximation of ψ (with a cubic polynomial in p and x) to ensure that beliefs were relatively smooth and that consumers expect a higher hedonic index x when prices increase (which was already the case for most price levels in the data). We now look at estimation results based on the ψ function estimated by kernel (see Section 4.2). Table 14 shows that the conclusions are qualitatively similar to those reached with the parametric specification, apart from the following two differences. First, the fit on transactions with 'better' adverts decreases substantially, from 76% to 46%. Secondly, the distribution of minimum search costs $\underline{s}$ shifts to the left (i.e. search costs are lower), which follows from the fact that we cannot explain as many transactions with 'better' adverts. The overall fit of the model remains quite high (84%).

This is intuitive as transactions with 'better' adverts are often such that $p_i > p_j$ (the advert sold is more expensive than the better alternative). For this transaction to be explained by the model, it must be that i is drawn and j is not. If i is more expensive, consumers must expect a better x in order to sample i first. This will not always be the case if $\psi_p(x)$ does not increase with p. Hence the jagged patterns shown in Figure 2b for the kernel specification negatively affect the fit of the model. This will not happen with the parametric specification so we can find a value of (s, γ) such that i is drawn before j. However we also need a high enough search cost so that j is not drawn. This is why the cdf of search costs shifts to the right when using the parametric specification.

Table 14: Results using alternative specifications of β or ψ

	Specification			
	β bench. ψ param.	β price ψ param.	β logit ψ param.	β bench. ψ kernel
pass rate (%) among transactions				
- any	93.69	93.66	94.06	83.77
- with no 'better' advert	99.50	99.47	99.61	96.42
- with ≥ 1 'better' advert	76.30	76.37	77.29	45.86
share (%) of explained transactions with $\underline{s} > 0$				
- any	26.28	26.48	26.47	17.54
- with no 'better' advert	9.44	9.53	9.54	6.54
- with ≥ 1 'better' advert	92.08	92.17	92.31	86.82
share (%) of explained transactions with ≥ 1 'better' advert and:				
- $\underline{s} = 0$	7.92	7.83	7.69	13.18
- $0 < \underline{s} \leq 0.5$	3.23	2.90	3.04	15.27
- $0.5 < \underline{s} \leq 1$	3.57	3.67	4.74	9.07
- $1 < \underline{s} \leq 2$	12.19	11.66	12.21	12.62
- $2 < \underline{s} \leq 3$	7.89	7.73	7.09	5.91
- $3 < \underline{s} \leq 4$	21.83	22.10	23.01	14.97
- $4 < \underline{s} \leq 5$	12.11	12.02	12.07	8.46
- $5 < \underline{s}$	31.25	32.08	30.16	20.52
share (%) of explained transactions with ≥ 1 'better' advert and:				
- $\underline{\gamma} = 0$	7.92	7.83	7.69	13.18
- $0 < \underline{\gamma} \leq 0.5$	5.98	5.38	7.47	10.01
- $0.5 < \underline{\gamma} \leq 1$	29.71	23.83	48.21	24.71
- $1 < \underline{\gamma} \leq 1.5$	37.28	39.53	28.22	29.70
- $1.5 < \underline{\gamma} \leq 2$	17.47	20.57	7.37	14.15
- $2 < \underline{\gamma} \leq 2.5$	1.01	2.00	0.63	3.87
- $2.5 < \underline{\gamma} \leq 3$	0.30	0.40	0.20	1.10
- $3 < \underline{\gamma}$	0.33	0.45	0.21	3.28

β bench. is the benchmark value found in Section 4.1.

β price is found using relative prices as weights in (14).

β logit is found by regressing a "sold" dummy on advert characteristics for transactions with $J = 2$ and $p_i > p_j$.

ψ param. and ψ kernel refer to the parametric (benchmark) and kernel specifications presented in Section 4.2.

Alternative estimation sample. So far we have focused on CD transactions during the third quarter of 2007. We now show that our results still hold when considering other time periods or another product category (DVD). We use three alternative samples: CD transactions during the first quarter of 2007, CD transactions during the second quarter of 2007 and DVD transactions during the third quarter of 2007. In all cases, we use our benchmark specification for β and ψ. We also use the same selection criteria for products (catalog price between 10 and 25€) and transactions (no advert has a price below 1 or above 20€) as in our benchmark sample (see section 3.3). We present in Table 15 the main estimation results obtained when using each of these samples and, for comparison, the benchmark results from Section 5.2.

Table 15: Results using alternative estimation samples

	Estimation sample			
	CD 2007Q3	CD 2007Q2	CD 2007Q1	DVD 2007Q3
pass rate (%) among transactions				
- any	93.69	92.40	93.36	86.34
- with no 'better' advert	99.50	99.59	99.52	99.16
- with ≥ 1 'better' advert	76.30	70.97	76.35	60.52
share (%) of explained transactions with $\underline{s} > 0$				
- any	26.28	24.75	27.95	31.49
- with no 'better' advert	9.44	8.92	10.21	13.08
- with ≥ 1 'better' advert	92.08	90.98	91.72	92.30
share (%) of explained transactions with ≥ 1 'better' advert and:				
- $\underline{s} = 0$	7.92	9.02	8.28	7.70
- $0 < \underline{s} \leq 0.5$	3.23	1.09	1.61	5.32
- $0.5 < \underline{s} \leq 1$	3.57	1.23	3.15	4.85
- $1 < \underline{s} \leq 2$	12.19	5.77	7.98	9.43
- $2 < \underline{s} \leq 3$	7.89	5.06	6.66	4.56
- $3 < \underline{s} \leq 4$	21.83	21.61	22.76	15.78
- $4 < \underline{s} \leq 5$	12.11	13.52	15.49	11.73
- $5 < \underline{s}$	31.25	42.70	34.08	40.63
share (%) of explained transactions with ≥ 1 'better' advert and:				
- $\underline{\gamma} = 0$	7.92	9.02	8.28	7.70
- $0 < \underline{\gamma} \leq 0.5$	5.98	3.61	2.83	3.86
- $0.5 < \underline{\gamma} \leq 1$	29.71	10.40	5.30	3.08
- $1 < \underline{\gamma} \leq 1.5$	37.28	31.86	19.39	11.27
- $1.5 < \underline{\gamma} \leq 2$	17.47	29.04	55.78	43.62
- $2 < \underline{\gamma} \leq 2.5$	1.01	9.57	7.62	16.83
- $2.5 < \underline{\gamma} \leq 3$	0.30	3.34	0.64	7.77
- $3 < \underline{\gamma}$	0.33	3.16	0.16	5.87

The main results on search costs and preferences using these alternative samples are similar to those we obtain in our benchmark case (shown in the first column of Table 15). The second and third columns of Table 15 show that results for CDs still hold if we consider the first two quarters of 2007. In the last column,

we see that consumer preferences and search costs also play an important role in DVD transactions. Even though the fit is slightly lower than for CDs,[29] we note that our model can explain 86% of all transactions and 60% of transactions where a 'better' advert was available. Results also show that some consumers are willing to pay substantially more for better advert characteristics and can face relatively high search costs. In particular, 31% of the explained transactions for DVDs are not consistent with a perfect information model.

6 Conclusion

In this paper we have conducted a structural analysis of consumer preferences and search costs using a directed sequential search model with flexible heterogeneity along these two dimensions. Our approach can account for a wide range of search patterns, where the sampling order depends on both the individual preferences of consumers and on their search cost. In particular consumers may not necessarily sample adverts by monotonous (decreasing or increasing) price order. Indeed, having strong preferences for non-price characteristics and expecting these characteristics to improve with the advert price may lead one to sample expensive adverts first. We are not aware of an empirical analysis, not necessarily using Internet data, based on this type of sequential, directed and preference-driven search model.

As far as we know, this paper also innovates on the methodological front by taking a revealed-preferences-based empirical approach to a search model. To this end, we choose to set identify and estimate our model, using a characterisation of the sets of preference and search cost parameters that follows from the optimal search-and-purchase strategies. This allows us to highlight the important role played by search costs and individual preferences in the transactions taking place on PriceMinister. In particular, we show that a flexible modeling of unobserved heterogeneity is important. Indeed, while the majority of transactions could be explained by a perfect information model, we find that a substantial share of purchases (more than a fourth) must have been made by consumers facing positive, sometimes high, search costs.

There are two directions in which we could extend our work. First, keeping a partial equilibrium analysis, one could try to enrich further the modeling of consumer search by allowing for consumers' beliefs to change across draws (i.e. learning) or for an unobserved advert characteristics. Each of these two extensions would be interesting but raises challenging modeling issues that would lead to a more parametric approach and thus do not really fit in the flexible framework used in this paper.

A possible direction for future extension of this framework would consist in closing our model and thus solving for sellers' optimal price posting strategy given consumers' search behaviour. The characterization of the identified sets that we derived from the optimal search-and-purchase strategies could be used to compute a seller's probability of making a transaction at a given price conditionally on the distribution of preferences and search costs. This distribution could be parametrized whilst making sure that it fits within the bounds found in our partial equilibrium analysis. We could then try to assess the respective roles of the heterogeneity in consumer preferences, search costs and seller characteristics in the substantial price dispersion observed

[29] This small loss in the fit of the model comes from the fact that the expected scalar hedonic index increases less with price for DVDs than it does for CDs. We would thus need higher values of the marginal willingness to pay γ to explain more transactions. We prefer to document the fit performance of our model whilst keeping reasonable values of γ and s.

in the Internet and documented in this paper. Further issues, however, would arise from the fact that sellers advertise many products at once and may have a global strategy in terms of their reputation and prices that goes beyond what we observe for a specific product.

References

Baye, M., J. Morgan, and P. Scholten (2004): "Price Dispersion in the Small and in the Large: Evidence from an Internet Price Comparison Site," *Journal of Industrial Economics*, 52, 463–496.

Blow, L., M. Browning, and I. Crawford (2008): "Revealed Preference Analysis of Characteristics Models," *Review of Economic Studies*, 75, 371–89.

Cherchye, L. Crawford, I., B. De Rock, and F. Vermeulen (2009): "The Revealed Preference Approach to Demand," in *Quantifying Consumer Preferences*, ed. by D. Slottje, Bingley, UK: Emerald Books.

Cosaert, S. and T. Demuynck (2014): "Revealed preference theory for finite choice sets," *Economic Theory*, 56, 1–32.

De Los Santos, B., A. Hortaçsu, and M. Wildenbesst (2012): "Testing Models of Consumer Search Using Data on Web Browsing and Purchasing Behavior," *American Economic Review*, 102, 2955–80.

——— (2013): "Search with Learning," mimeo.

Dinerstein, M., L. Einav, J. Levin, and N. Sundaresan (2014): "Consumer Price Search and Platform Design in Internet Commerce," mimeo.

Hong, H. and M. Shum (2006): "Using price distributions to estimate search costs," *The RAND Journal of Economics*, 37, 257–275.

Hortaçsu, A. and C. Syverson (2004): "Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds," *Quarterly Journal of Economics*, 119, 403–56.

Janssen, M. and J. Moraga-Gonzales (2004): "Strategic Pricing, Consumer Search and the Number of Firms," *Review of Economic Studies*, 71, 1089–1118.

Koulayev, S. (2014): "Estimating demand in online search markets, with application to hotel bookings," *RAND Journal of Economics*, 45, 553–75.

Moraga-Gonzales, J. and V. Petrikaite (2013): "Search Costs, Demand-Side Economies and the Incentives to Merge under Bertrand Competition," *Rand Journal of Economics*, 44, 391–424.

Morgan, P. and R. Manning (1985): "Optimal Search," *Econometrica*, 53, 923–44.

Silverman, B. (1998): "Density Estimation for Statistics and Data Analysis," London: Chapmanhall.

Weitzman, M. (1979): "Optimal Search for the Best Alternative," *Econometrica*, 47, 641–54.

Zhou, J. (2011): "Ordered Search in Differentiated Markets," *International Journal of Industrial Organization*, 29, 253–262.

APPENDIX

A Construction of the sample

The administrative data we obtained from PriceMinister consist of essentially two tables. In the first table, all transactions that took place on the website until December 2008 are recorded. For each transaction, we observe, among other things, the seller id, the product id, the advert id (not the buyer's), the price, the exact date when the transaction was initiated/completed, the seller's status (professional or individual) and the feedback. With this information we can thus compute for each seller at any given date his size (number of completed transactions so far) and his reputation (average feedback received so far). We observe a seller's status unless he has no transactions, initated or completed. We assume that sellers who never appear in the transaction table are private individuals (expecting professional sellers to have at least one contact with a buyer during the observation period).

The second main table contains all the adverts posted on the website, with information on advert id, seller id, product id, the condition of the good (new, as new, etc.), list price, the precise date when the advert was posted and whether the advert is still active at the data extraction date.

We combine these two tables leads to produce a dataset that, for any transaction in a given time period and product category (in our benchmark case, CDs during the last quarter of 2007), provides information on all the relevant information on the adverts for the exact same product available at the time when the transaction took place. To this end, we had to solve two problems, as we explain below.

The first issue is that the match on advert id between the transaction and the advert tables is not perfect. For a small proportion of CDs (which is the product category we are interested in), there exists one or several transactions for which the advert id is not found in the advert table. We cannot just get rid of these transactions because the advert that was bought on this occasion may have been available to consumers in other transactions. After some investigation, we think that this problem, which again only concerns a small minority of CDs, can be caused by adverts that are sold very quickly, within a few hours of being posted and thus appear in the transaction table but have not yet been included in the advert table. To make sure that these adverts do not interfere with other transactions for the same product, we drop all observations for a given product during the day when such a mismatched transaction takes place. A more drastic solution would consist in leaving out of the sample all the CDs for which this mismatch takes place, at any time. This would however take out the bestselling products and severily decrease the number of transactions. A previous version of the paper used that correction and the results were qualitatively similar to those shown in this version. Hence, we chose to keep as many products as possible, including the bestselling ones, and to apply the first, less drastic, solution to solve the mismatch problem.

The second problem we had to face pertains to adverts' end date i.e. when adverts disappear from the website. With our advert data we know when an advert is created and whether it is still active in December 2008. There are thus adverts for which the end date is not directly observed and we had to construct these dates based on a few assumptions. If an advert is still active in December 2008, we assume that it has been active since its creation. If an advert is no longer active at the extraction date and has led to at least one transaction, we assume that it was closed (taken off the screen) right after its last observed transaction. This is to reflect the fact that the seller ran out of stocks after the last transaction. Indeed, most sellers are individuals who probably have only one copy to sell and, if they were professional, they would not take out an advert for a product that sells and that they still have in stock. The last and most difficult case is an advert inactive in December 2008 and which was never sold. We assume that these adverts were taken off by sellers after a given time period which we compute as follows. We consider the distribution of durations between a CD advert's first transaction and its creation (conditionally on being sold at least once) and we take the 95% quantile of this distribution. Hence, we assume that sellers take out adverts that do not generate at least one transaction within a relatively long time interval.