

Inferring Cognitive Heterogeneity from Aggregate Choices*

Valentino Dardanoni

Università degli Studi di Palermo

Paola Manzini

University of Sussex and IZA

Marco Mariotti

Queen Mary University of London

Christopher J. Tyson

Queen Mary University of London

March 23, 2019

Abstract

Theoretical work on bounded rationality typically assumes a rich dataset of choices from many overlapping menus, limiting its practical applicability. In contrast, we study the problem of identifying the distribution of cognitive characteristics in a population of agents from a minimal dataset consisting of aggregate choice shares from a single menu. Under homogeneous preferences, we show that both “consideration capacity” and “consideration probability” distributions can be recovered effectively when the menu is sufficiently large. This remains generically true when tastes are heterogeneous but their distribution is known. When the preference distribution is unknown, we demonstrate that joint choice share data from as few as three “occasions” are generically sufficient for full identification of the cognitive capacity distribution and provide substantial information about tastes.

Keywords: attention, bounded rationality, consideration set, stochastic choice.

J.E.L. codes: D01, D12, D91.

*For valuable comments and suggestions that have improved this paper, we would like to thank the co-editor and three anonymous referees; Jason Abaluck, Abi Adams, Levon Barseghyan, Aluma Dembo, Reyer Gerlagh, Alessandro Iaria, Rod McCrorie, Yusufcan Masatlioglu, Irina Merkurieva, Francesca Molinari, and Ted O’Donoghue; seminar audiences at Edinburgh, Leuven, Naples, Oxford, St Andrews, Sussex, Tilburg, ULB (ECARES), Vienna, and Warwick (CRETA); and audiences at the Barcelona GSE Summer Forum (2016), D-TEA (2016), ICEEE (2017), the workshop on “Identification and Inference in Limited Attention Models” at Cornell University (2018), the Petralia Summer Workshop (2018), and the ASSA meetings (2019). Manzini and Mariotti acknowledge financial support from the ESRC through grant ES/J012513/1, and Dardanoni acknowledges support from MIUR through grant PRIN 2015.

1 Introduction

1.1 Motivation

Classical revealed preference analysis has yielded a fine-grained understanding of the relationship between unobserved tastes and observed choices, and of how to infer the former from the latter. More recently, theoretical work on bounded rationality has extended this methodology to incorporate a range of cognitive factors that may affect decision making.¹ One drawback of such theories is that they typically presume access to a very rich dataset—comprising a single individual’s choices from a large number of different overlapping menus—that can be used to identify the latent components of the cognitive model of interest. For instance, Aguiar et al. [2], Cattaneo et al. [13], and Masatlioglu et al. [32] require data for all possible menus drawn from a universal set of alternatives; Manzini and Mariotti [30] impose a stringent “richness” assumption on their dataset; and Caplin and Dean [10] postulate the observability of state-dependent stochastic choice data.²

Identification results developed using such assumptions on the choice domain are often formally elegant, and can be particularly useful for designing and interpreting experiments (as in Aguiar et al. [2] and Caplin and Dean [10]). They are less obviously relevant to field data, however, especially when the type of decision arises rarely (e.g., choice of hospital provider for elective surgery) or the menu is slow to change (e.g., choice of daily newspaper). Indeed, in settings with such features many characterization results from the literature on boundedly rational choice may appear implausibly data-hungry. In practice there may be insufficient menu variation to infer the model components of interest, and for this reason it is desirable to devise approaches to identification that create a direct link between theory and what may be feasible empirically.

¹This literature examines cognitive factors such as computational constraints, norms and heuristics, reference points and other framing effects, and various conceptions of attention. Contributions include those of Apesteguia and Ballester [4], Baigent and Gaertner [6], Caplin et al. [11], Cherepanov et al. [14], de Oliveira et al. [18], Echenique et al. [19], Manzini and Mariotti [29], Masatlioglu and Nakajima [31], Ok et al. [35], Salant and Rubinstein [38], and Tyson [44, 45], among numerous others.

²Even stronger assumptions about data availability are commonplace in the theory of choice under uncertainty, where the decision maker is typically imagined to express preferences over a highly structured mathematical space specifically designed to facilitate identification.

In this paper we focus on models of limited attention, where agents consider only a subset of the available alternatives, known as the “consideration set.”³ To address the data-voracity issue noted above, we propose a novel framework that postulates a minimal dataset comprising (in its basic version) *a single, fixed menu* from which we observe only the *aggregate choice shares* of a population of decision makers.⁴ Members of the population may (or may not) differ in their preferences over the alternatives, and they may also differ in cognitive characteristics that affect the allocation of attention. The latter “cognitive heterogeneity” is taken to be unobserved, and our principal goal is to infer the distribution of these characteristics from the aggregate choice shares.

We stress that this paper is concerned with the extent to which the distribution of cognitive characteristics is identified by a given model of bounded rationality *per se*; that is, prior to any ancillary econometric specification that may include covariates for the individuals or the alternatives. In this respect our primitives and objectives both remain typical of those in conventional abstract choice theory. Indeed, this is one way that our contribution can be distinguished from other recent work (see, e.g., Abaluck and Adams [1] and Barseghyan et al. [7, 8]) in which identification is facilitated by the presence of observable covariates.

1.2 Cognitive models

In our general framework, each agent has a *cognitive type* parameter $\theta \in \Theta \subset \mathfrak{R}$ that is distributed in the population according to a cumulative distribution function F . Given preferences over the menu, an individual of type θ will choose alternative x with probability $p_\theta(x)$, and the corresponding aggregate choice share will be $p(x) = \int_\Theta p_\theta(x) dF$. When the cognitive type is used to capture some form of bounded rationality, the indi-

³This usage follows the marketing literature; see, e.g., Roberts and Lattin [37] and Shocker et al. [39]. While we view the consideration set as a manifestation of bounded rationality, 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 [10] and Sims [41]).

⁴Alternatively, the framework could model a single individual choosing repeatedly from the same menu in different attentional states, where the variation may arise, for example, from a merchandising strategy of the retailer designed to manipulate customers’ consideration sets. In Section 4.2 we extend this framework to allow for richer “multi-occasion” choice data, but only after the informational value of our basic dataset has been completely exhausted.

vidual choice distribution will not generally assign all probability to the best available option, and neither will the aggregate distribution even when the population has homogeneous tastes. Indeed, the very fact that suboptimal alternatives will sometimes be chosen is what will enable us to infer features of the cognitive distribution F from the observed aggregate shares.⁵

As mentioned above, we study bounded rationality in the form of limited attention, where the cognitive parameter θ influences the formation of the decision maker’s consideration set. In the “consideration capacity” model, the parameter $\gamma \in \{0, 1, 2, \dots\}$ controls the maximum cardinality of the consideration set and is interpreted as a limit on the number of alternatives that the agent can actively investigate at any one time. We also examine in detail an important special case, the “consideration probability” model, in which the parameter $\rho \in [0, 1]$ controls the likelihood that each option is considered and is interpreted as the decision maker’s general awareness of the choice environment. We hypothesize that preferences are maximized over the consideration set, and full rationality can be restored by letting $\gamma \rightarrow \infty$ or $\rho = 1$, as appropriate.⁶

1.3 Preview of results

We begin by assuming that the population has homogeneous preferences. Under this assumption, we find that our cognitive model is fully identified by a small number of observed choice shares for several natural parameterizations of F . More specifically, if the consideration capacity γ has a Poisson or Pascal distribution, or if the consideration probability ρ has a uniform or Beta distribution, then between one and three aggregate choice shares are needed to recover all of the parameters of the cognitive distribution (see Examples 1–4). We then proceed to show that even in the absence of a parametric

⁵Note that our framework has similarities to mixed models in the discrete choice literature, where θ would be a taste parameter such as the agent’s unobserved marginal utility of some observed characteristic. (See Train [43] and McFadden [33].) However, since we shall use θ to control cognition instead of tastes, our setting calls for different functional form assumptions. In particular, p_θ will not have a logit specification (see Luce [27]), as would typically be assumed in relation to tastes.

⁶Variants of the consideration capacity model are employed by Barseghyan et al. [7] to study discrete choice with heterogeneous consideration sets, and by de Clippel et al. [17] to study price competition in a setting where consumers exhibit limited attention. A version of the consideration probability model appears in Manzini and Mariotti [30].

specification, the cognitive distribution can for practical purposes be fully recovered provided the menu of alternatives is sufficiently large. In the context of the consideration capacity model, the choice shares identify the probabilities of all capacities less than the cardinality n of the menu (see Proposition 2). Similarly, in the context of the consideration probability model the choice shares identify the first n raw moments of F (see Proposition 3), which—using maximum entropy methods and results from sparsity theory—can be exploited to reconstruct or to closely approximate the cognitive distribution itself (see Propositions 4–5). In each context, identification follows from the system of equations that define the choice shares being recursive and linear in the relevant quantities (namely, the capacity probabilities or the raw moments), so that closed-form expressions for these quantities can be obtained by inverting a triangular or anti-triangular matrix.

Turning to the case of heterogeneous preferences, we first note that our identification results continue to hold generically if the taste distribution is known (see Propositions 6–7). For heterogeneous and unknown tastes, we extend our dataset to include the joint distribution of choices by the same population of agents on at least three distinct “occasions”. Making use of an algebraic result on the uniqueness of tensor decompositions, we show that joint choice share data of this sort are generically sufficient for full identification of the cognitive capacity distribution, and also provide substantial information about the distribution of preferences (see Proposition 8).

1.4 Related empirical literature

While remaining entirely theoretical, this paper contributes to a growing literature on estimating consideration-set models from consumer demand or other choice data, reviewed briefly in this section.

Abaluck and Adams [1] construct a general econometric framework in which product characteristics are observable, and exploit asymmetries in cross-characteristic choice probability responses to identify consideration sets. Aguiar et al. [2] test random consideration models at the population level in an online experiment, finding support for a specification with heterogeneous preferences and logit attention. In the context

of choice under risk, Barseghyan et al. [7] obtain partial identification of preferences using minimal assumptions about the process of consideration-set formation, while Barseghyan et al. [8] obtain point identification of both preferences and attention in a discrete choice model. Cattaneo et al. [13] postulate “monotonic attention,” a restriction on how stochastic consideration sets change across menus, and use this assumption to derive testable restrictions on choice probabilities. Crawford et al. [16] 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. Gaynor et al. [20] exploit institutional changes to identify consideration sets in hospital choice, while Honka et al. [23] (among others) treat consideration sets as the outcome of a search process.⁷ Lu [26] describes an approach to estimating multinomial choice models that employs known upper and lower bounds on the consideration set. Sovinsky Goeree [42] studies the impact of marketing on the consideration set, using advertising data to separate the utility and attentional components of demand. And Van Nierop et al. [46] propose a model of brand choice accommodating both stated and revealed consideration-set data, which they apply to an experiment on merchandising strategies.

1.5 Outline

The remainder of the paper is structured as follows. Section 2 describes our framework and sets out both the consideration capacity model and the special case of the consideration probability model. Section 3 pursues cognitive inference under the simplifying assumption of homogeneous tastes, Section 4 extends the analysis to allow for taste heterogeneity, and Section 5 concludes.

⁷The search literature typically deals with datasets that include information about the composition of a consumer’s consideration set, although there are exceptions. For example, in Hastings et al. [22] exposure to a sales force influences the probability that financial products are considered.

2 Cognitive heterogeneity and consideration-set models

2.1 General framework

Let X denote the (finite) universal set of alternatives. A *menu* is any nonempty $A \subseteq X$, with which is associated a default outcome $d_A \notin A$. When presented 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 product, and the default is not to buy.
- (ii) The menu contains banks offering fixed deposits, and the default is to hold cash.
- (iii) The menu contains risky lotteries, and the default is a risk-free payment.

When deriving our main theoretical results (in Sections 2–3), we shall assume that all agents share the same linear order preferences \succsim over X . This assumption (relaxed in Section 4) can be interpreted as using the average utilities of the alternatives in the population, ignoring individual variation. In this sense our approach is complementary 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. Note 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) provided all agents are approximately risk neutral over the relevant stakes.

When imposing homogeneous tastes, we number the alternatives so that a higher position in the preference order implies a lower index. We thus write k_A for the k th best option on A , and the full menu appears as $A = \{1_A, 2_A, \dots, n_A\}$, where $n_A = |A|$.

We introduce cognitive heterogeneity by assigning each agent a cognitive type $\theta \in \Theta \subset \mathfrak{R}$, drawn independently across agents from the distribution F . We write $p_\theta(k_A)$ for the probability that type θ chooses alternative k_A , and $p(k_A) = \int_\Theta p_\theta(k_A) dF$ for the overall share in the population. Similarly, we write $p_\theta(d_A)$ for the probability that type θ accepts the default, and $p(d_A) = \int_\Theta p_\theta(d_A) dF$ for the population share. For each

$\theta \in \Theta$ we have $[\sum_{k=1}^{n_A} p_\theta(k_A)] + p_\theta(d_A) = 1$, and likewise $[\sum_{k=1}^{n_A} p(k_A)] + p(d_A) = 1$ in aggregate. If wishing 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 a large population choosing from a fixed menu M with $|M| = n_M = n \geq 2$. The analyst observes the aggregate choice shares, but knows neither the common preference order nor the distribution of cognitive types. 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 distribution F from $\langle p(1), p(2), \dots, p(n), p(d) \rangle$, and to use this information to predict aggregate choices from menus other than M .

We proceed now to specialize this framework to a more concrete model in which the cognitive heterogeneity relates to limited attention. Each agent will consider (i.e., pay attention to) a subset of the alternatives, and among those considered will choose the best option according to the common preference order. If the preference-maximizing alternative is not in the consideration set, this will result in a sub-optimal decision.

2.2 The consideration capacity model

Let $\gamma \in \{0, 1, 2, \dots\} = \Theta$ denote a limit on the cardinality of the agent's consideration set; that is, the *consideration capacity*. When $1 \leq \gamma < n$ we assume that the agent is equally likely to consider each $\Gamma \subset M$ with $|\Gamma| = \gamma$, and when $\gamma \geq n$ we know with certainty that the entire menu M will be considered. In the former case there are $\binom{n}{\gamma}$ candidate sets, of which $\binom{n-k}{\gamma-1}$ contain alternative k and do not contain any superior alternative $\ell < k$. For $1 \leq \gamma < n$, 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 full menu is considered we know that alternative 1 will be chosen regardless of the value of $\gamma \geq n$.⁸

⁸We have assumed that the common preference relation \succsim is a linear order; i.e., that no two distinct alternatives are indifferent. If we allow for indifference then, defining $\omega_k(R) = |\{j : jRk\}|$, for $1 \leq \gamma < n$ the probability of option k being chosen is $[(\binom{\omega_k(\succsim)}{\gamma} - \binom{\omega_k(\succ)}{\gamma})][\binom{n}{\gamma}[\omega_k(\succsim) - \omega_k(\succ)]]^{-1}$ (with Equations 1, 6, and 14 below modified accordingly). While this generalization causes no significant difficulty for the

The type-conditional choice frequencies can now be expressed as

$$p_\gamma(k) = \begin{cases} \binom{n-k}{n-1} & \text{if } \gamma \geq n, \\ \binom{n-k}{\gamma-1} / \binom{n}{\gamma} & \text{if } 1 \leq \gamma < n, \\ 0 & \text{if } \gamma = 0; \end{cases} \quad (1)$$

$$p_\gamma(d) = \begin{cases} 0 & \text{if } \gamma > 0, \\ 1 & \text{if } \gamma = 0. \end{cases} \quad (2)$$

Defining the probability masses

$$\pi(0) = F(0), \quad (3)$$

$$\forall \gamma \in \{1, 2, \dots, n-1\}, \quad \pi(\gamma) = F(\gamma) - F(\gamma-1), \quad (4)$$

$$\pi(n) = 1 - F(n-1); \quad (5)$$

the corresponding aggregate shares are then

$$p(k) = \sum_{\gamma=1}^{n-k+1} p_\gamma(k) \pi(\gamma) = \sum_{\gamma=1}^{n-k+1} \frac{\binom{n-k}{\gamma-1}}{\binom{n}{\gamma}} \pi(\gamma), \quad (6)$$

$$p(d) = \pi(0). \quad (7)$$

Observe that for $1 \leq k < n$ we can use Equation 6 to compute

$$p(k) - p(k+1) = \frac{\pi(n-k+1)}{\binom{n}{n-k+1}} + \sum_{\gamma=2}^{n-k} \frac{\binom{n-k-1}{\gamma-2}}{\binom{n}{\gamma}} \pi(\gamma). \quad (8)$$

This relation shows that, when we move one ordinal step up the preference scale, the aggregate choice share increases for two reasons: Firstly, the k th best alternative can be chosen when $\gamma = n - k + 1$, unlike the next best option. And secondly, for values of γ smaller than this the better alternative is chosen more frequently, since there are more

derivation of choice shares, we shall nevertheless maintain the linear ordering assumption so as to avoid our main objective of *cognitive* identification being hampered by a feature of *preferences* alone. We also view the prohibition on indifference as relatively innocuous in the present, finite-menu setting.

ways of populating the rest of the consideration set with inferior options.

Note also that setting $k = n$ in Equation 6 yields $p(n) = \frac{\pi(1)}{n}$ and hence

$$\pi(1) = np(n). \quad (9)$$

Similarly, setting $k = n - 1$ in Equation 8 yields $p(n - 1) - p(n) = \frac{2\pi(2)}{n[n-1]}$ and hence

$$\pi(2) = \frac{n[n-1]}{2} [p(n-1) - p(n)]. \quad (10)$$

Equations 9–10 prefigure the recursive method employed in Section 3 to identify the cognitive type distribution, in which the probabilities $\pi(1), \dots, \pi(n-1)$ are deduced sequentially, with one additional choice share used at each step.

Finally, using Equation 9, we can write Equation 8 in terms of probability ratios as

$$\frac{p(k) - p(k+1)}{p(n)} = \frac{n}{\binom{n}{n-k+1}} \frac{\pi(n-k+1)}{\pi(1)} + n \sum_{\gamma=2}^{n-k} \frac{\binom{n-k-1}{\gamma-2}}{\binom{n}{\gamma}} \frac{\pi(\gamma)}{\pi(1)}. \quad (11)$$

For instance, when $k = n - 1$ we find that the probability mass ratio

$$\frac{\pi(2)}{\pi(1)} = \frac{n-1}{2} \left[\frac{p(n-1)}{p(n)} - 1 \right] \quad (12)$$

between the two smallest (nonzero) values of the consideration capacity depends only on the aggregate choice share ratio between the two worst alternatives on the menu.

2.3 A special case: The consideration probability model

One special case of the consideration capacity model is a version of the *consideration probability* model studied by Manzini and Mariotti [30]. To see this, denote by $\rho \in [0, 1]$ the probability that the agent considers each alternative on the menu, with consideration independent across agents and alternatives. Since the same consideration probability applies independently to each alternative, all subsets of the menu of a given size are equally likely to be the consideration set. Moreover, the probability of a consideration set of size $\gamma \leq n$ is

$$\pi(\gamma) = \binom{n}{\gamma} \int_0^1 \rho^\gamma [1 - \rho]^{n-\gamma} dF, \quad (13)$$

and clearly $\pi(\gamma) = 0$ for $\gamma > n$. Adapting Equations 6–7 to this special case, we obtain the aggregate choice shares

$$p(k) = \sum_{\gamma=1}^{n-k+1} \binom{n-k}{\gamma-1} \int_0^1 \rho^\gamma [1 - \rho]^{n-\gamma} dF = \int_0^1 \rho [1 - \rho]^{k-1} dF, \quad (14)$$

$$p(d) = \int_0^1 [1 - \rho]^n dF; \quad (15)$$

for the consideration probability model. As in the general case, alternative k 's choice share is the probability that this option and nothing better is considered, and the share of the default outcome is the probability that nothing at all is considered.

3 Inference from aggregate choices

3.1 Preference identification

In the context of our limited attention model, the agents' common preferences over the alternatives are fully revealed by the observed choice shares under weak conditions. Indeed, the following result is a simple consequence of Equation 8.

Proposition 1. *For the consideration capacity model, with $1 \leq k < n$:*

- (i) $p(k) \geq p(k+1)$.
- (ii) *If $\sum_{\gamma=2}^{n-k+1} \pi(\gamma) > 0$ then $p(k) > p(k+1)$.*
- (iii) *If $\pi(2) > 0$ then $p(1) > p(2) > \dots > p(n)$.*

Using Equation 13, we can also specialize (iii) to the consideration probability model.

Corollary 1. *For the consideration probability model, if the support of F intersects $(0, 1)$ then $p(1) > p(2) > \dots > p(n)$.*

We conclude that, under the homogeneous tastes assumption, the preferences are for practical purposes fully revealed by aggregate choice data, and efforts can be focused squarely on the cognitive identification problem. For the remainder of Section 3 we assume tacitly that $\pi(2) > 0$, ensuring that the choice shares $p(1) > p(2) > \dots > p(n)$ faithfully reflect the underlying preference order.

3.2 Cognitive identification: Parametric analysis

To examine the cognitive inference problem in its most concrete manifestation, we first consider several natural functional forms for the type distribution. Our aim here is to show that in such cases the cognitive parameters can be revealed in a straightforward fashion from a small number of appropriately selected choice-share observations. As well as increasing our familiarity with the limited-attention model under investigation, the examples below will serve to highlight non-obvious ways that aggregate choices can convey information about the cognitive type distribution.

For both the consideration capacity model and the special case of the consideration probability model, we consider simple one- and two-parameter distributions for the cognitive type.

Example 1. [*Poisson* γ] For $\mu > 0$, let the consideration capacity γ have the Poisson distribution $\pi(\gamma) = \frac{\mu^\gamma}{\gamma!} e^{-\mu}$ for $0 \leq \gamma < n$. In this case Equation 7 yields default share $p(d) = \pi(0) = e^{-\mu}$, and thus $\mu = -\log p(d)$. Alternatively, Equation 12 yields

$$\frac{n-1}{2} \left[\frac{p(n-1)}{p(n)} - 1 \right] = \frac{\pi(2)}{\pi(1)} = \frac{\mu}{2}, \quad (16)$$

and so $\mu = [n-1] \left[\frac{p(n-1)}{p(n)} - 1 \right]$. \square

Example 2. [*Pascal* γ] For $r \in \{1, 2, 3, \dots\}$ and $q \in (0, 1)$, let the consideration capacity γ have the Pascal (or “negative binomial”) distribution $\pi(\gamma) = \binom{\gamma+r-1}{\gamma} [1-q]^r q^\gamma$ for $0 \leq \gamma < n$. Equation 12 then yields

$$\frac{n-1}{2} \left[\frac{p(n-1)}{p(n)} - 1 \right] = \frac{\pi(2)}{\pi(1)} = \frac{q[r+1]}{2}. \quad (17)$$

We have also

$$\frac{np(n)}{p(d)} = \frac{\pi(1)}{\pi(0)} = qr, \quad (18)$$

and Equations 17–18 can be solved simultaneously for the parameters

$$q = [n - 1] \left[\frac{p(n-1)}{p(n)} - 1 \right] - \frac{np(n)}{p(d)}, \quad (19)$$

$$r = \frac{np(n)^2}{p(d)[n-1][p(n-1) - p(n)] - np(n)^2}. \quad \square \quad (20)$$

Example 3. [*uniform* ρ] For $\rho_{\min} \in [0, 1)$, let the consideration probability ρ be distributed uniformly on $[\rho_{\min}, 1]$. Since $F(\rho) = \frac{\rho - \rho_{\min}}{1 - \rho_{\min}}$, Equation 14 becomes

$$p(k) = \frac{1}{1 - \rho_{\min}} \int_{\rho_{\min}}^1 \rho [1 - \rho]^{k-1} d\rho. \quad (21)$$

The first choice share is then $p(1) = \frac{1 + \rho_{\min}}{2}$, yielding the parameter $\rho_{\min} = 2p(1) - 1$. \square

Example 4. [*Beta* ρ] For $a, b > 0$, let the consideration probability have the Beta distribution $F(\rho) = \frac{1}{B(a, b)} \int_0^\rho t^{a-1} [1 - t]^{b-1} dt$ (where B is the Beta function). Here Equation 14 appears as

$$p(k) = \frac{1}{B(a, b)} \int_0^1 \rho^a [1 - \rho]^{b+k-2} d\rho = \frac{B(a+1, b+k-1)}{B(a, b)}. \quad (22)$$

The first two choice shares are

$$p(1) = \frac{B(a+1, b)}{B(a, b)} = \frac{a}{a+b}, \quad (23)$$

$$p(2) = \frac{B(a+1, b+1)}{B(a, b)} = \frac{ab}{[a+b][a+b+1]}; \quad (24)$$

and we can solve for the parameters

$$a = \frac{p(1)p(2)}{p(1)[1 - p(1)] - p(2)}, \quad (25)$$

$$b = \frac{[1 - p(1)]p(2)}{p(1)[1 - p(1)] - p(2)}. \quad \square \quad (26)$$

Observe that for parameterizations of the consideration capacity γ we have used the choice shares $p(n)$ and $p(n - 1)$, corresponding to the least attractive alternatives, to elicit information about the cognitive type distribution. In contrast, for parameterizations of the consideration probability ρ we have used $p(1)$ and $p(2)$, corresponding to the most attractive options. This mirrors our elicitation procedure below in the non-parametric setting, where each mass $\pi(\gamma)$ is seen to depend on the choice shares of a group of sufficiently unattractive alternatives (cf., Equation 28), and each moment of the ρ -distribution is seen to depend on the shares of a sufficiently attractive group (cf., Equation 32).

3.3 Cognitive identification: Nonparametric analysis

3.3.1 The nonparametric inference problem

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

In view of these features of the inference problem, we can decode the information about the cognitive capacity distribution encoded in the choice share data by inverting a triangular $n \times n$ matrix. This will yield the probability of each capacity value strictly less than n , and adding one more option to the menu will give us knowledge of one additional probability mass. In the consideration-probability setting, we can then invert a second triangular $n \times n$ matrix to deduce the first n raw moments of F from the capacity probabilities. Finally, well-established tools (specifically, sparse matrix theory

and maximum entropy methods) will permit us to approximate F from its moments with increasing precision as the size of the menu grows (see Section 3.4).

3.3.2 Recovering n probability masses

Absent parametric assumptions, the aggregate choice shares are given by Equation 6. These relations can be written together in matrix form as

$$\underbrace{\begin{bmatrix} p(1) \\ \vdots \\ p(k) \\ \vdots \\ p(n) \end{bmatrix}}_{\mathbf{p}} = \underbrace{\begin{bmatrix} \frac{1}{n} & \cdots & \frac{\gamma}{n} & \cdots & 1 \\ \vdots & & \vdots & & \vdots \\ \frac{1}{n} & \cdots & \frac{\binom{n-k}{\gamma-1}}{\binom{n}{\gamma}} & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ \frac{1}{n} & \cdots & 0 & \cdots & 0 \end{bmatrix}}_{\mathbf{C}} \underbrace{\begin{bmatrix} \pi(1) \\ \vdots \\ \pi(\gamma) \\ \vdots \\ \pi(n) \end{bmatrix}}_{\boldsymbol{\pi}}. \quad (27)$$

The upper anti-triangular and left-stochastic matrix \mathbf{C} has a lower anti-triangular inverse, allowing us to write $\boldsymbol{\pi} = \mathbf{C}^{-1}\mathbf{p}$.⁹ Indeed, we can calculate the components of $\boldsymbol{\pi}$ explicitly as

$$\pi(\gamma) = \binom{n}{\gamma} \sum_{k=n-\gamma+1}^n [-1]^{[\gamma-1]-[n-k]} \binom{\gamma-1}{n-k} p(k), \quad (28)$$

and of course $\pi(0) = p(d) = 1 - \sum_{k=1}^n p(k)$. Observe that since $\pi(n) = 1 - F(n-1)$, it is in fact the probabilities of the capacities $\gamma = 0, 1, \dots, n-1$ that are revealed; and $\gamma = n$ cannot be disambiguated from higher values. Indeed, all capacities greater than or equal to the number of alternatives will always be behaviorally indistinguishable. We summarize our conclusions as follows.

Proposition 2. *In the consideration capacity model, the probability masses $\boldsymbol{\pi}$ are uniquely determined by the aggregate choice shares \mathbf{p} .*

⁹A matrix is left (resp., right) stochastic if all entries are nonnegative and all columns (resp., all rows) sum to one.

3.3.3 Consideration probability: Recovering n moments

Returning to the special case of the consideration probability model, let us write the j th raw moment of the type distribution as $m_j = \int_0^1 \rho^j dF$. The binomial in Equation 13 can then be expanded to yield

$$\pi(\gamma) = \binom{n}{\gamma} \int_0^1 \rho^\gamma \left[\sum_{i=0}^{n-\gamma} \binom{n-\gamma}{i} [-\rho]^i \right] dF = \binom{n}{\gamma} \sum_{j=\gamma}^n \binom{n-\gamma}{j-\gamma} [-1]^{j-\gamma} m_j. \quad (29)$$

In matrix form, these relations appear as

$$\underbrace{\begin{bmatrix} \pi(1) \\ \vdots \\ \pi(\gamma) \\ \vdots \\ \pi(n) \end{bmatrix}}_{\boldsymbol{\pi}} = \underbrace{\begin{bmatrix} n & \cdots & n \binom{n-1}{j-1} [-1]^{j-1} & \cdots & n [-1]^{n-1} \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & \binom{n}{\gamma} \binom{n-\gamma}{j-\gamma} [-1]^{j-\gamma} & \cdots & \binom{n}{\gamma} [-1]^{n-\gamma} \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}}_{\mathbf{Q}} \underbrace{\begin{bmatrix} m_1 \\ \vdots \\ m_j \\ \vdots \\ m_n \end{bmatrix}}_{\mathbf{m}}. \quad (30)$$

The upper triangular matrix \mathbf{Q} has an upper triangular inverse, so we have

$$\mathbf{m} = \mathbf{Q}^{-1} \boldsymbol{\pi} = \mathbf{Q}^{-1} [\mathbf{C}^{-1} \mathbf{p}] = [\mathbf{CQ}]^{-1} \mathbf{p}. \quad (31)$$

Performing this calculation, the raw moments are given explicitly by

$$m_j = \sum_{k=1}^j [-1]^{k-1} \binom{j-1}{k-1} p(k). \quad (32)$$

We summarize our conclusions for the special case as follows.

Proposition 3. *In the consideration probability model, the raw moments \mathbf{m} are uniquely determined by the aggregate choice shares \mathbf{p} .*

3.4 Consideration probability: Beyond moments

3.4.1 From moments to distributions

Continuing to focus on the consideration probability model, we shall throughout Section 3.4 treat as known a finite number of raw moments of the cognitive type distribution F , appealing to Proposition 3 for justification. We proceed to outline two different strategies for ensuring that this moment information adequately captures F itself. The first strategy will rely on discreteness of the type distribution and ensure a unique characterization of F , while the second will rely on the existence of a density and guarantee convergence to F as $n \rightarrow \infty$.

3.4.2 Discrete type distributions

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

The realized values of ρ have probabilities $\langle \xi(\rho_1), \xi(\rho_2), \dots, \xi(\rho_L) \rangle$, each strictly positive and together summing to one, so that the j th raw moment of F appears as

$$m_j = \sum_{\ell=1}^L \xi(\rho_\ell) \rho_\ell^j. \quad (33)$$

Treating the first n moments as known, Equation 33 supplies a system of n equalities in $2L$ unknowns; namely, the values ρ_ℓ and their associated probabilities $\xi(\rho_\ell)$. This system can be solved for n sufficiently large, but it is not obvious that the solution will be unique.

Assume that the grid of admissible values for ρ is $\langle 0, \frac{1}{N}, \frac{2}{N}, \dots, 1 \rangle$, with the fineness parameter N large relative to L .¹⁰ Then F is a discrete distribution defined entirely by the probability masses $\langle \xi(\frac{\ell}{N}) \rangle_{\ell=0}^N$, of which exactly $L \ll N$ are nonzero. Recovering the distribution thus amounts to finding a solution ξ of the system

¹⁰For notational simplicity we use an evenly spaced grid of admissible values, but this is not essential for our conclusions.

$$\begin{array}{c} \left[\begin{array}{c} 1 \\ m_1 \\ \vdots \\ m_j \\ \vdots \\ m_n \end{array} \right] \\ \underbrace{\hspace{1.5cm}}_{\langle 1, \mathbf{m} \rangle} \end{array} = \begin{array}{c} \left[\begin{array}{cccccc} 1 & 1 & \cdots & 1 & \cdots & 1 \\ 0 & \frac{1}{N} & \cdots & \frac{\ell}{N} & \cdots & 1 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & [\frac{1}{N}]^j & \cdots & [\frac{\ell}{N}]^j & \cdots & 1 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & [\frac{1}{N}]^n & \cdots & [\frac{\ell}{N}]^n & \cdots & 1 \end{array} \right] \\ \underbrace{\hspace{1.5cm}}_{\mathbf{V}} \end{array} \begin{array}{c} \left[\begin{array}{c} \zeta(0) \\ \zeta(\frac{1}{N}) \\ \vdots \\ \zeta(\frac{\ell}{N}) \\ \vdots \\ \zeta(1) \end{array} \right] \\ \underbrace{\hspace{1.5cm}}_{\boldsymbol{\zeta}} \end{array}, \quad (34)$$

with each component $\zeta(\frac{\ell}{N})$ weakly positive and exactly L components strictly positive. Here \mathbf{V} is a Vandermonde matrix with many more columns (i.e., grid points) than rows (known moments), implying an under-determined system.¹¹ But the number L of grid points actually used could in principle be larger or smaller than n .

A result of Cohen and Yeredor [15, Theorem 1] applies to precisely this situation, stating that Equation 34 has a unique solution if $n \geq 2L$. We conclude the following.

Proposition 4. *In the consideration probability model, if F is a discrete distribution over L admissible types, with $n \geq 2L$, then F is uniquely determined by the aggregate choice shares \mathbf{p} .*

This result means that in practice any discrete distribution for the consideration probability ρ can be fully recovered from aggregate choice share data provided the number of alternatives is large relative to the number of cognitive types.

3.4.3 Type distributions with a density

Now suppose that the cognitive type distribution F admits a density f . In this case we will clearly not be able to recover F fully from a finite number n of moments. Instead, we aim to ensure that the known moments yield a reliable approximation of the true distribution.

Our method relies on standard techniques from the ‘‘Hausdorff moment problem’’ for distributions on a closed interval. Adopting a maximum entropy approach, define

¹¹See, e.g., Macon and Spitzbart [28] for the definition and properties of Vandermonde matrices.

the n th approximating density \hat{f}_n as the solution to the optimization problem

$$\max_{f_n} \int_0^1 [-\log f_n(\rho)] f_n(\rho) d\rho \quad (35)$$

subject to the (j th-moment) constraint

$$\int_0^1 \rho^j f_n(\rho) d\rho = m_j \quad (36)$$

for $j = 0, 1, \dots, n$. Mead and Papanicolaou [34, Theorem 2] show that such a solution exists and is unique;¹² and that for each bounded, continuous $\psi : [0, 1] \rightarrow \Re$ we have

$$\lim_{n \rightarrow \infty} \int_0^1 \psi(\rho) \hat{f}_n(\rho) d\rho = \int_0^1 \psi(\rho) f(\rho) d\rho. \quad (37)$$

Write \hat{F}_n for the distribution function associated with the approximating density \hat{f}_n . For any menu A and each $k \leq \min\{n, |A|\}$, we now have that

$$p(k_A; \hat{F}_n) = p(k_M; \hat{F}_n) = p(k_M; F) = p(k_A; F). \quad (38)$$

Here the first and third equalities follow from the observation that in the consideration probability model an alternative's choice share depends only on its rank on the menu according to the preference order. Moreover, in this model we have $\mathbf{p} = \mathbf{CQm}$ and the shares of the n best alternatives are determined by the first n moments. The constraints in Equation 36 guarantee that these moments coincide for the distributions \hat{F}_n and F , yielding the second equality in Equation 38. We summarize our findings as follows.

Proposition 5. *In the consideration probability model, if F admits a density then there exists a map $\mathbf{m} \mapsto \hat{F}_n$ such that:*

(i) The sequence $\langle \hat{F}_n \rangle_{n=1}^\infty$ converges weakly to F .

(ii) For any menu A and each $k \leq \min\{n, |A|\}$, we have $p(k_A; \hat{F}_n) = p(k_A; F)$.

¹²Indeed, the solution takes the form $\hat{f}_n(\rho) = \exp[-\sum_{j=0}^n \lambda_j \rho^j]$, where the quantities $\langle \lambda_j \rangle_{j=0}^n$ are the Lagrange multipliers on the constraints in Equation 36.

As already noted, the constraints in Equation 36 require each approximation \hat{F}_n to be observationally indistinguishable from the true distribution F in the sense that they generate the same first n moments, and hence the same aggregate choice shares over menu M . Proposition 5 reinforces this by guaranteeing that the cognitive heterogeneity in the population is reflected in two additional ways: Firstly, as the size of the observed menu increases, our approximation approaches (in the sense of weak convergence) the true distribution of the consideration probability. And secondly, each approximation F_n matches the true F not just over M , but also over the n best alternatives on any other menu A about which we may wish to make predictions.

3.5 Unobserved default outcome

3.5.1 Conditional choice shares

In this section we consider the prospects for cognitive identification when the default outcome is unobserved. Under this assumption our data set consists of the aggregate shares $\bar{p}(k) = \frac{p(k)}{1-p(d)}$ conditional on an active choice being made. Of course, any ratio of aggregate shares of the form $\tilde{p}(k, \ell) = \frac{\bar{p}(k)}{\bar{p}(\ell)} = \frac{p(k)}{p(\ell)}$ is unaffected by the conditioning, and so Equations 11–12 remain valid when restated in terms of the conditional shares and the associated probability masses $\bar{\pi}(\gamma) = \frac{\pi(\gamma)}{1-\pi(0)}$.

3.5.2 Parametric analysis

We begin by adapting each of the parametric examples introduced in Section 3.2 to the unobserved default scenario.

Example 1. [*Poisson γ ; continued*] Here $\mu = [n - 1] [\tilde{p}(n - 1, n) - 1]$, as above. \square

Example 2. [*Pascal γ ; continued*] Equation 17 can be written as $\tilde{p}(n - 1, n) = \frac{q[r+1]}{n-1} + 1$, and similarly from Equation 11 we obtain

$$\begin{aligned} \tilde{p}(n - 2, n) - \tilde{p}(n - 1, n) &= \frac{2}{n - 1} \left[\frac{\bar{\pi}(2)}{\bar{\pi}(1)} + \frac{3}{n - 2} \frac{\bar{\pi}(3)}{\bar{\pi}(1)} \right] \\ &= \frac{q[r + 1]}{n - 1} \left[1 + \frac{q[r + 2]}{n - 2} \right]. \end{aligned} \quad (39)$$

These equations can be solved simultaneously for the parameters

$$q = \frac{2\tilde{p}(n-1, n) - [n-1]\tilde{p}(n-1, n)^2 + [n-2]\tilde{p}(n-2, n) - 1}{\tilde{p}(n-1, n) - 1}, \quad (40)$$

$$r = \frac{2n\tilde{p}(n-1, n) - 2[n-1]\tilde{p}(n-1, n)^2 + [n-2]\tilde{p}(n-2, n) - n}{-2\tilde{p}(n-1, n) + [n-1]\tilde{p}(n-1, n)^2 - [n-2]\tilde{p}(n-2, n) + 1}. \quad \square \quad (41)$$

Example 3. [*uniform ρ ; continued*] From Equation 21 we have both $p(1) = \frac{1+\rho_{\min}}{2}$ and $p(2) = \frac{[2\rho_{\min}+1][1-\rho_{\min}]}{6}$. Hence $\tilde{p}(1, 2) = \frac{3[1+\rho_{\min}]}{[2\rho_{\min}+1][1-\rho_{\min}]}$, and it follows that

$$\rho_{\min} = \frac{\tilde{p}(1, 2) - 3 + \sqrt{3[3\tilde{p}(1, 2) - 1][\tilde{p}(1, 2) - 3]}}{4\tilde{p}(1, 2)}. \quad \square \quad (42)$$

Example 4. [*Beta ρ ; continued*] Equations 23–24 yield $\tilde{p}(2, 1) = \frac{b}{a+b+1}$, and likewise we can compute $\tilde{p}(3, 2) = \frac{b+1}{a+b+2}$. Solving for the parameters, we obtain

$$a = \frac{1 - 2\tilde{p}(3, 2) + \tilde{p}(3, 1)}{\tilde{p}(3, 2) - \tilde{p}(2, 1)}, \quad (43)$$

$$b = \frac{\tilde{p}(2, 1)[1 - \tilde{p}(3, 2)]}{\tilde{p}(3, 2) - \tilde{p}(2, 1)}. \quad \square \quad (44)$$

3.5.3 Recovering $n - 1$ probability mass ratios

In the non-parametric setting, it is straightforward to adapt Equation 28 to the case of an unobserved default outcome. Indeed, for each $\gamma = 2, 3, \dots, n$ we have that

$$\begin{aligned} \frac{\pi(\gamma)}{\pi(1)} &= \frac{\binom{n}{\gamma}}{n} \sum_{k=n-\gamma+1}^n [-1]^{[\gamma-1]-[n-k]} \binom{\gamma-1}{n-k} \frac{p(k)}{p(n)} \\ &= \frac{\binom{n}{\gamma}}{n} \sum_{k=n-\gamma+1}^n [-1]^{[\gamma-1]-[n-k]} \binom{\gamma-1}{n-k} \frac{\bar{p}(k)}{\bar{p}(n)}. \end{aligned} \quad (45)$$

Thus we can use the conditional choice shares to recover $n - 1$ probability mass ratios, though without knowledge of the default share $p(d) = \pi(0)$ we are of course unable to determine the masses themselves.

3.5.4 Consideration probability: Recovering $n - 1$ moment ratios

For the special case of the consideration probability model, Equation 32 can likewise be adapted to the unobserved default scenario. Here, for each $j = 2, 3, \dots, n$, we have

$$\frac{m_j}{m_1} = \sum_{k=1}^j [-1]^{k-1} \binom{j-1}{k-1} \frac{p(k)}{p(1)} = \sum_{k=1}^j [-1]^{k-1} \binom{j-1}{k-1} \frac{\bar{p}(k)}{\bar{p}(1)}. \quad (46)$$

This yields $n - 1$ ratios of raw moments, and we could proceed to use methods such as those in Section 3.4 above to approximate the shape of the cognitive type distribution F (the mean m_1 of which would remain undetermined without knowledge of the default share).¹³

4 Preference heterogeneity

4.1 Known taste distribution

Section 3 has studied the identification properties of our model of consideration set formation under the assumption that preferences are homogeneous. We now aim to show that the preceding analysis can be extended to allow for heterogeneous preferences, provided the taste distribution is known and statistically independent of the cognitive (type) distribution.¹⁴ We then proceed (in Section 4.2) to investigate the prospects for identification when both the taste and cognitive distributions are unknown.

To incorporate preference heterogeneity into the present framework, we order the alternatives arbitrarily as $M = \{1, 2, \dots, n\}$ and write $\varphi : M \rightarrow \{1, 2, \dots, n\}$ for the map that associates each option with its preference rank.¹⁵ We enumerate the possible rankings as $\langle \varphi_h \rangle_{h=1}^{n!}$, write τ_h for the probability of φ_h , and denote by $\mathbf{P}(h)$ the $n \times n$

¹³Horan [24] considers an unobserved default outcome in the context of a dataset with choices from multiple menus, showing that the identification properties of Manzini and Mariotti's [30] independent random consideration model remain largely intact.

¹⁴The distribution of taste parameters—such as discount factors or risk-aversion coefficients—may be treated as known for our purposes if these characteristics can be elicited from agents separately, in a setting (e.g., a laboratory experiment) where limited attention is thought to be irrelevant or controllable to an acceptable degree.

¹⁵This formulation maintains the assumption of linear order preferences imposed in Section 2.

permutation matrix corresponding to φ_h .¹⁶ With preference heterogeneity, Equation 27 then becomes

$$\mathbf{p} = \sum_{h=1}^{n!} \tau_h [\mathbf{P}(h)\mathbf{C}] \boldsymbol{\pi} = \left[\sum_{h=1}^{n!} \tau_h \mathbf{P}(h) \right] \mathbf{C} \boldsymbol{\pi}. \quad (47)$$

Provided the “average preference permutation matrix” $\mathbf{B} = \sum_{h=1}^{n!} \tau_h \mathbf{P}(h)$ has full rank, it follows that $\boldsymbol{\pi} = [\mathbf{BC}]^{-1} \mathbf{p}$, and similarly Equation 31 becomes $\mathbf{m} = [\mathbf{BCQ}]^{-1} \mathbf{p}$ for the special case of the consideration probability model. We conclude that the aggregate choice shares can still be used to find the probability masses in $\boldsymbol{\pi}$ and the raw moments in \mathbf{m} , as long as the taste distribution yields a nonsingular matrix \mathbf{B} .

As a convex combination of permutation matrices, the average preference permutation \mathbf{B} is always bistochastic.¹⁷ Clearly there exist taste distributions $\boldsymbol{\tau} = \langle \tau_h \rangle_{h=1}^{n!}$ for which this matrix is not invertible; e.g., the uniform distribution (with each $\tau_h = \frac{1}{n!}$) yields a singular \mathbf{B} (with each entry equal to $\frac{1}{n}$). However, invertibility is the generic situation, implying the following extensions of Propositions 2–3.¹⁸

Proposition 6. *In the consideration capacity model with known preference heterogeneity, for almost all taste distributions $\boldsymbol{\tau}$ the probability masses $\boldsymbol{\pi}$ are uniquely determined by the aggregate choice shares \mathbf{p} .*

Proposition 7. *In the consideration probability model with known preference heterogeneity, for almost all taste distributions $\boldsymbol{\tau}$ the raw moments \mathbf{m} are uniquely determined by the aggregate choice shares \mathbf{p} .*

The following example illustrates the handling of known preference heterogeneity in the context of the consideration capacity model.

Example 5. [*exploded logit*] Let $n = 3$, define $u : M \rightarrow \Re$ by $u(k) = \log k$, and suppose that the distribution of tastes is determined by an exploded logit based on u . For

¹⁶More explicitly, the permutation matrix $\mathbf{P}(h)$ translates the k th row of an $n \times n$ target matrix \mathbf{A} into the $\varphi_h(k)$ th row of the product $\mathbf{P}(h)\mathbf{A}$. Similarly, postmultiplying by $\mathbf{P}(h)$ permutes the columns of \mathbf{A} .

¹⁷A matrix is bistochastic if it is both left and right stochastic. The Birkhoff-von-Neumann Theorem states that the class of $n \times n$ bistochastic matrices is the convex hull of the set of $n \times n$ permutation matrices.

¹⁸Observe that $\det(\mathbf{B})$ is a polynomial function of $\boldsymbol{\tau} \in \Re^{n!}$, and recall that any real-valued polynomial function on a Euclidean space is either identically zero or nonzero almost everywhere (see, e.g., Caron and Traynor [12]). Since $\det(\mathbf{B})$ is nonzero for the case of homogeneous preferences, it is not identically zero, and thus \mathbf{B} is generically invertible.

instance, the probability assigned to the ranking φ_2 given by $2 \succ 3 \succ 1$ is calculated as

$$\tau_2 = \frac{e^{u(2)}}{e^{u(1)} + e^{u(2)} + e^{u(3)}} \times \frac{e^{u(3)}}{e^{u(1)} + e^{u(3)}} \times \frac{e^{u(1)}}{e^{u(1)}} = \frac{2}{6} \times \frac{3}{4} \times \frac{1}{1} = \frac{1}{4}. \quad (48)$$

The average preference permutation matrix is then

$$\begin{aligned} \mathbf{B} = \sum_{h=1}^{n!} \tau_h \mathbf{P}(h) &= \frac{1}{3} \underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}}_{\varphi_1: 3 \succ 2 \succ 1} + \frac{1}{4} \underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{\varphi_2: 2 \succ 3 \succ 1} + \frac{1}{6} \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}}_{\varphi_3: 3 \succ 1 \succ 2} \cdots \\ &+ \frac{1}{10} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}}_{\varphi_4: 1 \succ 3 \succ 2} + \frac{1}{12} \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\varphi_5: 2 \succ 1 \succ 3} + \frac{1}{15} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\varphi_6: 1 \succ 2 \succ 3} = \frac{1}{60} \begin{bmatrix} 10 & 15 & 35 \\ 20 & 24 & 16 \\ 30 & 21 & 9 \end{bmatrix}, \quad (49) \end{aligned}$$

which is nonsingular (with $\det(\mathbf{B}) = -\frac{1}{30}$). Now it is straightforward to compute

$$\begin{bmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \end{bmatrix} = \boldsymbol{\pi} = [\mathbf{BC}]^{-1} \mathbf{p} = \begin{bmatrix} \frac{15}{2}p(1) - 6p(2) + \frac{3}{2}p(3) \\ -15p(1) + 30p(2) - 15p(3) \\ \frac{17}{2}p(1) - 23p(2) + \frac{29}{2}p(3) \end{bmatrix}, \quad (50)$$

and as always $\pi(0) = p(d)$. \square

4.2 Unknown taste distribution

Continuing to allow for heterogeneous preferences, we next consider the problem of identifying the cognitive distribution when the taste distribution too is unknown. Here the information in a single observation of aggregate choices is clearly insufficient to reveal both distributions nonparametrically. Indeed, Propositions 2–3 already consume all n degrees of freedom in order to infer probability masses or raw moments of F . The impracticality of deducing cognition and tastes simultaneously from our basic dataset is illustrated in the following simple example.

Example 6. Let $n = 2$ and $\varphi_1(1) = 1$, so that τ_1 is the probability of the ranking $1 \succ 2$. Equation 47 then takes the form

$$\begin{bmatrix} p(1) \\ p(2) \end{bmatrix} = \mathbf{p} = [\mathbf{BC}] \boldsymbol{\pi} = \begin{bmatrix} \frac{1}{2} & \tau_1 \\ \frac{1}{2} & 1 - \tau_1 \end{bmatrix} \begin{bmatrix} \pi(1) \\ \pi(2) \end{bmatrix}, \quad (51)$$

an underdetermined system in which the cognitive distribution $\langle \pi(1), \pi(2) \rangle$ and the taste distribution τ_1 cannot be disambiguated. \square

To gain some leverage on the unknown tastes scenario it will be necessary to relax the stringent assumption that our dataset consists of aggregate choice shares from a single menu, and a variety of relaxations are possible.¹⁹ The approach we shall adopt here is to suppose that the researcher has access to choice data from the same population of agents on multiple “occasions” across which the cognitive distribution is stable. While we assume for convenience that the size of the menu is constant, the alternatives themselves need not be identical across occasions. For instance, the objects of choice could be interpreted as the same physical items at time-varying prices; the current model of a product offered in successive periods by a fixed set of suppliers; or the options available in an experiment with multiple rounds or treatments.

We assume further that our dataset consists of the *joint* distribution of choices across occasions; as arising, for example, from discrete choice panel data or from a sequence of discrete choice experiments. Although joint choice shares of this sort comprise “aggregate” data only from a somewhat literalist point of view, the agents in the population can remain anonymous in the sense that no observations on individuals will be required for our analysis other than their observed choices.²⁰

The advantage of this new multi-occasion setting is that it will allow us to deploy a powerful mathematical result on tensor decompositions to determine the cognitive

¹⁹One strategy would be to supply the researcher with aggregate data on choices from multiple subsets of the menu (cf. Geng and Ozbay [21]), while assuming stable tastes. Another strategy—explored in an earlier version of this paper—would be to supplement the dataset with covariates and estimate a random utility model of preference determination.

²⁰With I occasions and n alternatives, a single agent’s joint choice can be described by a unit vector in n^I -dimensional space. The aggregate choice frequencies for the population are then given by the sum of these vectors, which is equivalent to the aggregate joint distribution of choices in our dataset.

distribution even in the context of unknown and possibly changing tastes. Indeed, we shall find (in Proposition 8) that joint choice share data from as few as three occasions suffices generically to infer the consideration capacity distribution in full as well as substantial information about the distribution of tastes.

Formally, we index the occasions by $i = 1, \dots, I$ and suppose that on each occasion our population of agents chooses from a menu $M = \{1, \dots, n\}$ with default $d \notin M$. Here neither $k \in M$ nor the default d need represent the same economic outcome on different occasions, but the cardinality n of the menu is assumed to be constant.²¹ The taste distribution on occasion i is denoted by $\tau_i = \langle \tau_{ih} \rangle_{h=1}^{n!}$, and agents are assumed to retain their cognitive types across occasions so that the distribution F is stable. We write $p_\theta(k_1 \cdots k_I)$ for the joint probability that on each occasion i an individual of type θ chooses alternative k_i . Our dataset then consists of the corresponding population shares $p(k_1 \cdots k_I) = \int_{\Theta} p_\theta(k_1 \cdots k_I) dF$, and as before our objective is to use this data to deduce information about the underlying cognitive distribution F .

In the context of the consideration capacity model, we assume that the realizations of the consideration set Γ and the preference ranking φ_h are independent across occasions conditional on the (capacity) type γ . The analog of Equation 1 is then

$$p_\gamma(k_1 \cdots k_I) = \begin{cases} \prod_{i=1}^I \sum_{h: \varphi_h(k_i)=1} \tau_{ih} & \text{if } \gamma \geq n, \\ \prod_{i=1}^I \sum_{r=1}^{n-\gamma+1} \frac{\binom{n-r}{\gamma-1}}{\binom{n}{\gamma}} \sum_{h: \varphi_h(k_i)=r} \tau_{ih} & \text{if } 1 \leq \gamma < n, \\ 0 & \text{if } \gamma = 0; \end{cases} \quad (52)$$

where (for $1 \leq \gamma < n$) the product is over the various occasions i , the outer sum is over the possible ranking positions r of the chosen alternative k_i , and the inner sum is over the rankings that place k_i in position r . Now the analog of Equation 6 appears as

$$p(k_1 \cdots k_I) = \sum_{\gamma=1}^n \pi(\gamma) p_\gamma(k_1 \cdots k_I) = \sum_{\gamma=1}^n \pi(\gamma) \prod_{i=1}^I \sum_{r=1}^{n-\gamma+1} \frac{\binom{n-r}{\gamma-1}}{\binom{n}{\gamma}} \sum_{h: \varphi_h(k_i)=r} \tau_{ih}, \quad (53)$$

showing explicitly how the population choice shares are determined by the cognitive

²¹This assumption simplifies our notation considerably, but is not essential for the analysis below.

distribution π in conjunction with the taste distributions $\langle \tau_i \rangle_{i=1}^I$.

The following example illustrates the multi-occasion framework and demonstrates how joint choice share data can be used to infer the cognitive and taste distributions.

Example 7. Let $n = 2$, $I = 3$, and each $\varphi_{i1}(1) = 1$. The joint probabilities are the weighted sums of the two type-conditional joint probabilities, that is:

$$p(111) = \frac{\pi(1)}{8} + \pi(2)\tau_{11}\tau_{21}\tau_{31}, \quad (54)$$

$$p(112) = \frac{\pi(1)}{8} + \pi(2)\tau_{11}\tau_{21}[1 - \tau_{31}], \quad (55)$$

$$p(121) = \frac{\pi(1)}{8} + \pi(2)\tau_{11}[1 - \tau_{21}]\tau_{31}, \quad (56)$$

$$p(122) = \frac{\pi(1)}{8} + \pi(2)\tau_{11}[1 - \tau_{21}][1 - \tau_{31}]; \quad (57)$$

and so on. For $\tau_{11}, \tau_{21}, \tau_{31} \in (0, 1)$, we can combine the four equations above to yield

$$\frac{8p(111) - \pi(1)}{8p(112) - \pi(1)} = \frac{\tau_{31}}{1 - \tau_{31}} = \frac{8p(121) - \pi(1)}{8p(122) - \pi(1)}, \quad (58)$$

enabling recovery of the probability mass

$$\pi(1) = \frac{8[p(111)p(122) - p(112)p(121)]}{p(111) - p(112) - p(121) + p(122)}. \quad (59)$$

Equation 58 and two analogous relations then yield the three taste distributions

$$\tau_{11} = \frac{p(111) - p(112)}{p(111) + p(211) - p(112) - p(212)}, \quad (60)$$

$$\tau_{21} = \frac{p(111) - p(112)}{p(111) + p(121) - p(112) - p(122)}, \quad (61)$$

$$\tau_{31} = \frac{p(111) - p(121)}{p(111) + p(112) - p(121) - p(122)}. \quad (62)$$

Finally, from Equation 54 (or any of the other joint probability expressions) we obtain

$$\begin{aligned} \pi(2) = \frac{[p(111) - p(122)]^2 - [p(121) - p(112)]^2}{p(111) - p(112)} \dots \\ \dots \times \frac{p(111) - p(112) + p(211) - p(212)}{p(111) - p(112) - p(121) + p(122)}. \quad \square \quad (63) \end{aligned}$$

We aim now to demonstrate that the cognitive identification seen in Example 7 is a generic feature of the multi-occasion setting. To this end, we condition on the event $\gamma > 0$ and represent our dataset as a tensor \mathbf{S} of order I with dimensions $n \times \dots \times n$ and entries $S_{k_1, \dots, k_I} = \frac{p(k_1 \dots k_I)}{1 - p(d \dots d)}$.²² Writing $\mathbf{B}_i = \sum_{h=1}^{n!} \tau_{ih} \mathbf{P}(h)$ for the average preference permutation matrix on occasion i , we can then express Equation 53 more compactly as

$$\mathbf{S} = \sum_{\gamma=1}^n \bar{\pi}(\gamma) \otimes_{i=1}^I [\mathbf{B}_i \mathbf{C}] \mathbf{1}_\gamma, \quad (64)$$

where \otimes is the outer product operator and $\mathbf{1}_\gamma$ the unit vector for component γ .²³

Equation 64 decomposes the joint shares into a linear combination of n rank-1 tensors. The uniqueness properties of such decompositions have been studied extensively, with Kruskal [25, Theorem 4a] supplying a fundamental theorem that has been elucidated and refined by Sidiropoulos and Bro [40], Allman et al. [3], and Rhodes [36], among others. We shall use a special case of the result due to Rhodes [36, Corollary 2], adapted for our setting as follows.

Lemma 1. [Kruskal; Rhodes] *Given any collection $\langle \mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3 \rangle$ of invertible $n \times n$ matrices, the tensor $\mathbf{T} = \sum_{\gamma=1}^n [\mathbf{Z}_1 \mathbf{1}_\gamma \otimes \mathbf{Z}_2 \mathbf{1}_\gamma \otimes \mathbf{Z}_3 \mathbf{1}_\gamma]$ uniquely determines each \mathbf{Z}_i up to column rescaling and permutation. That is, for any $\langle \hat{\mathbf{Z}}_1, \hat{\mathbf{Z}}_2, \hat{\mathbf{Z}}_3 \rangle$ such that $\sum_{\gamma=1}^n [\hat{\mathbf{Z}}_1 \mathbf{1}_\gamma \otimes \hat{\mathbf{Z}}_2 \mathbf{1}_\gamma \otimes \hat{\mathbf{Z}}_3 \mathbf{1}_\gamma] = \mathbf{T}$ there exist invertible diagonal matrices $\langle \mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3 \rangle$ and a permutation matrix \mathbf{P} such that $\mathbf{D}_1 \mathbf{D}_2 \mathbf{D}_3 = \mathbf{I}_n$ and each $\hat{\mathbf{Z}}_i = \mathbf{Z}_i \mathbf{D}_i \mathbf{P}$.²⁴*

²²A tensor is a multidimensional array that generalizes the concept of a matrix to allow for an arbitrary number of indices—this number being the order of the tensor. The dimensions of a tensor indicate the number of possible values of each index, generalizing the number of rows and columns of a matrix.

²³Recall that the outer product of a pair of vectors is the first multiplied by the transpose of the second, and similarly each further outer product operation adds another dimension to the resulting array. A tensor is said to be of rank 1 if it is an outer product of vectors.

²⁴Here \mathbf{I}_n denotes the $n \times n$ identity matrix. Note that the actual result in Rhodes [36] is substantially more general than this statement, since he allows the \mathbf{Z}_i matrices to have different numbers of rows and

Setting $I = 3$ and applying this tool to our joint choice share tensor \mathbf{S} , we can show generic cognitive identification in the multi-occasion environment.

Proposition 8. *In the consideration capacity model with unknown preference heterogeneity and three occasions, if $\pi \gg 0$ then for almost all taste distributions $\langle \boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3 \rangle$ the probability masses π and average preference permutation matrices $\langle \mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3 \rangle$ are uniquely determined by the joint choice shares $p(k_1 k_2 k_3)$ for $1 \leq k_1, k_2, k_3 \leq n$.²⁵*

Proof. Writing $\mathbf{D}(\bar{\pi})$ for the diagonal matrix with entries $\bar{\pi} = \langle \bar{\pi}(\gamma) \rangle_{\gamma=1}^n \gg 0$, we can (following [3, p. 3118]) set $\mathbf{Z}_1 = [\mathbf{B}_1 \mathbf{C}] \mathbf{D}(\bar{\pi})$, $\mathbf{Z}_2 = \mathbf{B}_2 \mathbf{C}$, and $\mathbf{Z}_3 = \mathbf{B}_3 \mathbf{C}$, whereupon

$$\sum_{\gamma=1}^n [\mathbf{Z}_1 \mathbf{1}_\gamma \otimes \mathbf{Z}_2 \mathbf{1}_\gamma \otimes \mathbf{Z}_3 \mathbf{1}_\gamma] = \sum_{\gamma=1}^n \bar{\pi}(\gamma) \otimes_{i=1}^3 [\mathbf{B}_i \mathbf{C}] \mathbf{1}_\gamma = \mathbf{S}. \quad (65)$$

Here each $\mathbf{B}_i \mathbf{C}$ has full rank for almost all taste distributions $\boldsymbol{\tau}_i$ (see Footnote 18), and since $\bar{\pi} \gg 0$ it follows that each \mathbf{Z}_i is invertible. For duplicate parameters $\langle \hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2, \hat{\mathbf{B}}_3 \rangle$ and $\hat{\pi} \gg 0$ such that the corresponding $\sum_{\gamma=1}^n [\hat{\mathbf{Z}}_1 \mathbf{1}_\gamma \otimes \hat{\mathbf{Z}}_2 \mathbf{1}_\gamma \otimes \hat{\mathbf{Z}}_3 \mathbf{1}_\gamma] = \mathbf{S}$, Lemma 1 ensures that there exist rescalings $\langle \mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3 \rangle$ and a permutation \mathbf{P} such that

$$[\hat{\mathbf{B}}_1 \mathbf{C}] \mathbf{D}(\hat{\pi}) = \hat{\mathbf{Z}}_1 = \mathbf{Z}_1 \mathbf{D}_1 \mathbf{P} = [\mathbf{B}_1 \mathbf{C}] \mathbf{D}(\bar{\pi}) \mathbf{D}_1 \mathbf{P}, \quad (66)$$

$$\hat{\mathbf{B}}_2 \mathbf{C} = \hat{\mathbf{Z}}_2 = \mathbf{Z}_2 \mathbf{D}_2 \mathbf{P} = [\mathbf{B}_2 \mathbf{C}] \mathbf{D}_2 \mathbf{P}, \quad (67)$$

$$\hat{\mathbf{B}}_3 \mathbf{C} = \hat{\mathbf{Z}}_3 = \mathbf{Z}_3 \mathbf{D}_3 \mathbf{P} = [\mathbf{B}_3 \mathbf{C}] \mathbf{D}_3 \mathbf{P}. \quad (68)$$

Writing the vector of ones as $\mathbf{1} = \sum_{\gamma=1}^n \mathbf{1}_\gamma$, we have $\mathbf{1}^\top \hat{\mathbf{B}}_i \mathbf{C} = \mathbf{1}^\top [\mathbf{B}_i \mathbf{C}] \mathbf{D}_i \mathbf{P}$ for $i = 2, 3$ from Equations 67–68. Since \mathbf{B}_i and $\hat{\mathbf{B}}_i$ are bistochastic and \mathbf{C} is left stochastic, it follows that $\mathbf{1}^\top = \mathbf{1}^\top \mathbf{D}_i \mathbf{P}$ and thus $\mathbf{1}^\top = \mathbf{1}^\top \mathbf{P}^\top = \mathbf{1}^\top \mathbf{D}_i$. We conclude that $\mathbf{D}_2 = \mathbf{D}_3 = \mathbf{I}_n$, and hence $\mathbf{D}_1 = [\mathbf{D}_2 \mathbf{D}_3]^{-1} = \mathbf{I}_n$ as well.

Similarly, we have $\hat{\pi}^\top = \mathbf{1}^\top \mathbf{D}(\hat{\pi}) = \mathbf{1}^\top \mathbf{D}(\bar{\pi}) \mathbf{P} = \bar{\pi}^\top \mathbf{P}$ from Equation 66 and hence

one of them to have linearly dependent columns. This necessitates an additional hypothesis imposing a lower bound on the “Kruskal rank” (see [36, p. 1819]) of the \mathbf{Z}_i matrix in question—a hypothesis that is trivially satisfied in the square, full-rank case.

²⁵Our result shows that three occasions suffice (generically) for determination of π , and data from additional occasions will neither help nor hinder cognitive identification. Extensions of Kruskal’s theorem to tensors of order higher than three have been studied; e.g., by Sidiropoulos and Bro [40, Theorem 3]. We do not pursue this extension at present, since it is orthogonal to our main goal of inferring cognitive heterogeneity from minimal data.

$\hat{\pi} = \mathbf{P}^\top \bar{\pi}$. It follows that $\mathbf{D}(\hat{\pi}) = \mathbf{D}(\mathbf{P}^\top \bar{\pi}) = \mathbf{P}^\top \mathbf{D}(\bar{\pi}) \mathbf{P}$, so that Equation 66 yields $[\hat{\mathbf{B}}_1 \mathbf{C}] \mathbf{P}^\top \mathbf{D}(\bar{\pi}) \mathbf{P} = [\mathbf{B}_1 \mathbf{C}] \mathbf{D}(\bar{\pi}) \mathbf{P}$ and $[\hat{\mathbf{B}}_1 \mathbf{C}] \mathbf{P}^\top = \mathbf{B}_1 \mathbf{C}$. Together with Equations 67–68, this shows that each $\hat{\mathbf{B}}_i = \mathbf{B}_i [\mathbf{C} \mathbf{P} \mathbf{C}^{-1}]$. The duplicate parameters $\langle \hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2, \hat{\mathbf{B}}_3 \rangle$ and $\hat{\pi}$ are thus seen to result from label swapping; i.e., a garbling of the type distribution $\bar{\pi}$ (via the permutation \mathbf{P}^\top) that is reversed by adjustments to the taste distributions \mathbf{B}_i . When labels are assigned correctly we have $\mathbf{P} = \mathbf{I}_n$, so that $\hat{\pi} = \bar{\pi}$ and each $\hat{\mathbf{B}}_i = \mathbf{B}_i$. \square

Observe that the joint choice share data do not fully determine the taste distributions $\langle \tau_1, \tau_2, \tau_3 \rangle$. Indeed, factorial explosion of the number of rankings of n alternatives makes it obvious that such identification cannot be possible in general. In relation to tastes, what the joint choice shares determine are the average preference permutation matrices $\langle \mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3 \rangle$, which record the probabilities $\sum_{h: \varphi_h(k_i)=r} \tau_{ih}$ of alternative k_i being ranked in position r on occasion i . Thus the remaining scope for non-uniqueness is limited to shifts between rankings that preserve these marginal position probabilities.

The identification of the cognitive distribution and the average preference permutations in Proposition 8 requires no parametric assumptions on the primitives of the consideration capacity model. Introducing such assumptions may enable us to refine the taste distribution beyond the probabilities in each \mathbf{B}_i ; a task that is simplified by the problem now amounting to one of determining the preferences of agents with known cognitive type. In fact, since the type-conditional choice distributions have already been recovered, we could simply focus on the behavior of full-attention types (with $\gamma \geq n$) and apply known techniques to elicit the distribution of preferences on each occasion. For instance, we could assume that the type-conditional choices result from a random utility model (RUM) with a known error distribution, or by a single-crossing RUM as defined in Apesteguia et al. [5]. In any event, such parametric assumptions would be unrelated to the limited-attention aspects of the model and unnecessary for achieving our primary goal in this section, which is to identify the *cognitive* distribution in the presence of unknown taste heterogeneity.

5 Conclusion

In this paper we have shown how aggregate choice shares can identify the distribution of cognitive characteristics in a population of agents exhibiting limited attention. A key advantage of our theory is that it uses minimal data: either choice shares from a single menu (for a known taste distribution), or joint shares from as few as three menus (for an unknown taste distribution). This contrasts with prior theoretical work on bounded rationality, much of which uses individual choices from a family of overlapping menus. Both approaches can be brought to bear on data, but in our view the present framework is better suited to the practice of empirical research on discrete choice.

Under homogeneous tastes, we find that both the consideration capacity model and the special case of the consideration probability model are highly tractable in terms of cognitive identification. In each model the aggregate choice shares are seen to be linear in quantities that are highly informative about the cognitive distribution; respectively, small capacity probabilities and low-order raw moments. The resulting linear systems are recursive and easily solved for the quantities in question. Indeed, our main findings demonstrate that for large menus the cognitive distribution is essentially fully identified, and even for smaller menus we can infer important features of this distribution (Propositions 2–5). These findings extend generically to heterogeneous tastes with a known distribution, and when the taste distribution is unknown a parsimonious extension to our dataset ensures generic cognitive identification (Propositions 6–8).

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 instance, by relaxing the assumption that each Γ with $|\Gamma| = \gamma$ is equally likely to occur in the consideration capacity model, or by allowing ρ to vary by alternative in the consideration probability model (see Brady and Rehbeck [9]). A second is to bring additional models of bounded rationality—incorporating phenomena such as framing effects or satisficing—into the present setting. And a third is to develop a complete econometric specification of the multi-occasion environment in Section 4.2 that can be used to estimate preference and attention characteristics from joint choice share data.

References

- [1] Jason Abaluck and Abi Adams (2017). What do consumers consider before they choose? Identification from asymmetric demand responses. Unpublished.
- [2] Victor H. Aguiar, Maria Jose Boccardi, Nail Kashaev, and Jeongbin Kim (2018). Does random consideration explain behavior when choice is hard? Evidence from a large-scale experiment. Unpublished.
- [3] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics* 37:3099–3132.
- [4] Jose Apesteguia and Miguel A. Ballester (2013). Choice by sequential procedures. *Games and Economic Behavior* 77:90–99.
- [5] Jose Apesteguia, Miguel A. Ballester, and Jay Lu (2017). Single-crossing random utility models. *Econometrica* 85:661–674.
- [6] Nicholas Baigent and Wulf Gaertner (1996). Never choose the uniquely largest: A characterization. *Economic Theory* 8:239–249.
- [7] Levon Barseghyan, Maura Coughlin, Francesca Molinari, and Joshua Teitelbaum (2018). Heterogeneous consideration sets and preferences. Unpublished.
- [8] Levon Barseghyan, Francesca Molinari, and Matthew Thirkettle (2019). Discrete choice under risk with limited consideration. Unpublished.
- [9] Richard L. Brady and John Rehbeck (2016). Menu-dependent stochastic feasibility. *Econometrica* 84:1203–1223.
- [10] Andrew Caplin and Mark Dean (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review* 105:2183–2203.
- [11] Andrew Caplin, Mark Dean, and Daniel Martin (2011). Search and satisficing. *American Economic Review* 101:2899–2922.

- [12] Richard Caron and Tim Traynor (2005). The zero set of a polynomial. Unpublished.
- [13] Matias D. Cattaneo, Xinwei Ma, Yusufcan Masatlioglu, and Elchin Suleymanov (2017). A random attention model. Unpublished.
- [14] Vadim Cherepanov, Timothy Feddersen, and Alvaro Sandroni (2013). Rationalization. *Theoretical Economics* 8:775–800.
- [15] 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*.
- [16] Gregory S. Crawford, Rachel Griffith, and Alessandro Iaria (2018). Preference estimation with unobserved choice set heterogeneity using sufficient sets. Unpublished.
- [17] Geoffroy de Clippel, Kfir Eliaz, and Kareen Rozen (2014). Competing for consumer inattention. *Journal of Political Economy* 122:1203–1234.
- [18] Henrique de Oliveira, Tommaso Denti, Maximilian Mihm, and Kemal Ozbek (2017). Rationally inattentive preferences and hidden information costs. *Theoretical Economics* 12:621–654.
- [19] Federico Echenique, Kota Saito, and Gerelt Tserenjigmid (2017). The perception-adjusted Luce model. *Mathematical Social Sciences* 93:67–76.
- [20] 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.
- [21] Sen Geng and Erkut Y. Ozbay (2018). Choice with limited capacity. Unpublished.
- [22] Justine S. Hastings, Ali Hortaçsu, and Chad Syverson (2017). Sales force and competition in financial product markets: The case of Mexico’s social security privatization. *Econometrica* 85:1723–1761.

- [23] Elisabeth Honka, Ali Hortaçsu, and Maria Ana Vitorino (2017). Advertising, consumer awareness, and choice: Evidence from the U.S. banking industry. *Rand Journal of Economics* 48:611–646.
- [24] Sean Horan (2018). Random consideration and choice: A case study of “default” options. Unpublished.
- [25] Joseph B. Kruskal (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications* 18:95–138.
- [26] Zhentong Lu (2016). Estimating multinomial choice models with unobserved choice sets. Unpublished.
- [27] R. Duncan Luce (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley.
- [28] Nathaniel Macon and Abraham Spitzbart (1958). Inverses of Vandermonde matrices. *The American Mathematical Monthly* 65:95–100.
- [29] Paola Manzini and Marco Mariotti (2007). Sequentially rationalizable choice. *American Economic Review* 97:1824–1839.
- [30] Paola Manzini and Marco Mariotti (2014). Stochastic choice and consideration sets. *Econometrica* 82:1153–1176.
- [31] Yusufcan Masatlioglu and Daisuke Nakajima (2013). Choice by iterative search. *Theoretical Economics* 8:701–728.
- [32] Yusufcan Masatlioglu, Daisuke Nakajima, and Erkut Y. Ozbay (2012). Revealed attention. *American Economic Review* 102:2183–2205.
- [33] Daniel McFadden (2001). Economic choices. *American Economic Review* 91:351–378.
- [34] Laurence R. Mead and Nikos Papanicolaou (1984). Maximum entropy in the problem of moments. *Journal of Mathematical Physics* 25:2404–2417.

- [35] Efe A. Ok, Pietro Ortoleva, and Gil Riella (2014). Revealed (p)reference theory. *American Economic Review* 105:299–321.
- [36] John A. Rhodes (2010). A concise proof of Kruskal’s theorem on tensor decomposition. *Linear Algebra and its Applications* 432:1818–1824.
- [37] John H. Roberts and James M. Lattin (1997). Consideration: Review of research and prospects for future insights. *Journal of Marketing Research* 34:406–410.
- [38] Yuval Salant and Ariel Rubinstein (2008). (A, f) : Choice with frames. *Review of Economic Studies* 75:1287–1296.
- [39] Allan Shocker, Moshe Ben-Akiva, Bruno Boccara, and Prakash Nedungadi (1991). Consideration set influences on consumer decision making and choice: Issues, models, and suggestions. *Marketing Letters* 2:181–198.
- [40] Nicholas D. Sidiropoulos and Rasmus Bro (2000). On the uniqueness of multilinear decomposition of N-way arrays. *Journal of Chemometrics* 14:229–239.
- [41] Christopher A. Sims (2003). Implications of rational inattention. *Journal of Monetary Economics* 50:665–690.
- [42] Michelle Sovinsky Goeree (2008). Limited information and advertising in the U.S. personal computer industry. *Econometrica* 76:1017–1074.
- [43] Kenneth E. Train (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.
- [44] Christopher J. Tyson (2008). Cognitive constraints, contraction consistency, and the satisficing criterion. *Journal of Economic Theory* 138:51–70.
- [45] Christopher J. Tyson (2013). Behavioral implications of shortlisting procedures. *Social Choice and Welfare* 41:941–963.
- [46] 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.