University of St Andrews

School of Economics and Finance Discussion Papers

Inferring Cognitive Heterogeneity from Aggregate Choices

Valentino Dardanoni, Paola Manzini, Marco Mariotti and Christopher J. Tyson

School of Economics and Finance
Online Discussion Paper Series
issn 2055-303X
http://ideas.repec.org/s/san/wpecon.html
info: econ@st-andrews.ac.uk

School of Economics and Finance Discussion Paper No. 1701
13 Feb 2017 (revised 25 May 2017)
JEL Classification: D01, D03, D12
Keywords: attention, bounded rationality, revealed preference, stochastic choice

Inferring Cognitive Heterogeneity from Aggregate Choices*

Valentino Dardanoni
Università degli Studi di Palermo

Paola Manzini
University of St Andrews and IZA

Marco Mariotti
Queen Mary University of London

Christopher J. Tyson
Queen Mary University of London

May 25, 2017

Abstract

We study the problem of recovering the distribution of cognitive characteristics in a population of boundedly rational agents from their aggregate choices. In contrast to much of the theoretical literature on bounded rationality, we make use of choices from a fixed menu of alternatives rather than from a rich collection of menus. We examine in detail two models of choice with limited attention, and we show that both "consideration probability" and "consideration capacity" distributions are substantially identified by aggregate choice shares. These two models are applied to data on over-the-counter painkiller sales, yielding concurrent estimates that on average two or three out of the eight available products are considered in this market. We also demonstrate how our framework can be used "counterfactually" for welfare analysis.

Keywords: attention, bounded rationality, revealed preference, stochastic choice.
J.E.L. codes: D01, D03, D12.

1 Introduction

1.1 Motivation

Individuals differ not only in their tastes, but also in their cognitive characteristics. They have different computational abilities, different levels of strategic and epistemic sophistication, and different capacities for extracting relevant information from their environment. In this paper we develop tools for studying the choices of agents who differ in the degree to which they

*We are grateful to Jason Abaluck, Abi Adams, Reyer Gerlagh, Alessandro Iaria, Rod McCrorie, and Irina Merkurieva for comments and suggestions, as well as to seminar audiences at CRETA, ECARES, Tilburg University, the University of Edinburgh, the University of Vienna, the University of St Andrews, the Barcelona GSE Summer Forum 2016, D-TEA 2016, and the Seventh Italian Congress of Econometrics and Empirical Economics. We are indebted to Vishal Singh for making available the dataset used in the empirical exercise in Section 4. Manzini and Mariotti thank the ESRC for financial support provided through grant ES/J012513/1.

attend to the available alternatives. In the marketing literature, the subset of alternatives that an agent actively investigates is known as the *consideration set*, and this will be the specific manifestation of attention that we focus on. But we view the present exercise as merely one example of how cognitive heterogeneity can be employed to broaden the scope of economic models.[1]

Classical revealed preference analysis has studied extensively the correspondence between unobserved tastes and observed choices. More recently, theoretical work on boundedly rational decision making has extended this methodology to incorporate a range of cognitive factors, including computational constraints, norms and heuristics, reference points and other framing effects, and various conceptions of attention.[2] One potential drawback of such theories is that they typically take as given a rich dataset—comprising the same individual's choices from many different overlapping menus—that can be used to identify the latent components of the cognitive model of interest.[3] But idealized data of this sort may not be available in practice, particularly when the category of decision problem arises rarely (e.g., choice of hospital provider for elective surgery) or the menu is slow to change (e.g., choice of daily newspaper). In these contexts "full identification" results furnished by the theoretical literature on bounded rationality may be unrealistically data hungry, and alternative approaches are needed to link theory to what is feasible empirically.

To address this problem, our theoretical approach takes as primitive a dataset consisting of a single, fixed menu faced by a population of agents, together with the observed aggregate choice shares of the alternatives. Cognitive heterogeneity leads individuals to choose differently from this menu even when they have homogeneous tastes—an assumption that we impose to clarify the theoretical analysis, but which can be relaxed.[4] Given a particular cognitive model, our research question then becomes whether the distribution of parameters in the model can be inferred from the choice shares, in the same way that preference parameters are revealed by individual choice data.

[1] Another example is supplied by Kneeland's [18] analysis of heterogeneity in the cognitive characteristic of higher-order rationality.

[2] Contributions to this literature include, among others, Apesteguia and Ballester [2], Baigent and Gaertner [3], Caplin and Dean [6], Caplin et al. [7], Cherepanov et al. [8], de Oliveira et al. [12], Eliaz and Spiegler [13], Manzini and Mariotti [24, 25], Masatlioglu and Nakajima [26], Masatlioglu et al. [27], Ok et al. [35], Salant and Rubinstein [36], and Tyson [42, 43].

[3] For instance, Masatlioglu et al. [27] require choice data for all possible menus, while Manzini and Mariotti [25] impose a weaker but still stringent "richness" assumption. Note that even stronger background assumptions about data availability are commonplace in the theory of choice under uncertainty, where the decision maker is typically asked to report preferences over a highly structured mathematical space designed to facilitate identification.

[4] Taste heterogeneity is discussed in Section 3.4.2 and allowed for in the empirical exercise in Section 4. Note that for this generalization to be feasible, it will be essential that the cognitive and taste heterogeneity be statistically independent.

1.2 Two models of consideration set formation

In our general framework, each agent has a *cognitive type* parameter $\theta \in \Theta$ that is distributed in the population according to F. Assuming that tastes are homogeneous and known, an individual of type θ chooses alternative x with probability $p_\theta(x)$, and the corresponding aggregate choice share is

$$p(x) = \int_\Theta p_\theta(x)\, dF. \tag{1}$$

To the extent that the cognitive model captures some form of bounded rationality, neither the individual probabilities nor the population shares will put all weight on the best available option (according to the common preferences of the agents). Indeed, the fact that suboptimal alternatives are sometimes chosen is what will allow us to infer the unobserved cognitive distribution F from the observed aggregate shares.[5]

More specifically, our interest in this paper is in bounded rationality as limited attention, and here the parameter θ will affect the formation of the consideration set.[6] We will study in detail two different models of how this occurs, referred to below as the *ρ-model* and the *γ-model*. The first is a variant of the consideration set structure in Manzini and Mariotti [25] and has cognitive parameter $\rho \in [0,1]$, expressing the probability that each alternative is considered and interpreted as the agent's degree of general awareness of the decision making environment. In contrast, the second model has parameter $\gamma \in \{0,1,2,\ldots\}$, equal to the cardinality of the consideration set and interpreted as a constraint on the number of alternatives that the agent can actively investigate at any one time.[7] Both models assume preferences are maximized over the consideration set; both include full rationality as a special case (respectively, $\rho = 1$ and $\gamma \to \infty$); and both specify a default consequence in case the consideration set is empty.

1.3 Preview of results

We show first that, for several natural parameterizations of our consideration set models, the cognitive distribution F can be inferred from a sufficient number of observed aggregate shares. For instance, if the consideration probability ρ is uniformly distributed on an interval, then the bounds of this interval can be recovered from the shares of the two most preferred alternatives (Example 3). We then proceed to show that for large menus, the entire cognitive distribution

[5] Note that our framework has similarities to mixed models in the discrete choice literature (see Train [41] and McFadden [30]), where θ would be a taste parameter such as the unobserved marginal utility of some observed characteristic. However, since we shall use p_θ to express hypotheses about cognition instead of tastes, our specification calls for different functional form assumptions. In particular, p_θ will not be a logit conditional on θ (see Luce [21]), as would typically be assumed in relation to tastes.

[6] Although we adopt a bounded-rationality interpretation of consideration sets, it is worth noting that other interpretations are possible. Indeed, alternatives may fail to be considered due to habit formation, search costs, or other forms of rational inattention (see, e.g., Caplin and Dean [6] and Sims [39]).

[7] In the context of a price competition game, de Clippel et al. [11] use a similar model to implement consumers with limited attention.

can for practical purposes be recovered even in the absence of parametric assumptions. More precisely, in the γ-model the aggregate shares identify the probabilities of all consideration set sizes that are less than the cardinality n of the menu (Proposition 1). Turning to the ρ-model, the shares are seen to identify the first n raw moments of F (Proposition 2). Using maximum entropy techniques and sparsity analysis, we find that these moments can under mild conditions be used to reconstruct or closely approximate the distribution itself (Propositions 3–4). In each case, the identification is due to the system of equations that define the aggregate shares being recursive and linear in the relevant quantities (i.e., the probabilities or raw moments), so that explicit formulas for these quantities can be obtained by inverting a triangular or anti-triangular matrix.

Finally, we provide a circumscribed empirical exercise that applies our theoretical results to data on over-the-counter painkiller sales. Allowing for both cognitive and taste heterogeneity (the latter using a standard logit specification), we estimate suitably parameterized versions of both models of consideration set formation using maximum likelihood. The average value of ρ is estimated at 0.32, and the average value of γ at 2.1. The estimated models roughly agree on the mean number of painkiller products considered—between two and three out of the eight available—and there is no evidence to support a hypothesis that the entire menu is considered in this setting. An advantage of the methodology is that our "deep parameter" estimates may be used to compute consumer surplus under both the status quo and counterfactual scenarios, a comparison that can form the basis for policy analysis. Moreover, our contribution can be seen as complementary to existing empirical work on limited attention: For example, Crawford et al. [10] show in a model-free context that ignoring consideration-set effects may lead to biased estimates of preference parameters, whereas we use structural models of attention to jointly estimate tastes and cognitive characteristics.[8]

1.4 Outline

The remainder of the paper is structured as follows. Section 2 describes our general framework and introduces the two specific models of consideration set formation. Section 3 derives our theoretical results on inferring the cognitive type distribution from aggregate choice shares, and also discusses the issues raised by indifference and taste heterogeneity. Section 4 contains the empirical exercise, and Section 5 concludes.

[8]Other relevant papers include those by Sovinsky Goeree [40], Van Nierop et al. [34], Abaluck and Adams [1], Lu [20], and Honka et al. [16]. See Section 4.5 for a brief review of the empirical literature in this area.

2 Two models of consideration set formation

2.1 General framework

Let X denote the universal set of alternatives. A *menu* is any nonempty $A \subseteq X$, with which is associated a default outcome d_A (not in A). When faced with the menu A, an agent either chooses exactly one of the available alternatives or chooses none and accepts d_A. For example, we could have that:

(i) the menu contains retailers selling a particular product, and the default is not to buy the product at all;

(ii) the menu contains banks offering fixed-term deposits, and the default is to keep the money in cash; or

(iii) the menu contains risky lotteries, and the default is a risk-free payment.

In order to concentrate on cognitive rather than taste heterogeneity, we shall assume in the theoretical part of the paper (i.e., Sections 2–3) that all agents share the same preferences over the alternatives. Equivalently, this can be thought of as using the average utilities of alternatives in the population, ignoring variation. In this sense our approach is dual to that of the classical stochastic-choice literature in economics, where preferences are allowed to vary but cognitive capabilities are implicitly assumed to be uniform. Observe that homogeneous tastes are plausible in examples (i) and (ii) above, where preferences will be determined largely by price and interest rate comparisons, as well as in example (iii) if all agents are approximately risk neutral over the relevant stakes. In view of this assumption (relaxed in Section 3.4.2), we denote by k_A the kth best option on menu A according to the unanimous preferences, so that the full menu appears as $A = \{1_A, 2_A, \ldots, |A|_A\}$. We emphasize (see Footnote 4) that both cognitive and taste heterogeneity are allowed for in the empirical exercise in Section 4.

We model cognitive heterogeneity by assigning each agent a cognitive type $\theta \in \Theta \subset \Re$, drawn independently across agents from the distribution F. We write $p_\theta(k_A)$ for the probability that type θ chooses alternative k_A from menu A, and

$$p(k_A) = \int_\Theta p_\theta(k_A)\, \mathrm{d}F \tag{2}$$

for the overall share in the population. Similarly, we write $p_\theta(d_A)$ for the probability that type θ accepts the default consequence, and

$$p(d_A) = \int_\Theta p_\theta(d_A)\, \mathrm{d}F \tag{3}$$

for the population share. For each $\theta \in \Theta$ we have $\sum_{k=1}^{|A|} p_\theta(k_A) = 1 - p_\theta(d_A)$, and likewise in

aggregate $\sum_{k=1}^{|A|} p(k_A) = 1 - p(d_A)$. When we wish to emphasize the role of the type distribution in determining the choice probabilities, we write $p(k_A; F)$ and $p(d_A; F)$.

The basic scenario of interest involves the members of a large population choosing from a fixed menu M, with $|M| = n$. The analyst observes the aggregate choice shares, but does not know the cognitive type distribution F. In this context we shall generally suppress dependence on M, writing $p_\theta(k)$ and $p_\theta(d)$ for the type-specific frequencies and $p(k)$ and $p(d)$ for the population shares. Our goal is to deduce information about the type distribution from the shares $\langle p(1), p(2), \ldots, p(n), p(d) \rangle$, and to use this knowledge to predict aggregate choice behavior from menus other than M.

We proceed now to specialize this framework to two more concrete models illustrating different ways that the agents' attention to the alternatives may be limited. Among the alternatives that are considered (i.e., that attract attention), each agent will choose the best option according to the shared preference order. But since the alternatives considered may be a strict subset of those actually available, the attention deficits captured in the two specialized models can lead to sub-optimal decision making.

2.2 Consideration probability: The ρ-model

Let $\rho \in [0, 1] = \Theta$ denote the probability that the agent considers each alternative on the menu, with consideration independent across agents and alternatives.[9] In this case alternative k will be chosen if and only if the agent both (i) notices k and (ii) fails to notice each alternative $l < k$; whereas the default consequence will arise if no alternatives at all are noticed. The type-conditional choice frequencies are therefore

$$p_\rho(k) = \rho[1 - \rho]^{k-1}, \tag{4}$$

$$p_\rho(d) = [1 - \rho]^n; \tag{5}$$

with corresponding aggregate shares

$$p(k) = \int_0^1 \rho[1 - \rho]^{k-1} \mathrm{d}F, \tag{6}$$

$$p(d) = \int_0^1 [1 - \rho]^n \mathrm{d}F. \tag{7}$$

Example 1. [*Uniform* ρ] If the consideration probability ρ is distributed uniformly on the interval $[\rho_{\min}, \rho_{\max}]$, with $0 \leq \rho_{\min} < \rho_{\max} \leq 1$, then the cognitive distribution is

$$F(\rho) = \frac{\rho - \rho_{\min}}{\rho_{\max} - \rho_{\min}}. \tag{8}$$

[9] Variants of this model have been studied by Manzini and Mariotti [25] and Brady and Rehbeck [5].

In this case Equations 6–7 take the form

$$p\,(k) = \int_{\rho_{\min}}^{\rho_{\max}} \frac{\rho[1-\rho]^{k-1}\mathrm{d}\rho}{\rho_{\max}-\rho_{\min}} = \frac{[1+k\rho_{\min}][1-\rho_{\min}]^k - [1+k\rho_{\max}][1-\rho_{\max}]^k}{k[k+1][\rho_{\max}-\rho_{\min}]}, \tag{9}$$

$$p(d) = \int_{\rho_{\min}}^{\rho_{\max}} \frac{[1-\rho]^n\mathrm{d}\rho}{\rho_{\max}-\rho_{\min}} = \frac{[1-\rho_{\min}]^{n+1} - [1-\rho_{\max}]^{n+1}}{[n+1][\rho_{\max}-\rho_{\min}]}. \square \tag{10}$$

2.3 Consideration capacity: The γ-model

Let $\gamma \in \{0,1,2,\ldots\} = \Theta$ denote the number of alternatives that the agent is able to consider; that is, the "consideration capacity." When $\gamma < n$ we assume that the agent is equally likely to consider each $\Gamma \subset M$ with $|\Gamma| = \gamma$, and when $\gamma \geq n$ we know that the entire menu M will be considered. In the former case there are clearly $\binom{n}{\gamma}$ candidate consideration sets. Of these, exactly $\binom{n-k}{\gamma-1}$ both contain alternative k and do not contain any superior alternative $\ell < k$. The probability of k being chosen is thus $\binom{n-k}{\gamma-1}/\binom{n}{\gamma}$. Note that this probability is 0 for $k > n-\gamma+1$, since here there are fewer than $\gamma-1$ alternatives inferior to k that can populate the consideration set in order to allow k to be chosen. Of course, whenever the entire menu is considered we can be certain that alternative 1 will be chosen, regardless of the precise value of $\gamma \geq n$.

The type-conditional choice frequencies can now be expressed as

$$p_\gamma(k) = \begin{cases} \binom{n-k}{\min\{\gamma,n\}-1}/\binom{n}{\min\{\gamma,n\}} & \text{if } \gamma > 0, \\ 0 & \text{if } \gamma = 0; \end{cases} \tag{11}$$

$$p_\gamma(d) = \begin{cases} 0 & \text{if } \gamma > 0, \\ 1 & \text{if } \gamma = 0. \end{cases} \tag{12}$$

Defining the probability masses

$$\pi(0) = F(0), \tag{13}$$

$$\forall \gamma \geq 1, \quad \pi(\gamma) = F(\gamma) - F(\gamma-1); \tag{14}$$

the corresponding aggregate shares are

$$p(1) = \sum_{\gamma=1}^{\infty} \frac{\min\{\gamma,n\}}{n}\pi(\gamma), \tag{15}$$

$$\forall k : 2 \leq k \leq n, \quad p(k) = \sum_{\gamma=1}^{n-k+1} \frac{\binom{n-k}{\gamma-1}}{\binom{n}{\gamma}}\pi(\gamma), \tag{16}$$

$$p(d) = \pi(0). \tag{17}$$

Example 2. [*Poisson* γ] Consider the Poisson specification $\pi(\gamma) = \frac{\mu^\gamma e^{-\mu}}{\gamma!}$ for $\mu \geq 0$. In this case Equations 15–17 take the form

$$p(1) = e^{-\mu} \sum_{\gamma=1}^{\infty} \frac{\min\{\gamma, n\}}{n} \frac{\mu^\gamma}{\gamma!}, \tag{18}$$

$$\forall k : 2 \leq k \leq n, \quad p(k) = e^{-\mu} \sum_{\gamma=1}^{n-k+1} \frac{\binom{n-k}{\gamma-1}}{\binom{n}{\gamma}} \frac{\mu^\gamma}{\gamma!}, \tag{19}$$

$$p(d) = e^{-\mu}. \square \tag{20}$$

3 Inferring the cognitive type distribution

3.1 Parametric analysis

We first consider a variety of plausible functional forms for the cognitive type distribution, aiming to find tractable expressions that relate cognitive parameters to aggregate choice shares. As we shall see, under several natural parameterizations it is possible to identify the type distribution uniquely from a small number of appropriately selected share observations. Apart from increasing our familiarity with the two models under investigation, the main purpose of the examples below is to highlight the non-obvious ways that choice shares can convey information about cognitive parameters.

Example 3. [*Uniform* ρ] Given the distribution in Example 1 for the consideration probability ρ, we have the choice shares

$$p(1) = \frac{[1+\rho_{\min}][1-\rho_{\min}] - [1+\rho_{\max}][1-\rho_{\max}]}{2[\rho_{\max} - \rho_{\min}]} = \frac{\rho_{\max} + \rho_{\min}}{2}, \tag{21}$$

$$\begin{aligned} p(2) &= \frac{[1+2\rho_{\min}][1-\rho_{\min}]^2 - [1+2\rho_{\max}][1-\rho_{\max}]^2}{6[\rho_{\max} - \rho_{\min}]} \\ &= \frac{\rho_{\max} + \rho_{\min}}{2} - \frac{\rho_{\max}^2 + \rho_{\max}\rho_{\min} + \rho_{\min}^2}{3}. \end{aligned} \tag{22}$$

Solving Equations 21–22 then yields the parameter values

$$\rho_{\max} = p(1) + \sqrt{3[p(1) - p(2) - p(1)^2]}, \tag{23}$$

$$\rho_{\min} = p(1) - \sqrt{3[p(1) - p(2) - p(1)^2]}. \square \tag{24}$$

Example 4. [*Poisson* γ] Given the distribution in Example 2 for the consideration capacity γ, we have the default share $p(d) = \exp[-\mu]$ and hence $\mu = -\log p(d)$. $\square$

In each of these two examples, the full type distribution can be retrieved from as many choice

share observations as there are cognitive parameters. In the uniform ρ-model the identifying shares are those of the two best alternatives, while in the Poisson γ-model it is the share of the default outcome.

We next supply two-parameter functional forms for our two cognitive models in which identification of the type distribution is more challenging. In the first example, to infer the distribution of the consideration capacity we can use the shares of the two worst alternatives, the default share, and the size of the menu.

Example 5. [*Pascal* γ] Consider the Pascal (or "negative binomial") specification with $\pi(\gamma) = \binom{\gamma+r-1}{\gamma}[1-q]^r q^\gamma$, for $r \in \{1,2,3,\ldots\}$ and $q \in (0,1)$. In this case Equations 15–17 take the form

$$p(1) = [1-q]^r \sum_{\gamma=1}^{\infty} \frac{\min\{\gamma, n\}}{n} \binom{\gamma+r-1}{\gamma} q^\gamma, \tag{25}$$

$$\forall k : 2 \leq k \leq n, \quad p(k) = [1-q]^r \sum_{\gamma=1}^{n-k+1} \frac{\binom{n-k}{\gamma-1}}{\binom{n}{\gamma}} \binom{\gamma+r-1}{\gamma} q^\gamma, \tag{26}$$

$$p(d) = [1-q]^r. \tag{27}$$

When $n \geq 3$ we can compute the share ratios

$$\frac{p(n)}{p(d)} = \frac{qr}{n}, \tag{28}$$

$$\frac{p(n-1)}{p(n)} = 1 + \frac{q[r+1]}{n-1}; \tag{29}$$

allowing us to express the parameters as

$$q = [n-1]\left[\frac{p(n-1)}{p(n)} - 1\right] - \frac{np(n)}{p(d)}, \tag{30}$$

$$r = \frac{np(n)^2}{p(d)[n-1]\left[p(n-1) - p(n)\right] - np(n)^2}. \square \tag{31}$$

The next example involves a distribution for the consideration probability that will be used later for the empirical exercise in Section 4. Here we can obtain closed-form expressions for the parameters in two special cases, though not in general.

Example 6. [*Kumaraswamy* ρ] Consider the Kumaraswamy specification having distribution $F(\rho) = 1 - \left[1 - \rho^a\right]^b$, for $a, b > 0$.

If $b = 1$, then we have $F(\rho) = \rho^a$. In this case Equations 6–7 appear as

$$p(k) = a \int_0^1 \rho^a [1-\rho]^{k-1} \, \mathrm{d}\rho = aB(a+1, k), \tag{32}$$

$$p(d) = a \int_0^1 \rho^{a-1} [1-\rho]^n \mathrm{d}\rho = aB(a, n+1); \tag{33}$$

where B is the beta function.[10] From $p(1) = aB(a+1, 1) = \frac{a}{a+1}$ we then obtain

$$a = \frac{p(1)}{1-p(1)}. \tag{34}$$

Alternatively, if $a = 1$ then we have $F(\rho) = 1 - [1-\rho]^b$. In this case the shares are

$$p(k) = b \int_0^1 \rho [1-\rho]^{k+b-2} \mathrm{d}\rho = bB(2, k+b-1) = \frac{b}{[k+b][k+b-1]}, \tag{35}$$

$$p(d) = b \int_0^1 [1-\rho]^{n+b-1} \mathrm{d}\rho = bB(1, n+b) = \frac{b}{n+b}. \tag{36}$$

From $p(d) = \frac{b}{n+b}$ we then obtain

$$b = \frac{np(d)}{1-p(d)}. \tag{37}$$

In the general case, Equations 6–7 take the form

$$p(k) = ab \int_0^1 \rho^a [1-\rho]^{k-1} [1-\rho^a]^{b-1} \, \mathrm{d}\rho, \tag{38}$$

$$p(d) = ab \int_0^1 \rho^{a-1} [1-\rho]^n [1-\rho^a]^{b-1} \, \mathrm{d}\rho. \tag{39}$$

Using Equation 38, we can write the first two moments of the ρ distribution as

$$m_1 = ab \int_0^1 \rho^a [1-\rho^a]^{b-1} \, \mathrm{d}\rho = p(1), \tag{40}$$

$$\begin{aligned} m_2 &= ab \int_0^1 \rho^{a+1} [1-\rho^a]^{b-1} \, \mathrm{d}\rho \\ &= ab \int_0^1 [1-[1-\rho]] \rho^a [1-\rho^a]^{b-1} \, \mathrm{d}\rho \\ &= p(1) - p(2). \end{aligned} \tag{41}$$

This suggests that the difficulty of expressing the parameters in terms of the choice shares is primarily due to the difficulty of inverting the map $\langle a, b \rangle \mapsto \langle m_1, m_2 \rangle$ for this functional form, rather than to any feature of the consideration probability model itself. □

The observation that moments m_j of the Kumaraswamy distribution can be expressed as

[10] Recall that the beta function is defined by $B(y, z) = \int_0^1 t^{y-1}[1-t]^{z-1}\mathrm{d}t$, for $y, z > 0$.

weighted sums of the choice shares extends to values of $j > 2$, and is in fact a general feature of the ρ-model. This property is exploited in the nonparametric analysis of the consideration probability model in Sections 3.2–3.3 below.

3.2 Nonparametric analysis

3.2.1 The nonparametric inference problem

The examples in the previous section have shown a variety of ways that information about the type distribution F can be encoded in the choice shares, depending on the cognitive model and the specific parameterization employed. In this section, in contrast, we turn to the general structure of the inference problem. We shall see that identification of the type distribution remains tractable in both models even without parametric assumptions on F. This is because the choice shares are linear in the type probabilities $\pi(\gamma)$ in the consideration capacity model, and linear in the moments m_j of F in the consideration probability model. Moreover, each system of equations has a simple triangular structure that enables it to be solved recursively, using one additional choice share at each step.

These observations about the inference problem tell us that under either of the two cognitive models, the information encoded in the choice shares can be decoded by inverting a triangular matrix of dimension n, the number of alternatives. The larger is the observed menu, the more detailed will be the picture of F that is revealed by aggregate choice data. In the γ-model increasing the size of the menu to $n+1$ will yield an extra type probability $\pi(n)$, while in the ρ-model such an increase will yield an extra moment m_n. In the latter case we can then use well-established technology (both maximum entropy methods and results from sparsity theory) to show that knowledge of the moments of F allows us to reconstruct—or to construct a good approximation of—the distribution itself (see Section 3.3).

3.2.2 The γ-model: Recovering n probabilities

Without functional form assumptions on F, choice shares in the consideration capacity model are given by Equations 15–17. The first two of these equations can be written together in matrix form as

$$\underbrace{\begin{bmatrix} p(1) \\ \vdots \\ p(k) \\ \vdots \\ p(n) \end{bmatrix}}_{\mathbf{p}} = \underbrace{\begin{bmatrix} \frac{1}{n} & \cdots & \frac{\gamma}{n} & \cdots & \frac{n-1}{n} & 1 \\ \vdots & & \vdots & & & \\ \frac{1}{n} & \cdots & \frac{\binom{n-k}{\gamma-1}}{\binom{n}{\gamma}} & & 0 & 0 \\ \vdots & & & & \vdots & \vdots \\ \frac{1}{n} & & 0 & \cdots & 0 & 0 \end{bmatrix}}_{\mathbf{C}} \underbrace{\begin{bmatrix} \pi(1) \\ \vdots \\ \pi(\gamma) \\ \vdots \\ \pi(n-1) \\ 1-F(n-1) \end{bmatrix}}_{\boldsymbol{\pi}}. \tag{42}$$

The upper anti-triangular matrix $\mathbf{C}$ has a lower anti-triangular inverse, allowing us to write $\boldsymbol{\pi} = \mathbf{C}^{-1}\mathbf{p}$. Indeed, we can compute the probabilities of the n smallest consideration capacities explicitly as

$$\pi(0) = p(d) = 1 - \sum_{k=1}^{n} p\left(k\right), \tag{43}$$

$$\forall \gamma : 1 \leq \gamma < n, \quad \pi\left(\gamma\right) = \binom{n}{\gamma} \sum_{k=n-\gamma+1}^{n} [-1]^{[\gamma-1]-[n-k]} \binom{\gamma-1}{n-k} p(k). \tag{44}$$

On the other hand, the probabilities $\pi(\gamma)$ for $\gamma \geq n$ cannot be inferred from the available data, since these consideration capacities are behaviorally indistinguishable in a setting with only n alternatives. We summarize our conclusions for this model as follows.

Proposition 1. *In the γ-model, the probabilities $\langle \pi\left(\gamma\right) \rangle_{\gamma=0}^{n-1}$ are uniquely determined by the aggregate choice shares $\langle p\left(k\right) \rangle_{k=1}^{n}$.*

3.2.3 The ρ-model: Recovering n moments

In the consideration probability model, the choice shares are given by Equations 6–7. We can expand the binomial in Equation 6 to yield

$$p(k) = \int_0^1 \rho \left[\sum_{j=0}^{k-1} \binom{k-1}{j} [-\rho]^j \right] \mathrm{d}F = \sum_{j=1}^{k} \left(-1\right)^{j-1} \binom{k-1}{j-1} m_j, \tag{45}$$

where $m_j = \int_0^1 \rho^j \mathrm{d}F$ is the jth raw moment of the type distribution. Similarly, Equation 7 can be expressed in terms of the moments as

$$p\left(d\right) = \int_0^1 \left[\sum_{j=0}^{n} \binom{n}{j} [-\rho]^j \right] \mathrm{d}F = 1 + \sum_{j=1}^{n} [-1]^j \binom{n}{j} m_j. \tag{46}$$

Equation 45 appears in matrix form as

$$\underbrace{\begin{bmatrix} p\left(1\right) \\ \vdots \\ p\left(k\right) \\ \vdots \\ p\left(n\right) \end{bmatrix}}_{\mathbf{p}} = \underbrace{\begin{bmatrix} 1 & & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 1 & \cdots & [-1]^{j-1}\binom{k-1}{j-1} & & 0 \\ \vdots & & \vdots & & \\ 1 & \cdots & [-1]^{j-1}\binom{n-1}{j-1} & \cdots & [-1]^{n-1} \end{bmatrix}}_{\mathbf{R}} \underbrace{\begin{bmatrix} m_1 \\ \vdots \\ m_j \\ \vdots \\ m_n \end{bmatrix}}_{\mathbf{m}}. \tag{47}$$

The lower triangular matrix $\mathbf{R}$ is involutory (i.e., equal to its own inverse), allowing us to write $\mathbf{m} = \mathbf{Rp}$. The n smallest moments of F are then given by

$$\forall j : 1 \leq j \leq n, \quad m_j = \sum_{k=1}^{j} [-1]^{k-1} \binom{j-1}{k-1} p(k). \tag{48}$$

We conclude the following.

Proposition 2. *In the ρ-model, the moments $\langle m_j \rangle_{j=1}^{n}$ are uniquely determined by the aggregate choice shares $\langle p(k) \rangle_{k=1}^{n}$.*

3.2.4 Relationship between the two models

In the ρ-model the same consideration probability applies independently to each option, and so all subsets of the menu of a given size are equally likely to be the consideration set. It follows that the ρ-model is a special case of the γ-model and Equation 42 holds in this case.

In the ρ-model, the probability that the consideration set contains exactly γ alternatives is

$$\begin{aligned}
\pi(\gamma) &= \int_0^1 \binom{n}{\gamma} \rho^\gamma [1-\rho]^{n-\gamma} \mathrm{d}F \\
&= \binom{n}{\gamma} \int_0^1 \rho^\gamma \left[\sum_{i=0}^{n-\gamma} \binom{n-\gamma}{i} [-\rho]^i \right] \mathrm{d}F \\
&= \binom{n}{\gamma} \sum_{i=0}^{n-\gamma} \binom{n-\gamma}{i} [-1]^i m_{\gamma+i} \\
&= \binom{n}{\gamma} \sum_{j=\gamma}^{n} \binom{n-\gamma}{j-\gamma} [-1]^{j-\gamma} m_j.
\end{aligned} \tag{49}$$

In matrix form, these equalities appear as

$$\underbrace{\begin{bmatrix} \pi(1) \\ \vdots \\ \pi(\gamma) \\ \vdots \\ \pi(n) \end{bmatrix}}_{\boldsymbol{\pi}} = \underbrace{\begin{bmatrix} n & \cdots & n\binom{n-1}{j-1}[-1]^{j-1} & \cdots & n[-1]^{n-1} \\ & & \vdots & & \vdots \\ 0 & & \binom{n}{\gamma}\binom{n-\gamma}{j-\gamma}[-1]^{j-\gamma} & \cdots & \binom{n}{\gamma}[-1]^{n-\gamma} \\ \vdots & & & & \vdots \\ 0 & \cdots & 0 & & 1 \end{bmatrix}}_{\mathbf{Q}} \underbrace{\begin{bmatrix} m_1 \\ \vdots \\ m_j \\ \vdots \\ m_n \end{bmatrix}}_{\mathbf{m}}. \tag{50}$$

Combining Equations 42 and 50, we then have $\mathbf{p} = \mathbf{C}\boldsymbol{\pi} = \mathbf{CQm}$. This is equivalent to the direct calculation of the choice probabilities in Equation 47, since it can be verified that $\mathbf{CQ} = \mathbf{R}$.

3.3 Beyond moments in the ρ-model

3.3.1 From moments to distributions

Throughout Section 3.3 we shall treat as known a finite number of moments of the type distribution F, appealing to Proposition 2 for justification. We proceed to outline two different strategies for ensuring that this moment information adequately captures the distribution itself. The first strategy relies upon discreteness of the distribution and guarantees a unique characterization of F, while the second relies upon differentiability and guarantees convergence to F in the limit (with respect to n).

3.3.2 Discrete distributions

Assume first that F is a discrete distribution, with ρ taking on values $\langle \rho_1, \rho_2, \ldots, \rho_L \rangle$. The number L of cognitive types is known, though the values themselves may be unknown. We assume, however, that the values must be located on a finite grid of admissible points in $[0, 1]$, which can be as fine as desired.

The realized values of ρ have probabilities $\langle \pi(\rho_1), \pi(\rho_2), \ldots, \pi(\rho_L) \rangle$, strictly positive and summing to one, so that the jth moment of F appears as

$$m_j = \sum_{\ell=1}^{L} \pi(\rho_\ell)\, \rho_\ell^j. \tag{51}$$

Since the first n moments are known, Equation 51 provides a system of n equalities in $2L$ unknowns (namely, the values ρ_ℓ and the associated probabilities $\pi(\rho_\ell)$). This system can be solved for n sufficiently large, but it is not obvious that the solution will be unique.

Assume now that the grid of admissible values of ρ is $\langle 0, 1/N, 2/N, \ldots, 1 \rangle$, with the fineness parameter N large relative to L.[11] In this case F is a discrete distribution fully defined by the probability masses $\langle \pi(\ell/N) \rangle_{\ell=0}^{N}$, of which exactly $L \ll N$ are nonzero. Recovering the distribution then amounts to finding a solution of the system

$$\underbrace{\begin{bmatrix} 1 \\ m_1 \\ \vdots \\ m_j \\ \vdots \\ m_n \end{bmatrix}}_{\overline{\mathbf{m}}} = \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ 0 & 1/N & \cdots & \ell/N & \cdots & 1 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & [1/N]^j & \cdots & [\ell/N]^j & \cdots & 1 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & [1/N]^n & \cdots & [\ell/N]^n & \cdots & 1 \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} \pi(0) \\ \pi(1/N) \\ \vdots \\ \pi(\ell/N) \\ \vdots \\ \pi(1) \end{bmatrix}}_{\overline{\boldsymbol{\pi}}}, \tag{52}$$

[11] We use an evenly spaced grid of admissible values for notational simplicity, but this is not essential for our conclusions.

with all components of the solution vector $\overline{\pi}$ weakly positive and exactly L components strictly positive. Here $\mathbf{V}$ is a Vandermonde matrix with many more columns (grid points) than rows (known moments), implying an underdetermined system.[12] But the number L of grid points actually used could in principle be larger or smaller than n.

A result of Cohen and Yeredor [9, Theorem 1] applies to precisely this situation, stating that Equation 52 has a unique solution whenever $n \geq 2L$. We thus conclude the following.

Proposition 3. *In the ρ-model, if F is a discrete distribution over admissible types, with $n \geq 2L$, then it is uniquely determined by the aggregate choice shares $\langle p(k) \rangle_{k=1}^{n}$.*

That is to say, for practical purposes any discrete distribution for ρ can be fully recovered from aggregate choice data provided the number of alternatives is large compared to the number of cognitive types.

3.3.3 Differentiable distributions

Now assume that the type distribution F possesses a probability density f. In this case we will not be able to fully recover the distribution from the first n moments. Instead, we wish to ensure that the known moments yield a reliable approximation of the true distribution.

Our analysis relies on standard techniques from the "Hausdorff moment problem" for distributions on a closed interval. Adopting a maximum entropy approach, define the nth approximating density f_n as the solution to the optimization problem

$$\max_{f_n} \int_0^1 \left[-\log f_n(\rho) \right] f_n(\rho) \, \mathrm{d}\rho \tag{53}$$

subject to the moment constraints

$$\forall j : 0 \leq j \leq n, \quad \int_0^1 \rho^j f_n(\rho) \, \mathrm{d}\rho = m_j. \tag{54}$$

Mead and Papanicolaou [31, Theorem 2] establish that a solution to this problem exists and is unique.[13] Moreover, for each continuous map $\psi : [0,1] \to \Re$ we have

$$\lim_{n \to \infty} \int_0^1 \psi(\rho) f_n(\rho) \, \mathrm{d}\rho = \int_0^1 \psi(\rho) f(\rho) \mathrm{d}\rho. \tag{55}$$

Write F_n for the distribution function associated with the approximating density f_n.

[12] See, e.g., Macon and Spitzbart [23] for the definition and properties of Vandermonde matrices.

[13] Indeed, the solution takes the form $f_n(\rho) = \exp[-\sum_{j=0}^{n} \lambda_j \rho^j]$, where the quantities $\langle \lambda_j \rangle_{j=0}^{n}$ are the Lagrange multipliers on the constraints in Equation 54.

Observe now that for any menu A and each $k \leq \min\{n, |A|\}$, we have

$$p(k_A; F_n) = p(k_M; F_n) = p(k_M; F) = p(k_A; F). \tag{56}$$

Here the first and third equalities follow from the fact that in the ρ-model an alternative's choice share depends only on its ordinal position on the menu according to the unanimous preferences. Moreover, the shares of the n most preferred alternatives are determined by the first n moments (see Equation 47), which coincide for F and F_n (see Equation 54). This yields the second equality above, and we can summarize our findings as follows.

Proposition 4. *In the ρ-model, if F is differentiable then there exists a sequence $\langle F_n \rangle_{n=1}^{\infty}$ of distributions such that:* (i) *each F_n is defined by $\langle m_j \rangle_{j=1}^{n}$;* (ii) *$F_n$ converges weakly to F; and* (iii) *for each menu A and $k \leq \min\{n, |A|\}$ we have $p(k_A; F_n) = p(k_A; F)$.*

Equation 54 ensures that each approximation F_n is observationally indistinguishable from the true F in the sense that the two distributions generate the same first n moments, and hence the same aggregate choice shares over the observed menu M. Proposition 4 reinforces this conclusion by guaranteeing that the cognitive heterogeneity in the population is accurately reflected in two additional ways: First, as the size of the observed menu increases, the resulting approximations approach the true distribution in the sense of weak convergence. And second, for any menu size n the approximation F_n matches the true F not just over M, but also over the n most preferred alternatives on any other menu A about which we may wish to make predictions.

3.4 Extensions

3.4.1 Indifference

In both of our models of cognitive heterogeneity, strict preferences between the alternatives are needed for the full identification results. Nevertheless, when indifference is allowed we still obtain partial restrictions on the type probabilities in the γ-model and the moments of the type distribution in the ρ-model.

To examine this case, number the alternatives consistently with the ranking (i.e., so that strict preference implies a lower index) and arbitrarily within each indifference class. For each index k, write $\omega_k \leq k$ for the smallest index such that k and ω_k are in the same indifference class, and $\omega^k \geq k$ for the largest such index. We assume that indifferent options are chosen with equal probability if they are jointly optimal among the subset of alternatives considered.

To allow for indifference in the γ-model, we generalize Equations 15–16 to

$$p(k) = \begin{cases} \frac{1}{\omega^k - \omega_k + 1} \left[1 - \sum_{\gamma=0}^{n-\omega^k} \frac{\binom{n-\omega^k}{\gamma}}{\binom{n}{\gamma}} \pi(\gamma) \right] & \text{if } \omega_k = 1, \\ \frac{1}{\omega^k - \omega_k + 1} \left[[\sum_{\gamma=1}^{n-\omega_k+1} \frac{\binom{n-\omega_k+1}{\gamma}}{\binom{n}{\gamma}} \pi(\gamma)] - [\sum_{\gamma=1}^{n-\omega^k} \frac{\binom{n-\omega^k}{\gamma}}{\binom{n}{\gamma}} \pi(\gamma)] \right] & \text{if } \omega_k > 1. \end{cases} \tag{57}$$

Here the denominator of the first factor is the size of the indifference class containing alternative k, and in each case the second factor is the probability that the best perceived option is in this class.

Example 7. [*Indifference in the γ-model*] Let $n = 4$ and suppose that the unanimous preferences over the alternatives are $1 \succ 2 \sim 3 \succ 4$. Using Equation 57, the analog of Equation 42 is then

$$\underbrace{\begin{bmatrix} p(1) \\ p(2) \\ p(3) \\ p(4) \end{bmatrix}}_{\mathbf{p}} = \underbrace{\begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{3}{4} & 1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{8} & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{8} & 0 \\ \frac{1}{4} & 0 & 0 & 0 \end{bmatrix}}_{\mathbf{C}} \underbrace{\begin{bmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \\ 1 - F(3) \end{bmatrix}}_{\pi}. \tag{58}$$

Here the matrix $\mathbf{C}$ is no longer upper anti-triangular and no longer invertible, since its second and third rows are identical. Observe that these are the rows that determine the choice shares of the indifferent alternatives 2 and 3, and any two indifferent options will lead to similar non-invertibility of $\mathbf{C}$.

From Equation 58 we have $\pi(1) = 4p(4)$ and hence $2\pi(2) + \pi(3) = 8[p(2) - p(4)]$. But we cannot separate $\pi(2)$ and $\pi(3)$ in this system, since the strict preference $2 \succ 3$ is needed to disambiguate the cases $\gamma = 2$ and $\gamma = 3$. □

To allow for indifference in the ρ-model, we generalize Equations 6 and 45 to

$$\begin{aligned} p(k) &= \frac{1}{\omega^k - \omega_k + 1} \int_0^1 \left[[1-\rho]^{\omega_k - 1} - [1-\rho]^{\omega^k} \right] \mathrm{d}F \\ &= \frac{1}{\omega^k - \omega_k + 1} \left[[\sum_{j=1}^{\omega_k - 1} [-1]^j \left[\binom{\omega_k - 1}{j} - \binom{\omega^k}{j} \right] m_j] - [\sum_{j=\omega_k}^{\omega^k} [-1]^j \binom{\omega^k}{j} m_j] \right]. \end{aligned} \tag{59}$$

Here once again the denominator of the first factor is the size of the indifference class containing alternative k, while the second factor is the probability that the best perceived option is in this class.

Example 8. [*Indifference in the ρ-model*] As in Example 7, let $n = 4$ and $1 \succ 2 \sim 3 \succ 4$. Using

Equation 59, the analog of Equation 47 is then

$$\underbrace{\begin{bmatrix} p(1) \\ p(2) \\ p(3) \\ p(4) \end{bmatrix}}_{\mathbf{p}} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & -\frac{3}{2} & \frac{1}{2} & 0 \\ 1 & -\frac{3}{2} & \frac{1}{2} & 0 \\ 1 & -3 & 3 & -1 \end{bmatrix}}_{\mathbf{R}} \underbrace{\begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{bmatrix}}_{\mathbf{m}}. \tag{60}$$

Here the matrix $\mathbf{R}$ is no longer lower triangular and no longer invertible, since the rows determining the choice shares of alternatives 2 and 3 are again identical.

From Equation 60 we have $m_1 = p(1)$ and hence $3m_2 - m_3 = 2[p(1) - p(2)]$, but we cannot separate m_2 and m_3 due to the indifference $2 \sim 3$. □

3.4.2 Taste heterogeneity

Our analysis can be extended to allow for taste heterogeneity, provided it is statistically independent of the cognitive heterogeneity. To demonstrate this, order the alternatives on the menu $M = \{1, 2, \ldots, n\}$ arbitrarily and write $\varphi : M \to \{1, 2, \ldots, n\}$ for the mapping that associates each option with its preference rank. We enumerate the possible rankings as $\langle \varphi_h \rangle_{h=1}^{n!}$, and write τ_h for the probability of ranking φ_h. In the γ-model, we can express the choice share of alternative k in the presence of taste heterogeneity as

$$p(k) = \sum_{h=1}^{n!} \tau_h [\mathbf{C}\boldsymbol{\pi}](\varphi_h(k)) = \sum_{h=1}^{n!} \tau_h [\varphi_h(\mathbf{C})\boldsymbol{\pi}](k), \tag{61}$$

where $\varphi_h(\mathbf{C})$ is the matrix constructed by permuting the rows of $\mathbf{C}$ according to the ranking φ_h. The full vector of choice shares is then

$$\mathbf{p} = \sum_{h=1}^{n!} \tau_h \varphi_h(\mathbf{C})\boldsymbol{\pi} = \left[\sum_{h=1}^{n!} \tau_h \varphi_h(\mathbf{C}) \right] \boldsymbol{\pi}, \tag{62}$$

generalizing Equation 42. Thus, provided the taste distribution creates a nonsingular matrix $\sum_{h=1}^{n!} \tau_h \varphi_h(\mathbf{C})$, we can still use the choice shares to compute the probabilities of the n smallest consideration capacities as in Section 3.2.2. Similarly, in the ρ-model we obtain the shares

$$\mathbf{p} = \left[\sum_{h=1}^{n!} \tau_h \varphi_h(\mathbf{R}) \right] \mathbf{m}, \tag{63}$$

generalizing Equation 47 and permitting recovery of the n smallest moments as in Section 3.2.3.

In summary, we conclude that Propositions 1–2 continue to hold under taste heterogeneity provided the distribution of tastes is known (or can be estimated), and provided this distribu-

tion does not make the relevant transition matrix singular.[14]

Example 9. [*Exploded logit*] Let $n = 3$, define $u : M \to \Re$ by $u(k) = \log k$, and suppose that the distribution of tastes is determined by an exploded logit based on u (see, e.g., Luce and Suppes [22]). For instance, the probability assigned to the ranking φ_2 given by $2 \succ 3 \succ 1$ is computed as

$$\begin{aligned}\tau_2 &= \frac{\exp[u(2)]}{\exp[u(1)]+\exp[u(2)]+\exp[u(3)]} \times \frac{\exp[u(3)]}{\exp[u(1)]+\exp[u(3)]} \times \frac{\exp[u(1)]}{\exp[u(1)]} \\ &= \frac{2}{1+2+3} \times \frac{3}{1+3} \times \frac{1}{1} = \frac{1}{3} \times \frac{3}{4} \times 1 = \frac{1}{4}. \end{aligned} \tag{64}$$

In the context of the γ-model we then have

$$\begin{aligned}\sum_{h=1}^{n!} \tau_h \varphi_h(\mathbf{C}) = \frac{1}{3}\begin{bmatrix} \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{2}{3} & 1 \end{bmatrix} + \frac{1}{4}\begin{bmatrix} \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 1 \\ \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix} + \frac{1}{6}\begin{bmatrix} \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 1 \end{bmatrix} \cdots \\ \cdots + \frac{1}{10}\begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 1 \\ \frac{1}{3} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix} + \frac{1}{12}\begin{bmatrix} \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{2}{3} & 1 \\ \frac{1}{3} & 0 & 0 \end{bmatrix} + \frac{1}{15}\begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 1 \\ \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & 0 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{7}{36} & \frac{1}{6} \\ \frac{1}{3} & \frac{16}{45} & \frac{1}{3} \\ \frac{1}{3} & \frac{9}{20} & \frac{1}{2} \end{bmatrix}, \end{aligned} \tag{65}$$

a nonsingular matrix (with determinant $1/270$). □

4 Empirical exercise: Demand for OTC painkillers

4.1 Data

In this section we put our two models of cognitive heterogeneity to use by estimating them with data on sales of over-the-counter (OTC) painkillers. Rather than to carry out a comprehensive study of limited consideration in the retail market for analgesic drugs—a task that is beyond the scope of this paper—our aim is to provide a minimal "worked example" demonstrating that the models are conducive to estimation with a standard dataset.[15]

Our dataset contains sales data from 77 stores in the Chicago area for a period of 48 weeks. Painkillers are sold in various quantities under a generic brand as well as three popular brands: Tylenol, Advil, and Aspirin. For the estimation we include only the eight products that have a

[14] Taste parameters such as risk aversion coefficients or discount factors can be elicited from subjects separately, in a setting (e.g., a field or laboratory experiment) where limited attention is not thought to be relevant, and the resulting values used in the estimation of cognitive heterogeneity. Alternatively, as demonstrated in the empirical exercise in Section 4, taste and attention parameters can be estimated together via maximum likelihood.

[15] Indeed, the dataset employed (graciously supplied by Vishal Singh) is typical of those used for estimation using the techniques of Berry et al. [4].

market share of at least three percent. These are the generic brand in sizes 50 and 100 (tablets per bottle); Tylenol in sizes 25, 50, and 100; Advil in sizes 25 and 50; and Aspirin in size 100.

We observe a total of 3504 store/week combinations, which we refer to as *markets*. For each product and market we have data on sales (denoted by *Sales*), price (*Price*), and any applicable promotion (*Promo*). The dataset also contains demographic information about each store's neighborhood: the percentage non-white population (*NWP*), average education (*Edu*), and average household income (*Inc*). For ease of interpretation, the demographic variables are standardized across the 3504 markets.

4.2 Model specification

Our model of demand allows for both cognitive and taste heterogeneity. We assume that the utility of individual i for product k in market s is given by

$$u_{iks} = [\alpha_0 + \alpha_1 \log Inc_s + \alpha_2 NWP_s + \alpha_3 Edu_s] Price_{ks} \cdots \\ \cdots + [\beta_0 + \beta_1 \log Inc_s + \beta_2 NWP_s + \beta_3 Edu_s] Promo_{ks} + \delta_k + \epsilon_{iks}, \quad (66)$$

where δ_k is a product constant and the error term ϵ_{iks} is extreme value distributed.[16] We collect the utility parameters in $\boldsymbol{\omega}_u = \left\langle \langle \alpha_j, \beta_j \rangle_{j=0}^{3}, \langle \delta_k \rangle_{k=1}^{8} \right\rangle$, the market-$s$ variables in

$$\mathbf{D}_s = \left\langle \langle Sales_{ks}, Price_{ks}, Promo_{ks} \rangle_{k=1}^{8}, Inc_s, NWP_s, Edu_s \right\rangle, \quad (67)$$

and the full dataset in $\mathbf{D} = \langle \mathbf{D}_s \rangle_{s=1}^{3504}$. The constant for Tylenol 25 is normalized to zero.

To estimate the ρ-model we use the Kumaraswamy distribution (see Example 6), which has jth raw moment $m_j = bB(1 + j/a, b)$.[17] For the γ-model we use the Conway-Maxwell-Poisson (CMP) distribution, a two-parameter extension of the Poisson distribution with probability mass function

$$\pi(\gamma) = \frac{\lambda^\gamma}{[\gamma!]^\nu} \left[\sum_{t=0}^{\infty} \frac{\lambda^t}{[t!]^\nu} \right]^{-1}, \quad (68)$$

Poisson parameter $\lambda > 0$, and dispersion parameter $\nu \geq 0$.[18]

The choice probabilities of the eight products included are computed from Equations 62–63.

[16] In this formulation individual taste heterogeneity in each market is captured by the error term, taste heterogeneity across markets is captured by market-level demographics, and unobserved product quality is captured by the product constant. A more flexible specification would include random coefficients, individual-level demographics, and an error term incorporating unobserved product quality (see, e.g., Berry et al. [4]). Adding cognitive heterogeneity to such a model poses significant computational challenges and is beyond the scope of this exercise.

[17] See Footnote 10 for the definition of the beta function B, and Mitnik [32] for discussion of the relative merits of the Kumaraswamy and beta distributions.

[18] See Sellers and Shmueli [38] for the main theoretical properties of the CMP distribution, and Sellers et al. [37] for applications. Setting $\nu = 1$ yields the Poisson distribution, $\nu < 1$ generates over-dispersion, and $\nu > 1$ generates under-dispersion. Note that the latter is not accommodated by the better known negative binomial distribution.

Since we have no data on potential customers who did not make a purchase, we condition the theoretical choice probabilities on the event that at least one option is considered.

4.3 Maximum likelihood estimation

We estimate the ρ-model, the γ-model, and a baseline logit model with the utility in Equation 66. Note that the baseline logit is a limiting case of each of the first two models, in which all options are considered with certainty. The conditional log-likelihood function is

$$\mathcal{L}(\boldsymbol{\omega}; \mathbf{D}) = \sum_{k=1}^{8} \sum_{s=1}^{3504} Sales_{ks} \log p_s(k; \boldsymbol{\omega}, \mathbf{D}_s), \tag{69}$$

where $\boldsymbol{\omega}$ is the parameter vector and $p_s(k; \boldsymbol{\omega}, \mathbf{D}_s)$ is the theoretical choice share of product k in market s. In the (Kumaraswamy) ρ-model we have cognitive parameters $\boldsymbol{\omega}_\rho = \langle a, b \rangle$, full parameter vector $\boldsymbol{\omega} = \langle \boldsymbol{\omega}_u, \boldsymbol{\omega}_\rho \rangle$, and theoretical choice shares

$$p_s(k; \boldsymbol{\omega}, \mathbf{D}_s) = \sum_{h=1}^{8!} \tau_h(\boldsymbol{\omega}_u, \mathbf{D}_s)[\varphi_h(\mathbf{R})\mathbf{m}(\boldsymbol{\omega}_\rho)](k). \tag{70}$$

Similarly, in the (CMP) γ-model we have cognitive parameters $\boldsymbol{\omega}_\gamma = \langle \lambda, \nu \rangle$, full parameter vector $\boldsymbol{\omega} = \langle \boldsymbol{\omega}_u, \boldsymbol{\omega}_\gamma \rangle$, and choice shares

$$p_s(k; \boldsymbol{\omega}, \mathbf{D}_s) = \sum_{h=1}^{8!} \tau_h(\boldsymbol{\omega}_u, \mathbf{D}_s)[\varphi_h(\mathbf{C})\boldsymbol{\pi}(\boldsymbol{\omega}_\gamma)](k). \tag{71}$$

4.4 Results

Maximum likelihood estimates for the ρ-model, the γ-model, and the baseline logit model are reported in Table 1. Observe that the coefficients on the fifteen taste-related variables are similar across the first two models, with a ratio of 0.8–1.2 in all but one case (Advil 50). Moreover, the corresponding coefficients in the baseline model are roughly proportional to those in the γ-model, with a ratio of 0.4–0.6 in all but two cases (*Price* · *NWP* and Advil 50).

The distribution of consideration set size for each model of cognitive heterogeneity is shown in Table 2 and Figure 1 (assuming average demographics).[19]

In the ρ-model, the parameter estimates imply an average ρ of 0.322 and a mean size of 2.70, while in the γ-model the estimates imply a mean size of 2.22. Both assign negligible probability to consideration sets of size more than six, but the ρ-model shows more dispersion. Indeed, the γ-model predicts that no individuals will consider more than three options, while the ρ-model

[19] For the γ-model, the size distribution can be computed from Equation 68 (conditioning on $\gamma \geq 1$). For the ρ-model, the probability of size j is $\int_0^1 \binom{8}{j} \rho^j [1-\rho]^{8-j} [1-[1-\rho]^8]^{-1} ab\rho^{a-1}[1-\rho^a]^{b-1} d\rho$.

	ρ-model		γ-model		baseline logit	
	coef.	s.e.	coef.	s.e.	coef.	s.e.
Price	-0.7885	0.0078	-0.7118	0.0190	-0.4060	0.0062
Price · log *Inc*	0.0295	0.0032	0.0299	0.0034	0.0150	0.0016
Price · *NWP*	-0.0120	0.0025	-0.0101	0.0026	-0.0075	0.0013
Price · *Edu*	0.0229	0.0022	0.0264	0.0027	0.0114	0.0012
Promo	0.4110	0.0161	0.4748	0.0228	0.2114	0.0076
Promo · log *Inc*	0.0546	0.0279	0.0632	0.0252	0.0298	0.0137
Promo · *NWP*	-0.1123	0.0212	-0.0974	0.0205	-0.0531	0.0110
Promo · *Edu*	-0.1271	0.0202	-0.1442	0.0193	-0.0651	0.0099
Tylenol 50	1.6151	0.0192	1.6183	0.0622	0.8161	0.0111
Tylenol 100	2.3757	0.0331	2.0950	0.0584	1.2098	0.0231
Advil 25	-0.7765	0.0116	-0.7814	0.0292	-0.3977	0.0070
Advil 50	0.2007	0.0269	0.0663	0.0236	0.0665	0.0130
Aspirin 100	-0.7419	0.0203	-0.7809	0.0308	-0.4213	0.0085
generic 50	-1.9914	0.0159	-1.8922	0.0549	-1.0433	0.0114
generic 100	-0.4817	0.0218	-0.5448	0.0274	-0.2951	0.0097
log a or log λ	2.7931	0.0113	25.0728	4.0323	—	—
log b or log ν	17.9750	0.1417	3.1774	0.1656	—	—
log-likelihood	-654,662.88		-654,548.33		-654,839.74	

Table 1: Maximum likelihood estimates.

size	1	2	3	4	5	6	7	8
ρ-model	0.180	0.294	0.278	0.166	0.064	0.016	0.002	0.000
γ-model	0.000	0.782	0.218	0.000	0.000	0.000	0.000	0.000

Table 2: Distribution of consideration set size for average demographics.

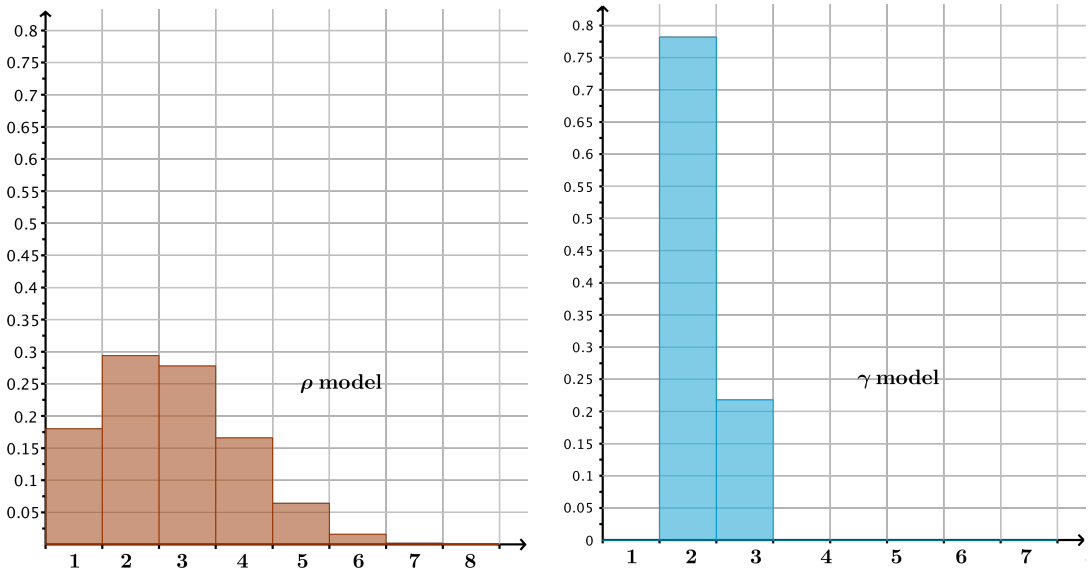

Figure 1: Distribution of consideration set size for average demographics.

model	BIC value
ρ-model	1.3095×10^6
γ-model	1.3093×10^6
baseline logit	1.3099×10^6

Table 3: Bayesian information criterion (BIC) values.

predicts that about one-quarter of the population will consider between four and six.[20] Neither model offers any support for the hypothesis that all alternatives are considered.

Table 3 reports Bayesian information criterion (BIC) values for each model estimated. This model selection criterion favors the γ-model, and ranks the baseline logit below both models of cognitive heterogeneity. In summary, the empirical exercise provides substantial support for limited attention on the part of OTC painkiller purchasers, and the models that allow for this phenomenon agree that the typical consideration set size in the market studied is two or three.

4.5 Related empirical literature

Our empirical exercise contributes to a growing literature on the estimation of consideration sets from demand data. Sovinsky Goeree [40] investigates the impact of marketing on the consideration set, using advertising data to separate utility and attentional components of demand. Van Nierop et al. [34] propose a model of brand choice that accommodates both stated and revealed consideration set data, and apply this framework to an online experiment simulating a variety of merchandising strategies. Abaluck and Adams [1] build a very general econometric framework that exploits asymmetries in the matrix of cross-partial derivatives to identify consideration-set effects. Crawford et al. [10] devise a model-free identification strategy based on reducing the menu of alternatives to a "sufficient set" of those that are certain to be considered. Lu [20] describes an approach to estimating multinomial choice models that employs known upper and lower bounds on the consideration set. Honka et al. [16], among others, model consideration sets as the outcome of a search process, while Gaynor et al. [14] exploit institutional changes to identify consideration sets in hospital choice.[21]

Our exercise is distinct from the above literature in that we use a different identification strategy: We rely on our theoretical results to establish a linear correspondence between the observed choice shares and the unobserved cognitive parameter (namely, the consideration

[20] Note that we have not used any demographic or other covariates in the estimation of ρ and γ. Doing so would allow these cognitive characteristics of the decision maker to be influenced by the environment. For instance, higher incomes could impose higher opportunity costs of attention and thus lead to smaller consideration sets. Studying how the cognitive parameters in our models are determined—both theoretically and empirically—is a promising avenue for future work.

[21] The search literature typically deals with datasets that include information about the composition of a consumer's consideration set, though there are exceptions. For example, in Hastings et al. [15] exposure to a sales force influences the probability that financial products are considered.

probability ρ or capacity γ). This enables us to retrieve the type distribution from demand data via either raw moments (in the ρ-model) or probability masses (in the γ-model).

4.6 Policy analysis

In this section we demonstrate how our empirical results can be used for policy analysis, by using the "deep parameter" estimates for our models to compute consumer welfare under both status-quo and counterfactual scenarios. For instance, we can examine the change in welfare induced by a shift in the distribution of the relevant attention parameter brought about by an advertising campaign (for painkillers generally), by news coverage, or by interventions in the market intended to reduce consumers' search costs.[22]

Generally speaking, any intervention that increases consumers' awareness of the products will shift the distribution of the attention parameter to the right, and for any specified counterfactual distribution the welfare impact can be easily computed. For simplicity we consider only the full-attention counterfactual in which each consumer considers all options on the menu, corresponding to $\rho = 1$ or $\gamma \geq n$ with certainty.

Consider an individual i with utility $u_{iks} = v_{iks} + \epsilon_{iks}$ for product k in market s. McFadden [29] shows that for extreme-value errors the consumer surplus from a subset A of products is given by

$$\frac{1}{\alpha_s} \log \left[\sum_{k \in A} \exp v_{iks} \right] + \text{constant}, \tag{72}$$

where α_s is the marginal utility of income in market s and the constant term reflects the lack of identification of the absolute level of utility. Writing $\mathcal{A}_j$ for the collection of subsets containing exactly j products, the expected consumer surplus in the ρ-model is then

$$\sum_{j=1}^{n} \left[\int_0^1 \frac{\binom{n}{j} \rho^j [1-\rho]^{n-j}}{1-[1-\rho]^n} \mathrm{d}F \right] \times \frac{1}{\binom{n}{j}} \sum_{A \in \mathcal{A}_j} \frac{1}{\alpha_s} \log \left[\sum_{k \in A} \exp v_{iks} \right] + \text{constant}. \tag{73}$$

Similarly, defining the conditional probability masses

$$\pi^+(\gamma) = \frac{1}{1-F(0)} \times \begin{cases} F(\gamma) - F(\gamma - 1) & \text{if } 1 \leq \gamma < n, \\ 1 - F(n-1) & \text{if } \gamma = n; \end{cases} \tag{74}$$

[22] For an example of news coverage plausibly having this effect, see Mele, Christopher, "Picking the Right Over-the-Counter Pain Reliever," *New York Times*, February 6, 2017.

the expected consumer surplus in the γ-model is

$$\sum_{\gamma=1}^{n} \pi^{+}(\gamma) \times \frac{1}{\binom{n}{\gamma}} \sum_{A \in \mathcal{A}_\gamma} \frac{1}{\alpha_s} \log \left[\sum_{k \in A} \exp v_{iks} \right] + \text{constant.} \tag{75}$$

We use Equation 66 to estimate v_{iks} and the price coefficient in each market to estimate the associated α_s. Combining these values with the share of consumers in each market and the estimated cognitive distribution F, we can then calculate consumer surplus for the limited attention case in each of our two models. As noted, our counterfactual scenario is that of full attention, with consumer surplus given by Equation 72 for A equal to the entire menu. This comparison yields an estimated (expected, per individual) welfare loss from limited attention of \$1.65 in the ρ-model and \$1.98 in the γ-model; sizable effects in view of the average price of \$4.27 in our dataset.

5 Concluding comments

This paper contributes to the theoretical literature on boundedly rational decision making by outlining a methodology for inferring the distribution of cognitive characteristics in a population using aggregate choice data. A major advantage of our approach is that it assumes a fixed menu of alternatives, in contrast to much earlier work in this area that assumes knowledge of a single individual's choices from a family of overlapping menus. While both theoretical frameworks yield results that can be brought to bear on data, our view is that the fixed-menu approach is closer to the practice of empirical research on discrete choice and hence lends itself particularly well to testing. We have sought to demonstrate this by means of a circumscribed empirical exercise applying models of consideration set formation to retail choice data.

A second message of the paper is that both the "consideration probability" ρ-model and the "consideration capacity" γ-model are surprisingly tractable within the fixed-menu framework. In both models the aggregate choice shares are linear functions of quantities that are highly informative about the cognitive distribution; namely, low-cardinality choice set probabilities in the γ-model and low-order raw moments in the ρ-model. These systems are recursive (provided all preferences are strict) and easily solved for the quantities in question. Indeed, our theoretical results show that for large menus the cognitive distribution is essentially fully identified, while for smaller menus we can still infer substantial useful information (and typically the full distribution in parameterized settings).

Finally, we mention three possible ways to build on the work reported in this paper. One is to generalize the models of consideration set formation that we have studied; for example, by allowing non-uniform consideration probabilities in the ρ-model, or by relaxing the assumption that all consideration sets with the same cardinality are equally likely to occur in

the γ-model.[23] Another is to bring additional models of bounded rationality—incorporating phenomena such as computational constraints and reference points—into the present framework. And a third is to enrich the econometric specification used in our empirical exercise (see Footnote 16), allowing more precise control of the interaction between cognitive and taste heterogeneity.

References

[1] Jason Abaluck and Abi Adams (2016). Discrete choice models with consideration sets: Identification from asymmetric cross-derivatives. Unpublished.

[2] Jose Apesteguia and Miguel A. Ballester (2013). Choice by sequential procedures. *Games and Economic Behavior* 77:90–99.

[3] Nicholas Baigent and Wulf Gaertner (1996). Never choose the uniquely largest: A characterization. *Economic Theory* 8:239–249.

[4] Steven Berry, James Levinsohn, and Ariel Pakes (1995). Automobile prices in market equilibrium. *Econometrica* 63:841–890.

[5] Richard L. Brady and John Rehbeck (2016). Menu-dependent stochastic feasibility. *Econometrica* 84:1203–1223.

[6] Andrew Caplin and Mark Dean (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review* 105:2183–2203.

[7] Andrew Caplin, Mark Dean, and Daniel Martin (2011). Search and satisficing. *American Economic Review* 101:2899–2922.

[8] Vadim Cherepanov, Timothy Feddersen, and Alvaro Sandroni (2013). Rationalization. *Theoretical Economics* 8:775–800.

[9] Anna Cohen and Arie Yeredor (2011). On the use of sparsity for recovering discrete probability distributions from their moments. *Proceedings of the 2011 IEEE Statistical Signal Processing Workshop*.

[10] Gregory S. Crawford, Rachel Griffith, and Alessandro Iaria (2016). Demand estimation with unobserved choice set heterogeneity. CEPR Discussion Paper 11675.

[11] Geoffroy de Clippel, Kfir Eliaz, and Kareen Rozen (2014). Competing for consumer inattention. *Journal of Political Economy* 122:1203–1234.

[23] On the extension to non-uniform consideration probabilities, see Brady and Rehbeck [5].

[12] Henrique de Oliveira, Tommaso Denti, Maximilian Mihm, and Kemal Ozbek (2016). Rationally inattentive preferences and hidden information costs. *Theoretical Economics*, forthcoming.

[13] Kfir Eliaz and Ran Spiegler (2011). Consideration sets and competitive marketing. *Review of Economic Studies* 78:235–262.

[14] Martin Gaynor, Carol Propper, and Stephan Seiler (2016). Free to choose? Reform, choice, and consideration sets in the English National Health Service. *American Economic Review* 106:3521–3557.

[15] Justine S. Hastings, Ali Hortaçsu, and Chad Syverson (2013). Sales force and competition in financial product markets: The case of Mexico's social security privatization. *Econometrica*, forthcoming.

[16] Elisabeth Honka, Ali Hortaçsu, and Maria Ana Vitorino (2016). Advertising, consumer awareness, and choice: Evidence from the U.S. banking industry. *Rand Journal of Economics*, forthcoming.

[17] Gil Kalai, Ariel Rubinstein, and Ran Spiegler (2002). Rationalizing choice functions by multiple rationales. *Econometrica* 70:2481–2488.

[18] Terri Kneeland (2015). Identifying higher-order rationality. *Econometrica* 83:2065–2079.

[19] P. Kumaraswamy (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology* 46:79–88.

[20] Zhentong Lu (2016). Estimating multinomial choice models with unobserved choice sets. Unpublished.

[21] R. Duncan Luce (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley.

[22] R. Duncan Luce and Patrick C. Suppes (1965). Preferences, utility and subjective probability. In: R. Duncan Luce, Robert R. Bush, and Eugene Galanter, eds., *Handbook of Mathematical Psychology*, Wiley.

[23] Nathaniel Macon and Abraham Spitzbart (1958). Inverses of Vandermonde matrices. *The American Mathematical Monthly* 65:95–100.

[24] Paola Manzini and Marco Mariotti (2007). Sequentially rationalizable choice. *American Economic Review* 97:1824–1839.

[25] Paola Manzini and Marco Mariotti (2014). Stochastic choice and consideration sets. *Econometrica* 82:1153–1176.

[26] Yusufcan Masatlioglu and Daisuke Nakajima (2013). Choice by iterative search. *Theoretical Economics* 8:701–728.

[27] Yusufcan Masatlioglu, Daisuke Nakajima, and Erkut Y. Ozbay (2012). Revealed attention. *American Economic Review* 102:2183–2205.

[28] Daniel L. McFadden (1974). Conditional logit analysis of quantitative choice behavior. In Paul Zarembka, ed., *Frontiers in Econometrics*, Academic Press.

[29] Daniel L. McFadden (1978). Modeling the choice of residential location. In: Anders Karlqvist, Lars Lundqvist, Folke Snickars, and Jorgen W. Weibull, eds., *Spatial Interaction Theory and Planning Models*, North-Holland.

[30] Daniel L. McFadden (2001). Economic choices. *American Economic Review* 91:351–378.

[31] Laurence R. Mead and Nikos Papanicolaou (1984). Maximum entropy in the problem of moments. *Journal of Mathematical Physics* 25:2404–2417.

[32] Pablo A. Mitnik (2013). New properties of the Kumaraswamy distribution. *Communications in Statistics—Theory and Methods* 42:741–755.

[33] Aviv Nevo (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics and Management Strategy* 9:513–548.

[34] Erjen van Nierop, Bart Bronnenberg, Richard Paap, Michel Wedel, and Philip Hans Franses (2010). Retrieving unobserved consideration sets from household panel data. *Journal of Marketing Research* 47:63–74.

[35] Efe A. Ok, Pietro Ortoleva, and Gil Riella (2014). Revealed (p)reference theory. *American Economic Review* 105:299–321.

[36] Yuval Salant and Ariel Rubinstein (2008). (A, f): Choice with frames. *Review of Economic Studies* 75:1287–1296.

[37] Kimberly F. Sellers, Sharad Borle, and Galit Shmueli (2012). The COM-Poisson model for count data: A survey of methods and applications. *Applied Stochastic Models in Business and Industry* 28:104–116.

[38] Kimberly F. Sellers and Galit Shmueli (2010). A flexible regression model for count data. *Annals of Applied Statistics* 4:943–961.

[39] Christopher A. Sims (2003). Implications of rational inattention. *Journal of Monetary Economics* 50:665–690.

[40] Michelle Sovinsky Goeree (2008). Limited information and advertising in the U.S. personal computer industry. *Econometrica* 76:1017–1074.

[41] Kenneth E. Train (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.

[42] Christopher J. Tyson (2008). Cognitive constraints, contraction consistency, and the satisficing criterion. *Journal of Economic Theory* 138:51–70.

[43] Christopher J. Tyson (2013). Behavioral implications of shortlisting procedures. *Social Choice and Welfare* 41:941–963.