Inferring cognitive heterogeneity from aggregate choices

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

Published in
Econometrica

Link to external publisher version
https://doi.org/10.3982/ECTA16382

INFERRING COGNITIVE HETEROGENEITY FROM AGGREGATE CHOICES

Valentino Dardanoni
Dipartimento SEAS, Università degli Studi di Palermo

Paola Manzini
Department of Economics, University of Sussex, and IZA

Marco Mariotti
School of Economics and Finance, Queen Mary University of London

Christopher J. Tyson
School of Economics and Finance, Queen Mary University of London

February 7, 2020

Abstract

Theories of bounded rationality often assume a rich dataset of choices from many overlapping menus, limiting their practical applicability. In contrast, we study the problem of identifying the distribution of cognitive characteristics in a population of agents from a minimal dataset that consists of aggregate choice shares from a single menu, and includes no observable covariates of any kind. With homogeneous preferences, we find that "consideration capacity" and "consideration probability" distributions can both be recovered effectively if the menu is sufficiently large. This remains true generically when tastes are heterogeneous with a known distribution. When the taste distribution is unknown, we show that joint choice share data from three "occasions" are generically sufficient for full identification of the cognitive distribution, and also provide substantial information about tastes.

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

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

Valentino Dardanoni: valentino.dardanoni@unipa.it
Paola Manzini: p.manzini@sussex.ac.uk
Marco Mariotti: m.mariotti@qmul.ac.uk
Christopher J. Tyson: c.j.tyson@qmul.ac.uk

For valuable comments and suggestions that have improved this paper, we would like to thank four 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 Alicante, Bilkent, CERGE-EI, Edinburgh, Leuven, LSE, Naples, Oxford, St Andrews, Sussex, Tilburg, UCL, ULB (ECARES), Vienna, Warwick (CRETA), and Zurich; 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 (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. Dardanoni acknowledges support from MIUR through grant PRIN 2015. This paper is a radically revised descendant of a 2017 working paper with the same title, from which it differs substantially.

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. (2018), Cattaneo et al. (2017), Gibbard (2019), and Masatlioglu et al. (2012) require data for all possible menus drawn from a universal set of alternatives; Manzini and Mariotti (2014) impose a stringent "richness" assumption on their dataset; and Caplin and Dean (2015) 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. 2018 and Caplin and Dean 2015). 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 more direct link between theory and what is feasible empirically.

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 examines the extent to which the cognitive distribution is identified by a given model of bounded rationality per se—once it has been stripped of the richness of menu variation—and *prior to any ancillary econometric specification that may include covariates for the individuals or the alternatives*. In this respect our primitives and objectives remain typical of those in conventional abstract choice theory, and this is one way that our contribution differs from recent work in which identification is facilitated by access to observable covariates (see, e.g., Abaluck and Adams 2017, Barseghyan et al. 2019a, and Barseghyan et al. 2019b). Our model can be extended and tailored to specific applications by introducing such covariates, as is often done when decision-theoretic models are brought to bear on data.[5] Although such extensions and specializations must be left for future work, at places we will note how the presence of additional data could aid our identification exercise.

1.2. *Cognitive models*

In our general framework, each agent has a *cognitive type* parameter $\theta \in \Theta \subset \Re$ that is distributed in the population according to a cumulative distribution function F. Given preferences over the menu, an

[1] This 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 (2013), Baigent and Gaertner (1996), Caplin et al. (2011), Cherepanov et al. (2013), de Oliveira et al. (2017), Echenique et al. (2018), Manzini and Mariotti (2007), Masatlioglu and Nakajima (2013), Ok et al. (2015), Salant and Rubinstein (2008), and Tyson (2008, 2013), among numerous others.

[2] Even 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.

[3] This usage follows the marketing literature; see, e.g., Roberts and Lattin (1997) and Shocker et al. (1991). While we view the consideration set as a manifestation of bounded rationality, other interpretations are possible: Alternatives may fail to be considered due to habit formation, search costs, or other forms of rational inattention (see, e.g., Caplin and Dean 2015 and Sims 2003).

[4] Alternatively, 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.

[5] For example, the Luce (1959) model of probabilistic choice is formulated in terms of abstract utilities, but is implemented empirically as the multinomial or conditional logit model in which utilities are linear functions of observable characteristics of the agents or alternatives.

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 individual 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 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.[6]

As already noted, we study bounded rationality in the form of limited attention. Here the cognitive parameter θ influences the formation of the decision maker's consideration set, and more specifically the number of alternatives that are considered. In the "consideration capacity" model, the parameter $\gamma \in \{0, 1, 2, \ldots\}$ 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 $\gamma \to \infty$ or $\rho = 1$, as appropriate.[7]

1.3. *Preview of results*

We begin by assuming that the population has homogeneous tastes. In this case preference identification is not challenging (Proposition 1), which enables us to concentrate entirely on the cognitive identification problem. Here our attention model is fully identified by a small number of observed choice shares under several natural functional forms for F (see Appendix A.1). But even without a parametric specification, the cognitive distribution can for practical purposes be fully recovered if the menu of alternatives is large enough. In the context of the consideration capacity model, the aggregate choice shares identify the probabilities of all capacities less than the cardinality n of the menu (Proposition 2). Similarly, for the consideration probability model the choice shares identify the first n raw moments of F (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 (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 raw moments), so that closed-form expressions for these quantities can be obtained by inverting an (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 (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." Here we employ a powerful mathematical result on the uniqueness of tensor decompositions, which to our knowledge has not previously been used in the bounded rationality literature. (This methodology may be of independent interest, since its potential extends well beyond the specific models studied in the present paper.) In the context of the consideration capacity model, we show that joint choice share data are generically sufficient for full identification of the cognitive distribution, and also provide substantial information about the taste distribution (Proposition 8).

1.4. *Related empirical literature*

While remaining entirely theoretical in orientation, 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 (2017) construct a general econometric framework in which product characteristics are observable (unlike our setting), and exploit asymmetries in cross-characteristic choice probability

[6] Note 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 2009 and McFadden 2001.) 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 1959), as typically assumed in relation to tastes.

[7] Variants of the consideration capacity model are used by Barseghyan et al. (2019a) and Barseghyan et al. (2019b) to study discrete choice with heterogeneous consideration sets, and by de Clippel et al. (2014) to study price competition in a setting where consumers exhibit limited attention. The consideration probability model employed here is the one sketched in Manzini and Mariotti (2014, Section 7.2).

responses to identify consideration sets. For choice under risk, Barseghyan et al. (2019a) study preferences and attention in an extremely general model, with minimal assumptions about the process of consideration-set formation; targeting *partial* identification of its components.

Cattaneo et al. (2017) postulate "monotonic attention," a constraint on how stochastic consideration sets can change across menus, and use this assumption to derive testable restrictions on choice probabilities. Aguiar et al. (2018) test random consideration models at the population level in a large scale online experiment, finding support for a specification with heterogeneous preferences and logit attention. Both of these contributions, however, depend on substantial menu variation.

Crawford et al. (2020) devise a model-free identification strategy based on reducing the menu to a "sufficient set" of alternatives that are certain to be considered. Gaynor et al. (2016) exploit institutional changes to identify consideration sets in hospital choice, while Honka et al. (2017) exemplify the approach of treating consideration sets as the outcome of a search process.[8]

Lu (2019) develops a methodology for estimating multinomial choice models that employs known upper and lower bounds on the consideration set. Sovinsky Goeree (2008) studies the impact of marketing on consideration, using advertising data and observable product characteristics to separate the utility and attentional components of demand. Van Nierop et al. (2010) propose a specific model of brand choice accommodating both stated and revealed consideration-set data, which they apply to an experiment on merchandising strategies.

The paper that relates most closely to our contribution is Barseghyan et al. (2019b), which uses the attention allocation process that we refer to as the "consideration capacity model." But here again their identification strategy relies upon observable covariates—a basic difference from our methodology.

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.

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 (identical versions of) 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 complements 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 kth best option on A, and the full menu appears as $A = \{1_A, 2_A, \ldots, n_A\}$, where $n_A = |A|$.

[8]The 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. (2017) exposure to a sales force influences the probability that financial products are considered.

We introduce 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, and $p(k_A) = \int_\Theta p_\theta(k_A)\mathrm{d}F$ for the overall share in the population. Similarly, we write $p_\theta(d_A)$ for the probability that type θ accepts the default, and $p(d_A) = \int_\Theta p_\theta(d_A)\mathrm{d}F$ 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 in aggregate we likewise have $[\sum_{k=1}^{n_A} p(k_A)] + p(d_A) = 1$. 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 $|M| = n_M \geq 2$. 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, $p(k)$ and $p(d)$ for the population shares, and $n = n_M$ for the cardinality of the menu. Our goal is then to deduce information about the cognitive distribution F from the data in $\langle p(1), p(2), \ldots, p(n), p(d)\rangle$.

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 suboptimal decision.

2.2. *The consideration capacity model*

Let $\gamma \in \{0, 1, 2, \ldots\} = \Theta$ 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 $\gamma \geq n$.[9]

The type-conditional choice frequencies can now be expressed as

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

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

Defining the probability masses

$$\begin{aligned} \pi(0) &= F(0), \\ \forall \gamma \in \{1, 2, \ldots, n-1\}\text{:}\ \pi(\gamma) &= F(\gamma) - F(\gamma - 1), \\ \pi(n) &= 1 - F(n-1); \end{aligned}$$

the corresponding aggregate shares are then

$$p(k) = \sum_{\gamma=1}^{n-k+1} p_\gamma(k)\pi(\gamma) = \sum_{\gamma=1}^{n-k+1} \frac{\binom{n-k}{\gamma-1}}{\binom{n}{\gamma}}\pi(\gamma), \tag{2}$$

$$p(d) = \pi(0). \tag{3}$$

[9] We 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, 2, and 10 below modified accordingly). While this generalization causes no significant difficulty for the 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.

Observe that for $1 \leq k < n$ we can use Equation 2 to compute

$$p(k) - p(k+1) = \frac{\pi(n-k+1)}{\binom{n}{n-k+1}} + \sum_{\gamma=2}^{n-k} \frac{\binom{n-k-1}{\gamma-2}}{\binom{n}{\gamma}} \pi(\gamma). \tag{4}$$

This relation shows that, when we move one ordinal step up the preference scale, the aggregate choice share increases for two reasons: Firstly, the kth 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 ways of populating the rest of the consideration set with inferior options.

Note also that setting $k = n$ in Equation 2 yields $p(n) = \pi(1)/n$ and hence

$$\pi(1) = np(n). \tag{5}$$

Similarly, setting $k = n-1$ in Equation 4 yields $p(n-1) - p(n) = \frac{2\pi(2)}{n[n-1]}$ and hence

$$\pi(2) = \frac{n[n-1]}{2}[p(n-1) - p(n)]. \tag{6}$$

Equations 5–6 prefigure the recursive method employed in Section 3 to identify the cognitive type distribution, in which the probabilities $\pi(1), \ldots, \pi(n-1)$ are deduced sequentially, with one additional choice share used at each step.

Finally, using Equation 5, we can write Equation 4 in terms of probability ratios as

$$\frac{p(k) - p(k+1)}{p(n)} = \frac{n}{\binom{n}{n-k+1}} \frac{\pi(n-k+1)}{\pi(1)} + n \sum_{\gamma=2}^{n-k} \frac{\binom{n-k-1}{\gamma-2}}{\binom{n}{\gamma}} \frac{\pi(\gamma)}{\pi(1)}. \tag{7}$$

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

$$\frac{\pi(2)}{\pi(1)} = \frac{n-1}{2}\left[\frac{p(n-1)}{p(n)} - 1\right] \tag{8}$$

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 (2014) that emphasizes the interpretation of attention as general awareness of the choice environment. 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

$$\pi(\gamma) = \binom{n}{\gamma} \int_0^1 \rho^\gamma [1-\rho]^{n-\gamma} \mathrm{d}F, \tag{9}$$

and clearly $\pi(\gamma) = 0$ for $\gamma > n$. Adapting Equations 2–3 to this special case, we obtain the aggregate choice shares

$$p(k) = \sum_{\gamma=1}^{n-k+1} \binom{n-k}{\gamma-1} \int_0^1 \rho^\gamma [1-\rho]^{n-\gamma} \mathrm{d}F = \int_0^1 \rho [1-\rho]^{k-1} \mathrm{d}F, \tag{10}$$

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

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. To see that revelation is not automatic, observe that each alternative's total choice probability aggregates the probability of being chosen conditional on each cognitive type, and the various types contribute for different members of M. In the extreme case where the population consists entirely of (fully rational) types $\gamma \geq n$, only the best alternative would be revealed; a population containing types $\gamma \geq n-1$ would reveal the best two alternatives; and so on. Since $\gamma = 2$ alone makes a contribution for every member of M, without this type the common preference between at least one pair of alternatives would not be identified.

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

(i) $p(k) \geq p(k+1)$.

(ii) $\sum_{\gamma=2}^{n-k+1} \pi(\gamma) > 0$ *if and only if* $p(k) > p(k+1)$.

(iii) $\pi(2) > 0$ *if and only if* $p(1) > p(2) > \cdots > p(n)$.

Here (i) holds since each term on the right-hand-side of Equation 4 is nonnegative, and (ii) follows since each such term is zero only if the relevant capacity probability is zero. Moreover $\pi(2)$ appears in Equation 4 for all $1 \leq k < n$, and thus strict positivity of this single probability suffices for full revelation of preferences. This fact, recorded in (iii), can also be specialized to the consideration probability model using Equation 9.

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

We conclude that, under the homogeneous tastes assumption, 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 $p(1) > p(2) > \cdots > p(n)$ faithfully reflect the underlying preference order.

3.2. *Cognitive identification*

3.2.1. *The nonparametric inference problem*

When the cognitive type distribution has a known functional form, its parameters can often be deduced from a few appropriately selected choice-share observations. (This is demonstrated in Appendix A.1 by means of four examples that highlight non-obvious ways that aggregate choices can convey information about F.) Yet even in the absence of functional-form assumptions, identification of the type distribution remains highly tractable for the consideration capacity model. 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. What's more, each linear system has a simple triangular structure that enables us to solve it 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 3.3).

3.2.2. *Recovering n probability masses*

Absent parametric assumptions, the aggregate choice shares are given by Equation 2. These relations can be written together in matrix form as

$$\underbrace{\begin{bmatrix} p(1) \\ \vdots \\ p(k) \\ \vdots \\ p(n) \end{bmatrix}}_{\boldsymbol{p}} = \underbrace{\begin{bmatrix} 1/n & \cdots & \gamma/n & \cdots & 1 \\ \vdots & & \vdots & & \vdots \\ 1/n & \cdots & \binom{n-k}{\gamma-1}/\binom{n}{\gamma} & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 1/n & \cdots & 0 & \cdots & 0 \end{bmatrix}}_{\boldsymbol{C}} \underbrace{\begin{bmatrix} \pi(1) \\ \vdots \\ \pi(\gamma) \\ \vdots \\ \pi(n) \end{bmatrix}}_{\boldsymbol{\pi}}. \tag{11}$$

The upper anti-triangular and left-stochastic matrix $\boldsymbol{C}$ has a lower anti-triangular inverse, allowing us to write $\boldsymbol{\pi} = \boldsymbol{C}^{-1}\boldsymbol{p}$.[10] Accordingly, we can calculate the components of $\boldsymbol{\pi}$ as

$$\pi(\gamma) = \binom{n}{\gamma} \sum_{k=n-\gamma+1}^{n} [-1]^{[\gamma-1]-[n-k]} \binom{\gamma-1}{n-k} p(k), \tag{12}$$

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 $\gamma = 0, 1, \ldots, n-1$ 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 $\boldsymbol{p}$.*

As an aside, note that the key features of the model for Proposition 2 are that $\boldsymbol{C}$ is known and invertible. These do not require the assumption that the agent is equally likely to consider each $\Gamma \subset M$ with $|\Gamma| = \gamma$, which we have imposed for simplicity as a point of departure. This observation is illustrated by the following example.

Example 1 (salience weights)**.** Let $n = 3$; assign to each alternative k a weight $w_k > 0$, assumed to be known to the researcher; and define the polynomials $W_1 = w_1 + w_2 + w_3$ and $W_2 = w_1w_2 + w_1w_3 + w_2w_3$. Conditional on $\gamma = 1$, let $\Gamma = \{k\}$ with probability w_k/W_1; and conditional on $\gamma = 2$, let $\Gamma = \{j, k\}$ (for $j \neq k$) with probability w_jw_k/W_2. In this case the analog of Equation 11 is

$$\underbrace{\begin{bmatrix} p(1) \\ p(2) \\ p(3) \end{bmatrix}}_{\boldsymbol{p}} = \underbrace{\begin{bmatrix} w_1/W_1 & [w_1w_2 + w_1w_3]/W_2 & 1 \\ w_2/W_1 & w_2w_3/W_2 & 0 \\ w_3/W_1 & 0 & 0 \end{bmatrix}}_{\boldsymbol{C}(w_1,w_2,w_3)} \underbrace{\begin{bmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \end{bmatrix}}_{\boldsymbol{\pi}}.$$

For instance, if $w_1 = 1$, $w_2 = 2$, and $w_3 = 4$, then $\boldsymbol{p} = \boldsymbol{C}(1, 2, 4)\boldsymbol{\pi}$ and we can compute

$$\begin{bmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \end{bmatrix} = \boldsymbol{\pi} = \boldsymbol{C}(1, 2, 4)^{-1}\boldsymbol{p} = \frac{1}{8} \begin{bmatrix} 14p(3) \\ 14p(2) - 7p(3) \\ 8p(1) - 6p(2) + p(3) \end{bmatrix}.$$

Here the matrix $\boldsymbol{C}(w_1, w_2, w_3)$ remains known, upper anti-triangular, and invertible for any salience weights. If salience is severely misaligned with the agents' tastes, then it is possible that the choice shares will no longer directly reveal the preference order. But note that the degree of misalignment embodied in $\boldsymbol{C}(1, 2, 4)$, together with the decidedly adverse capacity distribution $\boldsymbol{\pi} = \langle 1/2, 1/3, 1/6 \rangle$, is not enough to generate this phenomenon. (The resulting shares are $\boldsymbol{p} = \langle 8/21, 7/21, 6/21 \rangle$.) In any case, we shall return to this issue in Section 4, where we consider the prospects for preference revelation in a much more general setting that allows for taste heterogeneity. □

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

3.2.3. *Consideration probability: Recovering n moments*

Returning to the special case of the consideration probability model, let us write the jth raw moment of the type distribution as $m_j = \int_0^1 \rho^j \mathrm{d}F$. The binomial in Equation 9 can then be expanded to yield

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

In matrix form, these relations appear as

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

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

$$\boldsymbol{m} = \boldsymbol{Q}^{-1}\boldsymbol{\pi} = \boldsymbol{Q}^{-1}[\boldsymbol{C}^{-1}\boldsymbol{p}] = [\boldsymbol{C}\boldsymbol{Q}]^{-1}\boldsymbol{p}. \tag{13}$$

Performing this calculation, the raw moments are given explicitly by

$$m_j = \sum_{k=1}^{j} [-1]^{k-1} \binom{j-1}{k-1} p(k). \tag{14}$$

We summarize our conclusions for the special case as follows.

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

3.3. *Consideration probability: Beyond moments*

3.3.1. *From moments to distributions*

Continuing to focus on the consideration probability model, we now investigate what can be learned about F itself from the information supplied by Proposition 3. To this end, throughout Section 3.3 we shall treat as known a finite number of raw moments of the cognitive distribution. It is intuitive that this information constrains the shape of any sufficiently well behaved F, with more moments known generating tighter constraints. However, two distributions that share certain raw moments could in principle have differences that are relevant for questions we may wish to study. For instance, we might hope to measure the fraction of agents with very low or very high values of ρ, or to reveal the cognitive type of a particularly small subpopulation. For such purposes we may want to have confidence that the raw moment information can be turned into a distribution that matches or closely approximates F over the entire range of possible ρ-values.

We now proceed to outline two different strategies for ensuring that the moment information adequately captures the cognitive type distribution. The first will rely on discreteness of the distribution and ensure a unique characterization of F, while the second will rely on the existence of a density and guarantee convergence to F in the limit as $n \to \infty$.

3.3.2. *Discrete type distributions*

Suppose that F is a discrete distribution, with the consideration probability ρ taking on the list of 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 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 $\langle \xi(\rho_1), \xi(\rho_2), \ldots, \xi(\rho_L) \rangle$, each strictly positive and together summing to one, so that the jth raw moment of F appears as

$$m_j = \sum_{\ell=1}^{L} \xi(\rho_\ell) \rho_\ell^j. \tag{15}$$

Treating the first n moments as known, Equation 15 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 $\langle 0, 1/N, 2/N, \ldots, 1 \rangle$, with the fineness parameter N large relative to L.[11] Then F is a discrete distribution defined entirely by the probability masses $\langle \xi(\ell/N) \rangle_{\ell=0}^{N}$, of which exactly $L \ll N$ are nonzero. Recovering the distribution thus amounts to finding a solution $\boldsymbol{\xi}$ of the system

$$\underbrace{\begin{bmatrix} 1 \\ m_1 \\ \vdots \\ m_j \\ \vdots \\ m_n \end{bmatrix}}_{\langle 1, \boldsymbol{m} \rangle} = \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ 0 & 1/N & \cdots & \ell/N & \cdots & 1 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & [1/N]^j & \cdots & [\ell/N]^j & \cdots & 1 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & [1/N]^n & \cdots & [\ell/N]^n & \cdots & 1 \end{bmatrix}}_{\boldsymbol{V}} \underbrace{\begin{bmatrix} \xi(0) \\ \xi(1/N) \\ \vdots \\ \xi(\ell/N) \\ \vdots \\ \xi(1) \end{bmatrix}}_{\boldsymbol{\xi}}, \tag{16}$$

with each component $\xi(\ell/N)$ weakly positive and exactly L components strictly positive. Here $\boldsymbol{V}$ is a Vandermonde matrix with many more columns (i.e., 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 (2011, Theorem 1) applies to precisely this situation, stating that Equation 16 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 $\boldsymbol{p}$.*

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.

3.3.3. *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 the nth approximating density $\hat{f}_n$ as the solution to the optimization problem

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

subject to the (jth-moment) constraint

$$\int_0^1 \rho^j f_n(\rho) \mathrm{d}\rho = m_j \tag{17}$$

[11] For notational simplicity we use an evenly spaced grid of admissible values, but this is not essential for our conclusions.
[12] See, e.g., Macon and Spitzbart (1958) for the definition and properties of Vandermonde matrices.

for $j = 0, 1, \ldots, n$. Mead and Papanicolaou (1984, Theorem 2) show that such a solution exists and is unique;[13] and that for each bounded, continuous $\psi : [0,1] \to \Re$ we have

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

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

$$p(k_A; \hat{F}_n) = p(k_M; \hat{F}_n) = p(k_M; F) = p(k_A; F). \tag{18}$$

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 $\boldsymbol{p} = \boldsymbol{CQm}$ and the shares of the n best alternatives are determined by the first n moments. The constraints in Equation 17 guarantee that these moments coincide for $\hat{F}_n$ and F, yielding the second equality in Equation 18. We summarize our findings as follows.

Proposition 5. *In the consideration probability model, if F admits a density then there exists a map $\boldsymbol{m} \mapsto \hat{F}_n$ 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 $p(k_A; \hat{F}_n) = p(k_A; F)$.*

As already noted, the constraints in Equation 17 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.

3.4. *Unobserved default outcome*

3.4.1. *Conditional choice shares*

In this section we examine the feasibility of cognitive identification when the default outcome is unobserved. Under this assumption our data set consists of the aggregate shares $\overline{p}(k) = p(k)/[1 - p(d)]$ conditional on an active choice being made. Of course, any ratio of aggregate shares of the form $\widetilde{p}(k, \ell) = \overline{p}(k)/\overline{p}(\ell) = p(k)/p(\ell)$ is unaffected by the conditioning, and so Equations 7–8 remain valid when restated in terms of the conditional shares and the associated probability masses $\overline{\pi}(\gamma) = \pi(\gamma)/[1 - \pi(0)]$.

3.4.2. *Recovering $n-1$ probability mass ratios*

As in the basic model, several natural functional forms for the type distribution permit identification of their parameters even when the default outcome is unobserved. (See Appendix A.2 for examples.) In the nonparametric setting, it is simple to adapt Equation 12 to this case. Indeed, for each $\gamma = 2, 3, \ldots, n$ we have that

$$\frac{\pi(\gamma)}{\pi(1)} = \frac{\binom{n}{\gamma}}{n} \sum_{k=n-\gamma+1}^{n} [-1]^{[\gamma-1]-[n-k]} \binom{\gamma-1}{n-k} \frac{p(k)}{p(n)} = \frac{\binom{n}{\gamma}}{n} \sum_{k=n-\gamma+1}^{n} [-1]^{[\gamma-1]-[n-k]} \binom{\gamma-1}{n-k} \frac{\overline{p}(k)}{\overline{p}(n)}.$$

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.

[13] 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 17.

3.4.3. *Consideration probability: Recovering $n-1$ moment ratios*

For the special case of the consideration probability model, Equation 14 can likewise be adapted to the unobserved default scenario. Here, for each $j = 2, 3, \ldots, n$, we have

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

This yields $n-1$ raw-moment ratios, and we could proceed to use methods such as those in Section 3.3 to approximate the shape of the type distribution F (the mean m_1 of which would remain undetermined without knowledge of the default share).[14]

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.[15] 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 $M = \{1, 2, \ldots, n\}$ and write $\varphi : M \to \{1, 2, \ldots, n\}$ for the map that associates each option with its preference rank.[16] 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 $\boldsymbol{P}(h)$ the $n \times n$ permutation matrix corresponding to φ_h.[17] With preference heterogeneity Equation 11 then becomes

$$\boldsymbol{p} = \sum_{h=1}^{n!} \tau_h [\boldsymbol{P}(h)\boldsymbol{C}]\boldsymbol{\pi} = \underbrace{\left[\sum_{h=1}^{n!} \tau_h \boldsymbol{P}(h)\right]}_{\boldsymbol{B}} \boldsymbol{C}\boldsymbol{\pi}, \tag{19}$$

where $\boldsymbol{C}$ is the known, invertible matrix defined in Section 3.2.2.

Equation 19 differs from Equation 11 only in that the right-hand-side vector $\boldsymbol{C}\boldsymbol{\pi}$ is premultiplied by $\boldsymbol{B}$, which we refer to as the "average preference permutation matrix." A typical entry $B_{kr} = \sum_{h:\varphi_h(k)=r} \tau_h$ of this matrix is the total probability of alternative k being placed in position r, computed as the sum of the probabilities of all rankings φ_h that make this assignment. Provided $\boldsymbol{B}$ is invertible, we have $\boldsymbol{\pi} = [\boldsymbol{B}\boldsymbol{C}]^{-1}\boldsymbol{p}$ from Equation 19, and similarly Equation 13 becomes $\boldsymbol{m} = [\boldsymbol{B}\boldsymbol{C}\boldsymbol{Q}]^{-1}\boldsymbol{p}$ for the special case of the consideration probability model. We conclude that in the present context the aggregate choice shares can still be used to find the probability masses in $\boldsymbol{\pi}$ and the raw moments in $\boldsymbol{m}$, as long as the known taste distribution yields a nonsingular $\boldsymbol{B}$.

The matrix $\boldsymbol{B}$ is a convex combination of permutation matrices, and must therefore be bistochastic.[18] Of course there exist taste distributions $\boldsymbol{\tau} = \langle \tau_h \rangle_{h=1}^{n!}$ for which $\boldsymbol{B}$ is noninvertible; e.g., the uniform distribution (with each $\tau_h = 1/n!$) yields a singular $\boldsymbol{B}$ with each entry equal to $1/n$. However, invertibility is clearly the generic situation here. In fact, $\det(\boldsymbol{B})$ is a polynomial function of $\boldsymbol{\tau} \in \Re^{n!}$, and we know that any real-valued polynomial function on a Euclidean space is either identically zero or nonzero almost everywhere (see, e.g., Caron and Traynor 2005). Since $\det(\boldsymbol{B})$ is nonzero for the case of homogeneous

[14] Horan (2019) considers an unobserved default outcome in the context of a dataset with choices from multiple menus, showing that the identification properties of the independent random consideration model in Manzini and Mariotti (2014) remain largely intact.

[15] The 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.

[16] This formulation maintains the assumption of linear order preferences imposed in Section 2.

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

[18] A 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.

preferences, it is not identically zero, and thus $\boldsymbol{B}$ is generically invertible. This allows us to extend Propositions 2–3 as follows.

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 $\boldsymbol{p}$.*

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

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

Example 2 (exploded logit)**.** Let $n = 3$ and $u(k) = \log k$, and suppose that the distribution of tastes is determined by an exploded logit based on u. In this case the average preference permutation matrix is

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

where, for instance, the probability of the ranking φ_2 is calculated as

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

The matrix in Equation 20 is nonsingular (with $\det(\boldsymbol{B}) = -1/30$), whereupon we can compute

$$\begin{bmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \end{bmatrix} = \boldsymbol{\pi} = [\boldsymbol{BC}]^{-1}\boldsymbol{p} = \frac{1}{2}\begin{bmatrix} 30p(1) - 27p(2) + 3p(3) \\ -60p(1) + 75p(2) - 15p(3) \\ 32p(1) - 46p(2) + 14p(3) \end{bmatrix},$$

and as always $\pi(0) = p(d)$.[19] □

4.2. *Unknown taste distribution*

4.2.1. *The multiple occasion framework*

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 3 (identification failure)**.** Let $n = 2$ and $\varphi_1(1) = 1$, so that τ_1 is the probability of the ranking $1 \succ 2$. Equation 19 then takes the form

$$\begin{bmatrix} p(1) \\ p(2) \end{bmatrix} = \boldsymbol{p} = [\boldsymbol{BC}]\boldsymbol{\pi} = \begin{bmatrix} 1/2 & \tau_1 \\ 1/2 & \tau_2 \end{bmatrix}\begin{bmatrix} \pi(1) \\ \pi(2) \end{bmatrix},$$

an underdetermined system in which the cognitive distribution $\langle \pi(1), \pi(2) \rangle$ and the taste distribution $\tau_1 = 1 - \tau_2$ cannot be disambiguated. □

[19] Note that the invertibility of $\boldsymbol{B}$ in this example is not an accident. For any $u : M \to \Re$ it can be shown that

$$\det(\boldsymbol{B}) = \frac{e^{u(1)} - e^{u(2)}}{e^{u(1)} + e^{u(2)}} \times \frac{e^{u(1)} - e^{u(3)}}{e^{u(1)} + e^{u(3)}} \times \frac{e^{u(2)} - e^{u(3)}}{e^{u(2)} + e^{u(3)}},$$

and hence $\det(\boldsymbol{B}) \neq 0$ if and only if the function u is one-to-one.

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.[20] 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 notational simplicity 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 such joint choice shares 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 choice patterns.[21]

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 distribution even in the context of unknown and possibly changing tastes. We shall find (in Proposition 8) that joint choice share data from as few as three occasions is generically sufficient 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 constant (see Footnote 26). The taste distribution on occasion i is denoted by $\boldsymbol{\tau}_i = \langle \tau_{ih} \rangle_{h=1}^{n!}$, 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 $p(k_1 \cdots k_I) = \int_\Theta p_\theta(k_1 \cdots k_I) \mathrm{d}F$, and as before our objective is to use this data to deduce information about the underlying cognitive distribution F.

4.2.2. *Joint choice shares in the consideration capacity model*

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 type γ. The analog of Equation 1 is then

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

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 2 appears as

$$p(k_1 \cdots k_I) = \sum_{\gamma=1}^{n} \pi(\gamma) p_\gamma(k_1 \cdots k_I) = \sum_{\gamma=1}^{n} \pi(\gamma) \prod_{i=1}^{I} \sum_{r=1}^{n-\gamma+1} \frac{\binom{n-r}{\gamma-1}}{\binom{n}{\gamma}} \sum_{h:\varphi_h(k_i)=r} \tau_{ih}, \tag{21}$$

demonstrating how the population choice shares are determined by the cognitive type distribution $\boldsymbol{\pi}$ in conjunction with the taste distributions $\langle \boldsymbol{\tau}_i \rangle_{i=1}^{I}$.

[20] One strategy would be to supply the researcher with aggregate data on choices from multiple subsets of the menu (cf. Aguiar et al. 2018 and Geng and Ozbay 2018), 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.

[21] With 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.

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

Example 4 (three occasions)**.** Let $n = 2$ and $I = 3$, and for each occasion i let $\tau_{i1} \in (0, 1)$ denote the probability of $1_i \succ_i 2_i$. In other words, on each occasion we let the first ranking be the one that prefers the first alternative (according to the arbitrary initial ordering of M) to the second. Conditioning on $\gamma > 0$, as in Section 3.4, let $\boldsymbol{S}$ denote the $2 \times 2 \times 2$ array describing the joint distribution of choices on the three occasions. This array can be represented explicitly as

$$\boldsymbol{S} = \left[\!\!\left[\begin{array}{c|cc||c|cc} k_3 = 1 & k_2 = 1 & k_2 = 2 & k_3 = 2 & k_2 = 1 & k_2 = 2 \\ \hline k_1 = 1 & \overline{p}(111) & \overline{p}(121) & k_1 = 1 & \overline{p}(112) & \overline{p}(122) \\ k_1 = 2 & \overline{p}(211) & \overline{p}(221) & k_1 = 2 & \overline{p}(212) & \overline{p}(222) \end{array} \right]\!\!\right]. \tag{22}$$

Our goal is to use the eight joint choice shares $\overline{p}(k_1k_2k_3)$ to deduce both the cognitive distribution $\overline{\pi}(1)$ and the three occasion-specific taste distributions τ_{i1}.[22]

In Equation 22, consider the top row (associated with $k_1 = 1$) of each 2×2 subarray. Agents of type $\gamma = 1$ choose the two options with equal probability on each occasion, and so $p_1(111) = p_1(121) = p_1(112) = p_1(122) = 1/8$. In contrast, agents of type $\gamma \geq 2$ choose alternative 1 with probability τ_{i1} and alternative 2 with probability $\tau_{i2} = 1-\tau_{i1}$ on occasion i, so we obtain the expressions $p_2(111) = \tau_{11}\tau_{21}\tau_{31}$, $p_2(121) = \tau_{11}\tau_{22}\tau_{31}$, $p_2(112) = \tau_{11}\tau_{21}\tau_{32}$, and $p_2(122) = \tau_{11}\tau_{22}\tau_{32}$. Equation 21 can then be specialized to each of these four joint choice shares as

$$\overline{p}(111) = \overline{\pi}(1) \cdot [1/8] + \overline{\pi}(2) \cdot \tau_{11}\tau_{21}\tau_{31}, \tag{23}$$
$$\overline{p}(121) = \overline{\pi}(1) \cdot [1/8] + \overline{\pi}(2) \cdot \tau_{11}\tau_{22}\tau_{31}, \tag{24}$$
$$\overline{p}(112) = \overline{\pi}(1) \cdot [1/8] + \overline{\pi}(2) \cdot \tau_{11}\tau_{21}\tau_{32}, \tag{25}$$
$$\overline{p}(122) = \overline{\pi}(1) \cdot [1/8] + \overline{\pi}(2) \cdot \tau_{11}\tau_{22}\tau_{32}. \tag{26}$$

To recover the cognitive distribution, we combine Equations 23–26 to establish that

$$\frac{8\overline{p}(111) - \overline{\pi}(1)}{8\overline{p}(112) - \overline{\pi}(1)} = \frac{\tau_{31}}{\tau_{32}} = \frac{8\overline{p}(121) - \overline{\pi}(1)}{8\overline{p}(122) - \overline{\pi}(1)}.$$

We can then solve for

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

and of course $\overline{\pi}(2) = 1 - \overline{\pi}(1)$ since $n = 2$.

Next, to recover the taste distributions, we define the univariate marginals

$$G_1(k) = \overline{p}(k11) + \overline{p}(k12) + \overline{p}(k21) + \overline{p}(k22),$$
$$G_2(k) = \overline{p}(1k1) + \overline{p}(1k2) + \overline{p}(2k1) + \overline{p}(2k2),$$
$$G_3(k) = \overline{p}(11k) + \overline{p}(12k) + \overline{p}(21k) + \overline{p}(22k).$$

Adding Equations 23–26 yields $G_1(1) = \overline{\pi}(1) \cdot [1/2] + \overline{\pi}(2) \cdot \tau_{11}$, and more generally for each occasion i we have $G_i(1) = \overline{\pi}(1) \cdot [1/2] + \overline{\pi}(2) \cdot \tau_{i1}$. We can then express each

$$\tau_{i1} = \frac{2G_i(1) - \overline{\pi}(1)}{2\overline{\pi}(2)}$$

in terms of known functions of the array $\boldsymbol{S}$, as desired. □

[22] Here $\overline{\pi}(\gamma) = \pi(\gamma)/[1 - \pi(0)]$, as before, and we define $\overline{p}(k_1 \cdots k_I) = p(k_1 \cdots k_I)/[1 - p(d \cdots d)]$ analogously with our notation for the single-occasion setting.

4.2.3. *Cognitive identification from three occasions*

We proceed now to establish that the cognitive identification seen in Example 4 is a generic feature of the multi-occasion setting. Once again conditioning on the event $\gamma > 0$, we represent our dataset as a tensor $\boldsymbol{S}$ of order I with dimensions $n \times \cdots \times n$,[23] having typical entry

$$S_{k_1,\ldots,k_I} = \overline{p}(k_1 \cdots k_I) = \frac{p(k_1 \cdots k_I)}{1 - p(d \cdots d)}.$$

Writing $\boldsymbol{B}_i = \sum_{h=1}^{n!} \tau_{ih} \boldsymbol{P}(h)$ for the average preference permutation matrix on occasion i, we can then express Equation 21 more compactly as

$$\boldsymbol{S} = \sum_{\gamma=1}^{n} \overline{\pi}(\gamma) \otimes_{i=1}^{I} [\boldsymbol{B}_i \boldsymbol{C}] \mathbf{1}_\gamma, \tag{27}$$

where $\otimes$ is the outer product operator and $\mathbf{1}_\gamma$ denotes the unit vector for component γ (which here extracts the γth column of the matrix $\boldsymbol{B}_i \boldsymbol{C}$).[24] To illustrate this notation, we pause to revisit the preceding example.

Example 5 (three occasions; continued from Example 4)**.** For each occasion $i = 1, 2, 3$ we have

$$\boldsymbol{B}_i \boldsymbol{C} = \begin{bmatrix} 1/2 & \tau_{i1} \\ 1/2 & \tau_{i2} \end{bmatrix},$$

and so Equation 27 takes the form

$$\begin{aligned} \boldsymbol{S} = \overline{\pi}(1) \otimes_{i=1}^{3} \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix} + \overline{\pi}(2) \otimes_{i=1}^{3} \begin{bmatrix} \tau_{i1} \\ \tau_{i2} \end{bmatrix} &= \overline{\pi}(1) \left[\!\!\left[\begin{array}{cc||cc} 1/8 & 1/8 & 1/8 & 1/8 \\ 1/8 & 1/8 & 1/8 & 1/8 \end{array} \right]\!\!\right] \cdots \\ &\cdots + \overline{\pi}(2) \left[\!\!\left[\begin{array}{cc||cc} \tau_{11}\tau_{21}\tau_{31} & \tau_{11}\tau_{22}\tau_{31} & \tau_{11}\tau_{21}\tau_{32} & \tau_{11}\tau_{22}\tau_{32} \\ \tau_{12}\tau_{21}\tau_{31} & \tau_{12}\tau_{22}\tau_{31} & \tau_{12}\tau_{21}\tau_{32} & \tau_{12}\tau_{22}\tau_{32} \end{array} \right]\!\!\right]. \end{aligned} \tag{28}$$

Here $[\boldsymbol{B}_i \boldsymbol{C}]\mathbf{1}_1 = \langle 1/2, 1/2 \rangle$ is the choice distribution of cognitive type $\gamma = 1$ on occasion i; namely, uniform randomization between the two alternatives. Likewise, $[\boldsymbol{B}_i \boldsymbol{C}]\mathbf{1}_2 = \langle \tau_{i1}, \tau_{i2} \rangle$ is the choice distribution of type $\gamma = 2$, which simply reproduces the taste distribution on occasion i since the full-consideration type always chooses optimally. Note further that Equation 28 contains the four joint choice shares in Equations 23–26, as expected.

Finally—anticipating our formal result for the multi-occasion setting—we can bring the capacity probabilities $\overline{\pi}(\gamma)$ into the outer products by defining matrices

$$\boldsymbol{Z}_1 = [\boldsymbol{B}_1 \boldsymbol{C}] \begin{bmatrix} \overline{\pi}(1) & 0 \\ 0 & \overline{\pi}(2) \end{bmatrix},$$

$\boldsymbol{Z}_2 = \boldsymbol{B}_2 \boldsymbol{C}$, and $\boldsymbol{Z}_3 = \boldsymbol{B}_3 \boldsymbol{C}$; and by writing Equation 28 as $\boldsymbol{S} = \otimes_{i=1}^{3} \boldsymbol{Z}_i \mathbf{1}_1 + \otimes_{i=1}^{3} \boldsymbol{Z}_i \mathbf{1}_2$. □

As illustrated in Example 5, Equation 27 decomposes the joint choice share tensor $\boldsymbol{S}$ into a linear combination of n rank-1 tensors.[25] The advantage of this representation is that the uniqueness properties of such decompositions have been studied extensively, with Kruskal (1977, Theorem 4a) supplying a fundamental theorem that has been further refined by Sidiropoulos and Bro (2000) and Allman et al. (2009), among others. We shall use a corollary of the theorem due to Rhodes (2010), adapted for our setting as follows.

[23] A 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.

[24] Recall 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.

[25] A tensor is said to be of rank 1 if it is an outer product of vectors. See Supplemental Material Appendix B (Dardanoni et al. 2020) for a primer on tensor decompositions of the sort studied in Section 4.2.

Lemma 1 (Kruskal; Rhodes)**.** *Given any triad* $\langle \boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3 \rangle$ *of invertible* $n \times n$ *matrices, the tensor* $\boldsymbol{T} = \sum_{\gamma=1}^{n} [\boldsymbol{Z}_1 \mathbf{1}_\gamma \otimes \boldsymbol{Z}_2 \mathbf{1}_\gamma \otimes \boldsymbol{Z}_3 \mathbf{1}_\gamma]$ *uniquely determines each* $\boldsymbol{Z}_i$ *up to column rescaling and permutation. That is, for any* $\langle \hat{\boldsymbol{Z}}_1, \hat{\boldsymbol{Z}}_2, \hat{\boldsymbol{Z}}_3 \rangle$ *such that* $\sum_{\gamma=1}^{n} [\hat{\boldsymbol{Z}}_1 \mathbf{1}_\gamma \otimes \hat{\boldsymbol{Z}}_2 \mathbf{1}_\gamma \otimes \hat{\boldsymbol{Z}}_3 \mathbf{1}_\gamma] = \boldsymbol{T}$*, there exist invertible diagonal matrices* $\langle \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_3 \rangle$ *and a permutation matrix* $\boldsymbol{P}$ *such that* $\boldsymbol{D}_1 \boldsymbol{D}_2 \boldsymbol{D}_3 = \boldsymbol{I}_n$ *and each* $\hat{\boldsymbol{Z}}_i = \boldsymbol{Z}_i \boldsymbol{D}_i \boldsymbol{P}$*; where* $\boldsymbol{I}_n$ *is the* $n \times n$ *identity matrix.*[26]

Setting $I = 3$ and applying Lemma 1 to the tensor $\boldsymbol{S}$, we can show generic cognitive identification in the multi-occasion environment.[27]

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

Proof. Write $\boldsymbol{D}(\overline{\boldsymbol{\pi}})$ for the diagonal matrix with entries $\overline{\boldsymbol{\pi}} = \langle \overline{\pi}(\gamma) \rangle_{\gamma=1}^{n} \gg 0$. Following Allman et al. (2009, p. 3118) and Example 5 above, set $\boldsymbol{Z}_1 = [\boldsymbol{B}_1 \boldsymbol{C}] \boldsymbol{D}(\overline{\boldsymbol{\pi}})$, $\boldsymbol{Z}_2 = \boldsymbol{B}_2 \boldsymbol{C}$, and $\boldsymbol{Z}_3 = \boldsymbol{B}_3 \boldsymbol{C}$. We then have

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

As argued in connection with Proposition 6, each matrix $\boldsymbol{B}_i \boldsymbol{C}$ has full rank for almost all distributions $\boldsymbol{\tau}_i$, and since $\overline{\boldsymbol{\pi}} \gg 0$ it follows that each $\boldsymbol{Z}_i$ is generically invertible.

Suppose now that there exists a set of duplicate parameters $\langle \hat{\boldsymbol{B}}_1, \hat{\boldsymbol{B}}_2, \hat{\boldsymbol{B}}_3 \rangle$ and $\hat{\overline{\boldsymbol{\pi}}} \gg 0$ such that the corresponding $\sum_{\gamma=1}^{n} [\hat{\boldsymbol{Z}}_1 \mathbf{1}_\gamma \otimes \hat{\boldsymbol{Z}}_2 \mathbf{1}_\gamma \otimes \hat{\boldsymbol{Z}}_3 \mathbf{1}_\gamma] = \boldsymbol{S}$. By Lemma 1 there then exist rescalings $\langle \boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_3 \rangle$ and a permutation $\boldsymbol{P}$ such that

$$[\hat{\boldsymbol{B}}_1 \boldsymbol{C}] \boldsymbol{D}(\hat{\overline{\boldsymbol{\pi}}}) = \hat{\boldsymbol{Z}}_1 = \boldsymbol{Z}_1 \boldsymbol{D}_1 \boldsymbol{P} = [\boldsymbol{B}_1 \boldsymbol{C}] \boldsymbol{D}(\overline{\boldsymbol{\pi}}) \boldsymbol{D}_1 \boldsymbol{P}, \tag{29}$$

$$\hat{\boldsymbol{B}}_2 \boldsymbol{C} = \hat{\boldsymbol{Z}}_2 = \boldsymbol{Z}_2 \boldsymbol{D}_2 \boldsymbol{P} = [\boldsymbol{B}_2 \boldsymbol{C}] \boldsymbol{D}_2 \boldsymbol{P}, \tag{30}$$

$$\hat{\boldsymbol{B}}_3 \boldsymbol{C} = \hat{\boldsymbol{Z}}_3 = \boldsymbol{Z}_3 \boldsymbol{D}_3 \boldsymbol{P} = [\boldsymbol{B}_3 \boldsymbol{C}] \boldsymbol{D}_3 \boldsymbol{P}. \tag{31}$$

Writing $\mathbf{1}$ for the vector of ones (and $\mathbf{1}^\top$ for its transpose), from Equations 30–31 we have $\mathbf{1}^\top \hat{\boldsymbol{B}}_i \boldsymbol{C} = \mathbf{1}^\top [\boldsymbol{B}_i \boldsymbol{C}] \boldsymbol{D}_i \boldsymbol{P}$ for each $i = 2, 3$. Since $\boldsymbol{B}_i$ and $\hat{\boldsymbol{B}}_i$ are bistochastic and $\boldsymbol{C}$ is left stochastic, it follows that $\mathbf{1}^\top = \mathbf{1}^\top \boldsymbol{D}_i \boldsymbol{P}$ and thus $\mathbf{1}^\top = \mathbf{1}^\top \boldsymbol{P}^\top = \mathbf{1}^\top \boldsymbol{D}_i$. We conclude that $\boldsymbol{D}_2 = \boldsymbol{D}_3 = \boldsymbol{I}_n$, and therefore $\boldsymbol{D}_1 = [\boldsymbol{D}_2 \boldsymbol{D}_3]^{-1} = \boldsymbol{I}_n$ as well.[28]

Similarly, we have $\hat{\overline{\boldsymbol{\pi}}}^\top = \mathbf{1}^\top \boldsymbol{D}(\hat{\overline{\boldsymbol{\pi}}}) = \mathbf{1}^\top \boldsymbol{D}(\overline{\boldsymbol{\pi}}) \boldsymbol{P} = \overline{\boldsymbol{\pi}}^\top \boldsymbol{P}$ from Equation 29 and hence $\hat{\overline{\boldsymbol{\pi}}} = \boldsymbol{P}^\top \overline{\boldsymbol{\pi}}$. It follows that $\boldsymbol{D}(\hat{\overline{\boldsymbol{\pi}}}) = \boldsymbol{D}(\boldsymbol{P}^\top \overline{\boldsymbol{\pi}}) = \boldsymbol{P}^\top \boldsymbol{D}(\overline{\boldsymbol{\pi}}) \boldsymbol{P}$, so that Equation 29 yields $[\hat{\boldsymbol{B}}_1 \boldsymbol{C}] \boldsymbol{P}^\top \boldsymbol{D}(\overline{\boldsymbol{\pi}}) \boldsymbol{P} = [\boldsymbol{B}_1 \boldsymbol{C}] \boldsymbol{D}(\overline{\boldsymbol{\pi}}) \boldsymbol{P}$ and $[\hat{\boldsymbol{B}}_1 \boldsymbol{C}] \boldsymbol{P}^\top = \boldsymbol{B}_1 \boldsymbol{C}$. Together with Equations 30–31, this shows that $\hat{\boldsymbol{B}}_i \boldsymbol{C} = [\boldsymbol{B}_i \boldsymbol{C}] \boldsymbol{P}$ for all $i = 1, 2, 3$. The duplicate parameters $\langle \hat{\boldsymbol{B}}_1, \hat{\boldsymbol{B}}_2, \hat{\boldsymbol{B}}_3 \rangle$ and $\hat{\overline{\boldsymbol{\pi}}}$ are thus seen to result from label swapping; i.e., a garbling of the cognitive type distribution $\overline{\boldsymbol{\pi}}$ via the permutation $\boldsymbol{P}^\top$. This garbling is reversed by swapping labels in the $\boldsymbol{B}_i \boldsymbol{C}$ matrices, carried out by the transformations $\hat{\boldsymbol{B}}_i = \boldsymbol{B}_i [\boldsymbol{C} \boldsymbol{P} \boldsymbol{C}^{-1}]$. When labels are assigned correctly, we have $\boldsymbol{P} = \boldsymbol{I}_n$, $\hat{\overline{\boldsymbol{\pi}}} = \overline{\boldsymbol{\pi}}$, and each $\hat{\boldsymbol{B}}_i = \boldsymbol{B}_i$, as desired. □

[26] The result in Rhodes (2010, Corollary 2) is in fact substantially more general than this, since it allows the $\boldsymbol{Z}_i$ matrices to have different numbers of rows and one of them to have linearly dependent columns. This necessitates a restriction on the "Kruskal rank" (Rhodes 2010, p. 1819) of the latter matrix—a hypothesis that is trivially satisfied in the square, full-rank case. In view of Example 4, different numbers of rows in the $\boldsymbol{Z}_i$ matrices will correspond to different numbers of alternatives across the choice occasions, and so relaxing the assumption of constant menu cardinality is within the scope of our approach to cognitive identification. We do not pursue this extension at present, since it is tangential to our main purpose and would complicate our notation considerably.

[27] Since our goal is to infer cognitive heterogeneity from minimal data, we limit our agent's choices to the three occasions needed to apply Lemma 1. Data from additional occasions will neither help nor hinder cognitive identification in this context, since extensions of Kruskal's theorem to tensors of order higher than three are available (see, e.g., Sidiropoulos and Bro 2000, Theorem 3). On the other hand, two occasions are inadequate, with Allman et al. (2009, p. 3108) noting that "[t]his nonidentifiability is intimately related to the nonuniqueness of certain matrix factorizations." (See also Kruskal 1977, p. 122.)

[28] In other words, while Lemma 1 allows for rescaling of columns, our framework rules this out.

4.2.4. *Partial preference identification*

In connection with Proposition 8, it is important to note that joint choice share data do not fully pin down the taste distributions $\langle \boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3 \rangle$. On the contrary, factorial explosion of the number of rankings of n alternatives makes it clear that such identification cannot be possible in general. In relation to tastes, what the joint choice shares determine are the average preference permutation matrices $\langle \boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3 \rangle$. Recall that, for each occasion i, these matrices record the overall probability of each alternative k_i being assigned each rank position r. The kth entry in the rth column of $\boldsymbol{B}_i$ is given by $\sum_{h:\varphi_h(k_i)=r} \tau_{ih}$, which is the total probability of all preference rankings that make this assignment. Hence preferences are less than fully identified only to the extent that the taste distribution can be changed without affecting these rank-position probabilities.

For instance, when $n = 3$ we have six preference orders, resulting in average preference permutation matrices of the form

$$\boldsymbol{B}_i = \tau_{i1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \tau_{i2} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} + \tau_{i3} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \tau_{i4} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \cdots$$
$$\cdots + \tau_{i5} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} + \tau_{i6} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \tau_{i1} + \tau_{i2} & \tau_{i3} + \tau_{i4} & \tau_{i5} + \tau_{i6} \\ \tau_{i3} + \tau_{i5} & \tau_{i1} + \tau_{i6} & \tau_{i2} + \tau_{i4} \\ \tau_{i4} + \tau_{i6} & \tau_{i2} + \tau_{i5} & \tau_{i1} + \tau_{i3} \end{bmatrix}.$$

Here the perturbed taste distribution $\hat{\boldsymbol{\tau}}_i = \boldsymbol{\tau}_i + \epsilon\langle 1, -1, -1, 1, 1, -1 \rangle$ yields $\hat{\boldsymbol{B}}_i = \boldsymbol{B}_i$, an unchanged average preference permutation matrix, and it follows that $\boldsymbol{\tau}_i$ and $\hat{\boldsymbol{\tau}}_i$ cannot be distinguished using our methods.

Proposition 8 achieves identification of the cognitive distribution and the average preference permutation matrices with no parametric assumptions on the primitives of the consideration capacity model. Introducing such assumptions may enable us to refine our conclusions about $\langle \boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\tau}_3 \rangle$ beyond the rank-position probabilities recorded in $\langle \boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3 \rangle$, a task that is greatly simplified by knowledge of F. In fact, since the type-conditional choice distributions have already been recovered, we could 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. We could, for example, assume that the type-conditional choices result from a random utility model (RUM) with a given error distribution, or by a single-crossing RUM as defined in Apesteguia et al. (2017). In any event, such parametric assumptions are unrelated to the limited-attention aspects of our model and unnecessary to achieve our primary goal in this section, which is to identify the *cognitive* distribution F in the presence of unknown taste heterogeneity.

5. CONCLUSION

The main contribution of this paper is to show how aggregate choice shares can identify the distribution of cognitive characteristics in a population of agents who exhibit limited attention. A central advantage of our approach is that it requires minimal data: With homogeneous (or known) tastes we use choices shares from a single menu, and with heterogeneous (and unknown) tastes we use joint choice shares from three "occasions." In this respect our methodology differs from prior theoretical work on bounded rationality, much of which uses individual choice data from a rich family of overlapping menus. At the same time, it contrasts with the more applied, econometrically oriented literature on this topic, where identification is typically facilitated by the presence of observable covariates—an assumption that we deliberately eschew. Notwithstanding the parsimonious nature of our datasets, we find that aggregate choice shares can encode substantial information about the distribution of attention characteristics in the population. In particular, they can reveal the distribution of the consideration capacity γ up to the cardinality of the menu, and from this we can deduce the same number of raw moments of ρ in the special case of the consideration probability model.

There are numerous ways that we could build upon the work reported in this paper (some of which have been mentioned in passing; see Footnotes 20 and 26). We could, for instance, attempt to tighten the identification of preferences in the multiple occasion environment by postulating access to additional data. Alternatively, we could seek to bring other forms of bounded rationality—such as framing effects

or satisficing—into the present setting. A third variety of extension would be to weaken the assumptions needed to derive our main results, and in order to illustrate this possibility let us return briefly to the assumption of conditionally uniform consideration sets.

For $n = 3$, Example 1 outlines a generalized consideration capacity model in which salience weights $\boldsymbol{w} = \langle w_1, w_2, w_3 \rangle$ for the alternatives affect the relative probabilities of different consideration sets of the same size. Allowing for preference heterogeneity, a typical ranking φ_h permutes the salience weights to $\boldsymbol{P}(h)\boldsymbol{w}$, and the transition from $\boldsymbol{\pi}$ to $\boldsymbol{p}$ is then governed by the matrix $\boldsymbol{P}(h)\boldsymbol{C}(\boldsymbol{P}(h)\boldsymbol{w})$. Averaging over rankings yields the full transition matrix $\boldsymbol{T} = \sum_{h=1}^{6} \tau_h [\boldsymbol{P}(h)\boldsymbol{C}(\boldsymbol{P}(h)\boldsymbol{w})]$, and the analog of Equation 19 for this scenario can be written explicitly as

$$\underbrace{\begin{bmatrix} p(1) \\ p(2) \\ p(3) \end{bmatrix}}_{\boldsymbol{p}} = \underbrace{\begin{bmatrix} w_1/W_1 & [\tau_1+\tau_2+\tau_5]w_1w_2/W_2 + [\tau_1+\tau_2+\tau_3]w_1w_3/W_2 & \tau_1+\tau_2 \\ w_2/W_1 & [\tau_3+\tau_4+\tau_6]w_1w_2/W_2 + [\tau_1+\tau_3+\tau_4]w_2w_3/W_2 & \tau_3+\tau_4 \\ w_3/W_1 & [\tau_4+\tau_5+\tau_6]w_1w_3/W_2 + [\tau_2+\tau_5+\tau_6]w_2w_3/W_2 & \tau_5+\tau_6 \end{bmatrix}}_{\boldsymbol{T}} \underbrace{\begin{bmatrix} \pi(1) \\ \pi(2) \\ \pi(3) \end{bmatrix}}_{\boldsymbol{\pi}}.$$

Extending this generalized model to the multi-occasion setting, with $I = 3$ and

$$\boldsymbol{T}_i = \begin{bmatrix} w_1/W_1 & [\tau_{i1}+\tau_{i2}+\tau_{i5}]w_1w_2/W_2 + [\tau_{i1}+\tau_{i2}+\tau_{i3}]w_1w_3/W_2 & \tau_{i1}+\tau_{i2} \\ w_2/W_1 & [\tau_{i3}+\tau_{i4}+\tau_{i6}]w_1w_2/W_2 + [\tau_{i1}+\tau_{i3}+\tau_{i4}]w_2w_3/W_2 & \tau_{i3}+\tau_{i4} \\ w_3/W_1 & [\tau_{i4}+\tau_{i5}+\tau_{i6}]w_1w_3/W_2 + [\tau_{i2}+\tau_{i5}+\tau_{i6}]w_2w_3/W_2 & \tau_{i5}+\tau_{i6} \end{bmatrix},$$

Equation 27 becomes $\boldsymbol{S} = \sum_{\gamma=1}^{3} \overline{\pi}(\gamma) \otimes_{i=1}^{3} \boldsymbol{T}_i \boldsymbol{1}_\gamma$. As previously, we can use Lemma 1 to show that the probability masses $\boldsymbol{\pi}$ and transition matrices $\langle \boldsymbol{T}_1, \boldsymbol{T}_2, \boldsymbol{T}_3 \rangle$ are uniquely determined by the joint choice shares $p(k_1k_2k_3)$. The attention distribution therefore continues to be identified by our multi-occasion dataset, as are the relative salience weights (supplied by the first column of $\boldsymbol{T}_i$) when these are taken to be unknown to the researcher. In fact, all of this would remain true even if the salience weights were to depend on the occasion—a useful formulation if, for instance, each weight was defined as a function of characteristics of the corresponding alternative. Carrying through such a broad generalization of our attention model goes well beyond the scope of the present paper. However, its apparent feasibility serves to emphasize that our results on cognitive identification are robust to quite substantial extensions of the framework, including the sort of modifications that may be needed to use our contribution as the basis for fully fledged empirical applications.

APPENDIX A: PARAMETRIC IDENTIFICATION

A.1. *Basic models*

Both for the consideration capacity model and for the special case of the consideration probability model, we consider simple one- and two-parameter functional forms for F.

Example 6 (Poisson γ)**.** For $\mu > 0$, let the consideration capacity γ have the Poisson distribution $\pi(\gamma) = [\mu^\gamma/\gamma!]e^{-\mu}$ for $0 \leq \gamma < n$. In this case Equation 3 yields default share $p(d) = \pi(0) = e^{-\mu}$, and thus $\mu = -\log p(d)$. Alternatively, Equation 8 yields

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

and so $\mu = [n-1][p(n-1)/p(n) - 1]$. □

Example 7 (Pascal γ)**.** For $r \in \{1, 2, 3, \ldots\}$ 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 8 then yields

$$\frac{n-1}{2}\left[\frac{p(n-1)}{p(n)} - 1\right] = \frac{\pi(2)}{\pi(1)} = \frac{q[r+1]}{2}. \tag{32}$$

We have also

$$\frac{np(n)}{p(d)} = \frac{\pi(1)}{\pi(0)} = qr, \tag{33}$$

and Equations 32–33 can be solved simultaneously for the parameters

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

Example 8 (uniform ρ)**.** For $\rho_{\min} \in [0,1)$, let the consideration probability ρ be distributed uniformly on $[\rho_{\min}, 1]$. Since $F(\rho) = [\rho - \rho_{\min}]/[1 - \rho_{\min}]$, Equation 10 becomes

$$p(k) = \frac{1}{1-\rho_{\min}} \int_{\rho_{\min}}^{1} \rho[1-\rho]^{k-1} \mathrm{d}\rho. \tag{34}$$

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

Example 9 (Beta ρ)**.** For $a, b > 0$, let the consideration probability have the Beta distribution $F(\rho) = B(a,b)^{-1} \int_0^\rho t^{a-1}[1-t]^{b-1}\mathrm{d}t$ (where B is the Beta function). Here Equation 10 appears as

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

The first two choice shares are

$$p(1) = \frac{B(a+1,b)}{B(a,b)} = \frac{a}{a+b}, \tag{35}$$
$$p(2) = \frac{B(a+1,b+1)}{B(a,b)} = \frac{ab}{[a+b][a+b+1]}; \tag{36}$$

and we can solve for the parameters

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

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 alternatives. This mirrors our elicitation procedure in Section 3.2, where each mass $\pi(\gamma)$ is seen to depend on the choice shares of a group of sufficiently unattractive options (cf., Equation 12), and each moment of the ρ-distribution is seen to depend on the shares of a sufficiently attractive group (cf., Equation 14).

A.2. *Unobserved default*

Here we adapt each of the parametric examples in Section A.1 to the unobserved default scenario.

Example 10 (Poisson γ; continued from Example 6)**.** Here $\mu = [n-1][\widetilde{p}(n-1,n) - 1]$, as above. $\square$

Example 11 (Pascal γ; continued from Example 7)**.** Equation 32 can be written as $\widetilde{p}(n-1,n) = q[r+1]/[n-1] + 1$, and similarly from Equation 7 we obtain

$$\widetilde{p}(n-2,n) - \widetilde{p}(n-1,n) = \frac{2}{n-1}\left[\frac{\overline{\pi}(2)}{\overline{\pi}(1)} + \frac{3}{n-2}\frac{\overline{\pi}(3)}{\overline{\pi}(1)}\right] = \frac{q[r+1]}{n-1}\left[1 + \frac{q[r+2]}{n-2}\right].$$

These equations can be solved simultaneously for the parameters

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

Example 12 (uniform ρ; continued from Example 8)**.** From Equation 34 we have both $p(1) = [1+\rho_{\min}]/2$ and $p(2) = [2\rho_{\min}+1][1-\rho_{\min}]/6$. Hence

$$\widetilde{p}(1,2) = \frac{3[1+\rho_{\min}]}{[2\rho_{\min}+1][1-\rho_{\min}]},$$

and it follows that

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

Example 13 (Beta ρ; continued from Example 9)**.** Equations 35–36 yield $\widetilde{p}(2,1) = b/[a+b+1]$, and likewise we can compute $\widetilde{p}(3,2) = [b+1]/[a+b+2]$. Solving for the parameters, we obtain

$$a = \frac{1 - 2\widetilde{p}(3,2) + \widetilde{p}(3,1)}{\widetilde{p}(3,2) - \widetilde{p}(2,1)},$$
$$b = \frac{\widetilde{p}(2,1)[1-\widetilde{p}(3,2)]}{\widetilde{p}(3,2) - \widetilde{p}(2,1)}. \square$$

REFERENCES

Abaluck, J. and A. Adams (2017): "What do consumers consider before they choose? Identification from asymmetric demand responses," NBER Working Paper 23566.

Aguiar, V. H., M. J. Boccardi, N. Kashaev, and J. Kim (2018): "Does random consideration explain behavior when choice is hard? Evidence from a large-scale experiment," Working paper (arXiv:1812.09619).

Allman, E. S., C. Matias, and J. A. Rhodes (2009): "Identifiability of parameters in latent structure models with many observed variables," *Annals of Statistics*, 37, 3099–3132.

Apesteguia, J. and M. A. Ballester (2013): "Choice by sequential procedures," *Games and Economic Behavior*, 77, 90–99.

Apesteguia, J., M. A. Ballester, and J. Lu (2017): "Single-crossing random utility models," *Econometrica*, 85, 661–674.

Baigent, N. and W. Gaertner (1996): "Never choose the uniquely largest: A characterization," *Economic Theory*, 8, 239–249.

Barseghyan, L., M. Coughlin, F. Molinari, and J. C. Teitelbaum (2019a): "Heterogeneous choice sets and preferences," Working paper (arXiv:1907.02337).

Barseghyan, L., F. Molinari, and M. Thirkettle (2019b): "Discrete choice under risk with limited consideration," Centre for Microdata Methods and Practice CWP08/19.

Caplin, A. and M. Dean (2015): "Revealed preference, rational inattention, and costly information acquisition," *American Economic Review*, 105, 2183–2203.

Caplin, A., M. Dean, and D. Martin (2011): "Search and satisficing," *American Economic Review*, 101, 2899–2922.

Caron, R. and T. Traynor (2005): "The zero set of a polynomial," Working paper, Department of Mathematics and Statistics, University of Windsor.

Cattaneo, M. D., X. Ma, Y. Masatlioglu, and E. Suleymanov (2017): "A random attention model," Working paper (arXiv:1712.03448).

Cherepanov, V., T. Feddersen, and A. Sandroni (2013): "Rationalization," *Theoretical Economics*, 8, 775–800.

Cohen, A. and A. Yeredor (2011): "On the use of sparsity for recovering discrete probability distributions from their moments," in *Proceedings of the 2011 IEEE Statistical Signal Processing Workshop*, 753–756.

Crawford, G. S., R. Griffith, and A. Iaria (2020): "A survey of preference estimation with unobserved choice set heterogeneity," Working paper, Department of Economics, University of Manchester.

Dardanoni, V., P. Manzini, M. Mariotti, and C. J. Tyson (2020): "Supplement to 'Inferring cognitive heterogeneity from aggregate choices'," *Econometrica Supplemental Material*.

de Clippel, G., K. Eliaz, and K. Rozen (2014): "Competing for consumer inattention," *Journal of Political Economy*, 122, 1203–1234.

de Oliveira, H., T. Denti, M. Mihm, and K. Ozbek (2017): "Rationally inattentive preferences and hidden information costs," *Theoretical Economics*, 12, 621–654.

Echenique, F., K. Saito, and G. Tserenjigmid (2018): "The perception-adjusted Luce model," *Mathematical Social Sciences*, 93, 67–76.

Gaynor, M., C. Propper, and S. Seiler (2016): "Free to choose? Reform, choice, and consideration sets in the English National Health Service," *American Economic Review*, 106, 3521–3557.

Geng, S. and E. Y. Ozbay (2018): "Choice with limited capacity," Working paper, Wang Yanan Institute for Studies in Economics, Xiamen University.

Gibbard, P. J. (2019): "Disentangling preferences and limited attention: Random-utility models with consideration sets," Working paper, Research School of Economics, Australian National University.

Hastings, J. S., A. Hortaçsu, and C. Syverson (2017): "Sales force and competition in financial product markets: The case of Mexico's social security privatization," *Econometrica*, 85, 1723–1761.

Honka, E., A. Hortaçsu, and M. A. Vitorino (2017): "Advertising, consumer awareness, and choice: Evidence from the U.S. banking industry," *Rand Journal of Economics*, 48, 611–646.

Horan, S. (2019): "Random consideration and choice: A case study of 'default' options," *Mathematical Social Sciences*, 102, 73–84.

Kruskal, J. B. (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.

Lu, Z. (2019): "Estimating multinomial choice models with unobserved choice sets," SSRN Working Paper 3503554.

Luce, R. D. (1959): *Individual Choice Behavior: A Theoretical Analysis*, New York: Wiley.

Macon, N. and A. Spitzbart (1958): "Inverses of Vandermonde matrices," *The American Mathematical Monthly*, 65, 95–100.

Manzini, P. and M. Mariotti (2007): "Sequentially rationalizable choice," *American Economic Review*, 97, 1824–1839.

——— (2014): "Stochastic choice and consideration sets," *Econometrica*, 82, 1153–1176.

Masatlioglu, Y. and D. Nakajima (2013): "Choice by iterative search," *Theoretical Economics*, 8, 701–728.

Masatlioglu, Y., D. Nakajima, and E. Y. Ozbay (2012): "Revealed attention," *American Economic Review*, 102, 2183–2205.

McFadden, D. (2001): "Economic choices," *American Economic Review*, 91, 351–378.

Mead, L. R. and N. Papanicolaou (1984): "Maximum entropy in the problem of moments," *Journal of Mathematical Physics*, 25, 2404–2417.

Ok, E. A., P. Ortoleva, and G. Riella (2015): "Revealed (p)reference theory," *American Economic Review*, 105, 299–321.

Rhodes, J. A. (2010): "A concise proof of Kruskal's theorem on tensor decomposition," *Linear Algebra and its Applications*, 432, 1818–1824.

Roberts, J. H. and J. M. Lattin (1997): "Consideration: Review of research and prospects for future insights," *Journal of Marketing Research*, 34, 406–410.

Salant, Y. and A. Rubinstein (2008): "(A, f): Choice with Frames," *Review of Economic Studies*, 75, 1287–1296.

Shocker, A., M. Ben-Akiva, B. Boccara, and P. Nedungadi (1991): "Consideration set influences on consumer decision making and choice: Issues, models, and suggestions," *Marketing Letters*, 2, 181–198.

Sidiropoulos, N. D. and R. Bro (2000): "On the uniqueness of multilinear decomposition of N-way arrays," *Journal of Chemometrics*, 14, 229–239.

Sims, C. A. (2003): "Implications of rational inattention," *Journal of Monetary Economics*, 50, 665–690.

Sovinsky Goeree, M. (2008): "Limited information and advertising in the U.S. personal computer industry," *Econometrica*, 76, 1017–1074.

Train, K. E. (2009): *Discrete Choice Methods with Simulation*, New York: Cambridge University Press.

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

——— (2013): "Behavioral implications of shortlisting procedures," *Social Choice and Welfare*, 41, 941–963.

Van Nierop, E., B. Bronnenberg, R. Paap, M. Wedel, and P. H. Franses (2010): "Retrieving unobserved consideration sets from household panel data," *Journal of Marketing Research*, 47, 63–74.