WEIGHTED LINEAR DISCRETE CHOICE[†]

CHRISTOPHER P. CHAMBERS[*], YUSUFCAN MASATLIOGLU[§], PAULO NATENZON[†], AND COLLIN RAYMOND[*]

Abstract. We introduce a new model of stochastic choice. The model assigns each choice option a utility, along with a salience parameter reflecting economic frictions involved in choice. The model is consistent with many classical approaches, including preference maximization as in Machina (1985), as well as classic random utility. We characterize our model behaviorally and investigate its comparative statics properties. We show that the model generates intuitive closed-form solutions in equilibrium settings where firms can choose price, quality, and advertising. In addition, we show that the model is consistent with data that many typical discrete choice approaches cannot accommodate, and allows for simple preference parameter identification.

1. Introduction

This paper introduces a *simple and tractable* yet *flexible* model of probabilistic discrete choice, which we call the weighted linear (WL) model of discrete choice. In our model, each choice option is described by two parameters: one reflects the underlying quality or utility of an item, while the second captures the ease of choosing an item, as a result of its salience. Thus, our model can account for experimental findings from cognitive science and marketing regarding the importance of salience as prominence in choice (e.g., Milosavljevic et al. (2012), Sutherland and Galloway, 1981). The WL model adds a single parameter to describe each option compared to the widely used Luce model (also called the multinomial logit model). Although many

Date: Mar 2024.
[†] We thank Ian Krajbich, Joe Mazur, Yesim Orhun, Andre Veiga and Chenyu Yang for their helpful comments. We are especially grateful to Christopher Turansick for excellent research assistance.
[*] Department of Economics, Georgetown University, ICC 580 37th and O Streets NW, Washington DC 20057. E-mail: Christopher.Chambers@georgetown.edu.
[§] University of Maryland, 3147E Tydings Hall, 7343 Preinkert Dr., College Park, MD 20742. E-mail: yusufcan@umd.edu.
[†] Olin Business School, Washington University in St. Louis, 1 Brookings Drive, St. Louis, MO 63130, pnatenzon@wustl.edu.
[*] Johnson School of Management, Cornell University, Ithaca, NY 14853, collinbraymond@gmail.com.

models generalize the Luce model, we will demonstrate that our model has particularly strong explanatory and predictive power in environments where products exhibit asymmetries, especially when responding to changes in the choice set. The WL model can sidestep some of the "counter-intuitive implications" of commonly used models that S. Berry and Pakes (2007) highlight (see also Benkard and Bajari (2001)), and rationalizes flexible cross-price substitution effects, intuitive market share responses to the introduction of new products and the relationship between markups and advertising under oligopolistic competition. In addition, the WL also performs well when conducting demand estimations and counterfactual predictions relative to alternative formulations of random choice.

Section 2 introduces the WL model. We formulate the probabilistic choice behavior in the model as the solution to a simple utility maximization problem subject to the cost of choosing items too frequently. As in other recent models of deliberate randomization, the decision-maker chooses the probability with which each option is realized. Her goal is to maximize the expected utility of the chosen object while keeping down costs that are quadratic in individual probabilities.[1] The marginal cost of choosing an option is directly proportional to its choice probability and inversely proportional to its salience. Hence, in our formulation, the scale of the cost depends on the salience of each alternative (we discuss our interpretation of salience in more detail in two paragraphs).

Outcomes in our model are therefore ranked by two orderings: a utility function u, which captures "how good" an item is, and a salience function, m, which captures "how salient" an item is. Under a simple condition on these parameters, the decision maker's problem has an interior solution, where the probability of choosing x from S is equal to

$$\rho(x|S) = \underbrace{\frac{m(x)}{\sum_{y\in S} m(y)}}_{\text{Base Probability}} + \underbrace{m(x)[u(x) - \bar{u}_m(S)]}_{\text{Comparative Probability Transfer}} .$$

Here, $\bar{u}_m(S)$ is the weighted average utility of outcomes in S with respect to m; that is $\bar{u}_m(S) = \frac{\sum_{y\in S} m(y)u(y)}{\sum_{y\in S} m(y)}$. The probability of choosing x is the sum of two components: a base probability plus a comparative probability transfer. The first, the base probability, reflects how easy it is to think of x compared to the rest of the choice set and mirrors Luce's choice rule (Luce, 1959). The comparative probability transfer, in line with

[1]A general form of quadratic preferences over lotteries is given decision-theoretic foundations in Chew et al. (1991, 1994).

Fechernian stochastic choice (Fechner, 1860), reflects the difference in utility between the item and a weighted average of other items in the choice set, where the weights depend on the salience of the item. The salience of x then scales this difference.

We link our model to an object-specific notion of salience based on prominence and in line with intuitions in cognitive science on visual salience (e.g., Milosavljevic et al. (2012); Towal et al. (2013); Janiszewski et al. (2013); Weingarten and Hutchinson (2017); Dai et al. (2020)), and in marketing on brand salience (e.g., Hoyer and Brown, 1990) and has been tied to advertising (Sutherland and Galloway, 1981, Moran, 1990, Yi, 1990 Miller and Berry, 1998).[2] Consistent with these approaches, and Bordalo et al., 2022's definition, we think of salience as capturing a "bottom-up" mental process. We review some stylized facts from these literatures in Section 2 and how they accord with our approach. Although recent models in cognitive science have been developed to jointly explain attention and visual salience as well as choice, our model, which abstracts away from the choice process, has the advantage of being quite tractable, even for large choice sets, as well as matching many stylized facts.

At the same time, our approach is the stochastic choice equivalent of a linear demand system, widely used in economics. In our model, demand for a product reflects: (i) a base component independent of the utility of the item and (ii) a component that reflects the difference in utility between the item and the average item scaled by some number that captures how easy it is to consider alternative products (i.e. a kind of friction) in the market (as the number gets bigger, the best items eventually attract the entire market). The approach of linear demand systems has been used extensively in applied settings (early references include Shubik and Levitan (1980), Dixit and Stiglitz (1977), Spence (1976), and Singh and Vives (1984)).[3]

The WL model has straightforward behavioral foundations. Section 3 demonstrates that three simple axioms summarize all behavioral implications of the WL model. Structurally, the behavioral content of the model is characterized by a novel type of acyclicity condition. These axioms connect the (unobserved) components of

[2]Bordalo et al., 2022 describes salient stimuli as those that attracts attention through an involuntary mental process, and is typically driven by contrast, surprise or prominence. There is a recent body of work in economics on context and attribute-based salience driven by contrast and surprise (e.g., Bordalo et al., 2013) which embodies a distinct approach to ours.

[3]Our model naturally nests the multinomial logit model, as well as the "simple" version of linear demand systems, where all firms are symmetric in their ease of choice.

the model to observed choice behavior. We show that identification can be achieved by observing choices on relatively few choice sets.

Section 4 compares the WL to the additive perturbed utility (APU) approach of Fudenberg et al. (2015) and random utility model (RUM). Although logically distinct from the APU, our model has a non-trivial intersection with it, which nests both the multinomial logit model as well as the simple linear demand approach described above. Although our model has a relatively novel representation in terms of choice probabilities it is a subset of the well-known class of RUMs (characterized by Falmagne (1978)). It also shares a common interpretation with most existing models of random choice (including discrete choice models). We also define and demonstrate how the WL model naturally generates "rank polarization," which, given a RUM, is the sum of the proportion of consumers that rank an option as their first (best) option and as their last (worst) option among all the options. Rank polarization subsequently plays a role in interpreting the WL model's ability to explain empirical patterns, and to predict market shares.

Section 5 demonstrates the explanatory power of the WL model, showcasing its ability to accommodate several empirical regularities, and highlighting the potential benefits relative to the multinomial logit as well as a wider class of discrete choice models. First, we show that the binary choice comparisons in our model satisfy moderate stochastic transitivity, which means that it can match widely observed choice patterns that cannot be explained with the stronger notion of transitivity satisfied by models like logit and APU. Second, we examine substitution patterns in our model, including those derived from utility, salience, and price changes. We show how our model can allow for richer patterns of cross-price substitution relative to the logit approach. We show that in line with intuitions, items with higher salience induce higher cross-price elasticities. Third, we show that our model can accommodate natural shifts in market share when new products are introduced, and that these shifts are tied to rank polarization. We show that items which are rank polarized have more similar market shares across choice sets of different sizes, and that higher utility means that an item will lose relatively less market share when new products are introduced. In fact, high utility items can even maintain non-negligible market share when an infinite number of new products are introduced, unlike many other models of discrete choice. Finally, we extend the model to allow for firms to compete by influencing utility via setting price, and salience via advertising (in line with evidence from marketing that shows brand salience is influenced by advertising). We show that the model can tractably capture these kinds of oligopolistic competition, delivering closed form solutions, and that the model generates relationships between markups, the number of firms and advertising consistent with intuitions and existing evidence.

Section 6 highlights the predictive power of the WL model. We focus on comparing the performance of the WL model in applications to some of the most well known models in discrete choice estimation, namely classic logit, nested logit and (covariance) probit. To assess the strengths of each model in a variety of situations, we use simulated data. We estimate each model in 14,850,000 simulated datasets, covering the entire range of demand systems that come from a population of heterogeneous rational agents, indexed by their level of extremeness (that is, rank polarization) of tastes. We compare the out-of-sample predictions of these models using several standard metrics. The overall picture that emerges from the simulations is that, while retaining the simplicity and the tractability of the logit model, the WL is able to outperform more complex models across a wide range of situations. First, the WL model generally outperforms both the classic logit model as well as the nested logit model. Second, the WL model predicts best, relative to probit, in several specific instances: i) extreme degree of rank polarization, ii) no rank polarization, and iii) predicting market shares for intermediate choice set size (i.e. ternary sets). Thus, rank polarization is an important measure to track the predictive performance of the WL model versus the alternatives. Moreover, even outside of these situations, the WL model performs relatively close to probit, despite having fewer parameters. Hence, while the WL does not clearly dominate the Probit model, and performs slightly worse in some cases, it could prove a better model to use given its smaller number of parameters, closed form choice probabilities, and easy interpretation.

Finally, Section 7 concludes. The Appendices include proofs and discuss other relevant ideas, such as behavioral properties of the analog of our model, which allows for zero probabilities.

2. Model

2.1. **Functional Form.** We initially describe our model in an abstract environment, and then in later sections apply it to particular settings. Let X be a finite set of alternatives. Let $\mathcal{X}$ be the set of all probability measures on X. That is, $\rho(\cdot|X) \in \mathcal{X}$

implies $\rho(x|X) \geq 0$ and $\sum_{x \in X} \rho(x|X) = 1$. Let $\mathcal{D}$ denote the set of non-empty subsets of X. For every $S \in \mathcal{D}$, denote by $\mathcal{S}$ the elements in $\mathcal{X}$ which naturally induce probability measures on S, i.e., $\rho(\cdot|S) \in \mathcal{S}$ means $\rho(x|S) \in \mathcal{X}$ and $\rho(x|S) = 0$ whenever $x \notin S$. $\rho(x|S)$ denotes the choice probability (market share) of x in S. We also denote the sum of choice probabilities in $T \subset S$ as $\rho(T|S)$. Similarly, for any real function, f on X, $f(S)$ denotes the sum of $f(x)$ for all $x \in S$. We will denote binary choices as $\rho(x, y)$ instead of $\rho(x|\{x, y\})$. A *stochastic choice* (sometimes called a stochastic choice rule or stochastic choice function) is a family $\{\rho(\cdot|S)\}_{S \in \mathcal{D}}$, where each $\rho(\cdot|S) \in \mathcal{S}$.

We now introduce our parametric model of stochastic choice. In this model, each alternative x is represented by two values: its utility $u(x)$, and its salience $m(x)$ (for details on our interpretation of salience, please see the next subsection). Motivated by the "stochastic choice as optimization" paradigm of Machina (1985) and Cerreia-Vioglio et al. (2019), we suppose that for any set S, probabilities arise as the solution to maximizing expected utility less quadratic cost:

$$\mathcal{P}(S) = \underset{\rho(.|S) \in \mathcal{S}}{\operatorname{argmax}} \sum_{x \in S} \left(u(x)\rho(x|S) - \frac{1}{2m(x)} \rho(x|S)^2 \right) \tag{1}$$

The objective function in (1) describes individuals who want to maximize utility but face the cost of assigning too much weight to one particular product. The marginal benefit of choosing option x is given by its utility $u(x)$; while the marginal cost is $\rho(x|S)/m(x)$. Hence, the marginal cost is directly proportional to the current probability of choosing x and inversely proportional to its salience $m(x)$. Thus, we can interpret those items with high salience (high m) as those with a low penalty attached to choosing them. Our key assumption is that salience is always positive: $m(x) > 0$ for all x.

Figure 1 illustrates the model graphically for $X = \{x_1, x_2, x_3\}$. The point inside the triangle represents the choice probabilities from X, and is denoted by $\rho(x_1, x_2, x_3)$. Each green elliptical curve represents an indifference curve in the simplex. The preference is increasing towards $\rho(x_1, x_2, x_3)$. The DM maximizes his utility at this point when he is free to choose from the entire simplex. When x_3 is not available, the DM maximizes his utility subject to being on the $x_1 - x_2$ edge. The choice from $\{x_1, x_2\}$ is determined by the highest level curve having an intersection with this line, denoted by $\rho(x_1, x_2)$.

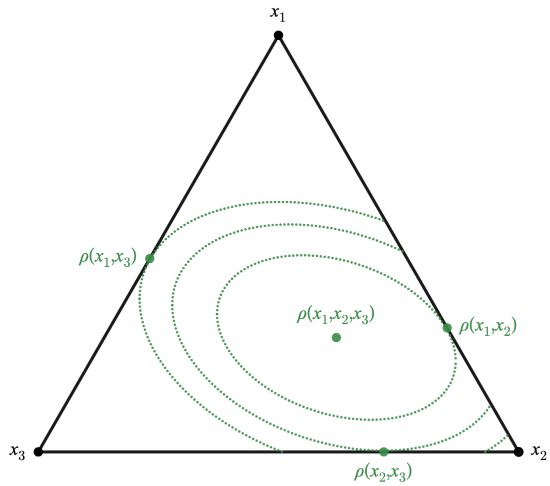

FIGURE 1. Choice probabilities in the WL model are obtained as a solution to the maximization of expected utility minus quadratic costs. The point $\rho(x_1, x_2, x_3)$ is optimal when all three options are available. When x_2 is removed from the choice set, optimal choice probabilities $\rho(x_1, x_3)$ lie on the indifference curve tangent to the edge of the simplex that contains x_1 and x_3.

We will focus on situations where the solution to (1) features only positive probabilities. Then, the first order conditions to (1) are

$$u(x) - \frac{1}{m(x)}\rho(x|S) = \Lambda(S) \text{ for all } x \in S$$

where $\Lambda(S)$ is the Lagrange multiplier of the constraint that the probabilities add up to 1. Summing across the elements of S and a bit of algebra implies

$$\Lambda(S) = \frac{\sum_{y\in S} m(y)u(y) - 1}{\sum_{y\in S} m(y)}.$$

Plugging back into FOC, we get $\rho(x|S) = m(x)(u(x) - \Lambda(S))$ or

$$\rho(x|S) = \frac{m(x)}{m(S)} + m(x)[u(x) - \bar{u}_m(S)] \tag{2}$$

where $\bar{u}_m(S) \equiv \frac{\sum_{y\in S} u(y)m(y)}{m(S)}$ is the weighted utility with respect to m in S.

Definition 1. A stochastic choice ρ is consistent with a *weighted linear* (WL) model on $\mathcal{D}$ if there exist functions $u : X \to \mathbb{R}$ and $m : X \to \mathbb{R}_{++}$ such that (2) holds for all $S \in \mathcal{D}$. We also say (u, m) *represents* ρ, or (u, m) is a *WL representation* of ρ.

Not every pair of functions (u, m) generates strictly positive choice probabilities as the solution to (1); hence, not every pair u, m generates a stochastic choice ρ with (2). It is straightforward to characterize the set of options that are chosen with positive probability in (1) for any pair (u, m) (see Appendix B). This characterization is directly related to shadow prices. We show in the appendix that the solution to the optimization problem (1) is unique. Moreover, we show that the set of options chosen with strictly positive probability is the unique subset S such that the utility of each option in the set is strictly larger than the shadow price for that set; that is, $u(x) > \Lambda(S)$ for every $x \in S$. Finally, we also show that if an option is chosen with strictly positive probability from the grand set X, then it must be chosen with strictly positive probability in every subset in which it appears. This directly leads to a necessary and sufficient restriction on any pair u, m to be the parameters of a WL model:

Proposition 1. *(u, m) is a WL representation for some ρ iff $u(x) > \Lambda(X)$ for all x.*

The closed-form solution (2) that defines our model has an intuitive interpretation. The model suggests that each product has a base probability of attracting consumers $m(x)/m(S)$. This can be thought of as how easily the product comes to the mind of consumers or the salience of the product in the marketplace. Of course, not all products might be available at any given time. So, conditional on a particular choice set S, the probability of product x getting noticed or x's presence in the market (given the set of available products) is proportional to $m(x)$. While this component is context-dependent, the relative base probabilities are fixed across different choice sets.

The base probabilities are not the only determinants of choice probabilities in our model. The utility of each product also plays an important role. Products gain or lose consumers in proportion to the difference between their utility $u(x)$ and the weighted average utility in the market $\bar{u}_m(S)$. The individuals' impression of what the average utility is in the market is influenced by the presence or salience of the products. The deviation from $\bar{u}_m(S)$ captures the market influence on probabilistic demand. In this formulation, the choice probability of a product increases linearly in its own utility. If a product offers a higher utility than the market average, it enjoys additional choice probability. Otherwise, being less than the market average reduces choice probabilities. Similarly, the probability of choice is linearly decreasing in the utility of other products. Unlike the base probabilities, there exists a non-trivial context dependence in this component.

Our formulation is a specific case of a larger class of demand functions where choice probabilities (i.e., market shares) are linear in the net utility of a product, which are often used in applications. Early examples are Shubik and Levitan (1980), Dixit and Stiglitz (1977), Spence (1976), and Singh and Vives (1984), while Choné and Linnemer (2020) provides a survey.

In interpreting the parameters, and linking them to empirical evidence (such as that discussed in the next subsection) it is useful to think about what happens to demand (i.e. the probability of being chosen) for a given object as both u and m change. First, we note that increasing either $u(x)$ or $m(x)$ results in an increase in the choice probability $\rho(x|S)$. For both of these changes, there is a direct effect of the change in the parameter, as well as an indirect effect, which emerges because of the change in $\bar{u}_m(S)$. Given that $\rho(x|S) > 0$, algebra shows that

$$\frac{\partial \rho(x|S)}{\partial u(x)} = \frac{m(x)m(S \setminus x)}{m(S)} \quad \text{and} \quad \frac{\partial \rho(x|S)}{\partial m(x)} = \frac{m(S \setminus x)}{m(x)m(S)} \rho(x|S)$$

Clearly, the probability of choosing x is an increasing function of both $u(x)$ and $m(x)$. Moreover, viewed as a function of u, for a given $S \in \mathcal{D}$, $\rho(x|S)$ is linear. This implies that by reducing $u(x)$ it is possible for x to be chosen with probability arbitrarily close to 0. Similarly, increasing $u(x)$ enough will eventually lead to all other items being chosen with 0 probability. It is also easy to see from this that $m(x)$ and $u(x)$ are complements: increases in utility have a larger impact on highly salient items, and vice versa.

Unlike utility u, the effect of salience m on choice probabilities is non-linear. If $u(x) \geq \bar{u}_m(S)$, then the probability of x being chosen will go to 1 as m increases. If not, the probability of x being chosen converges to an upper bound strictly below 1. In contrast, as $m(x)$ decreases, the probability of x being chosen falls to 0 (in line with the idea, discussed in the next subsection, that "unseen is is unsold"). Moreover, holding the initial choice probability constant the effect of a change in salience on choice is larger when $m(S)$ is larger — e.g. when the choice set is large.

2.2. **Interpreting Salience.** A novel feature of our approach is the importance of salience or market presence in determining choice probabilities. Salience, as a term, has been widely applied across disciplines to mean a variety of related concepts. Bordalo et al., 2022 summarizes salience as capturing an involuntary process of attracting attention through contrast, surprise, or prominence. Moreover, some work focuses on the salience of objects, others on the salience of attributes. For example, Bordalo et al., 2013 develops a model (that has been applied by a variety of researchers) where certain aspects of an item are more or less salient depending on the choice set (and so depends on contrast). In a distinct line of work Chetty et al., 2009 and related papers consider what happens when one aspect of a good may be less salient (e.g. the post-tax price, relative to the pre-tax price).

In our model, salience is an inherent feature of the object itself, rather than being derived from comparisons to other objects via, for example, contrast effects (although the weight of salience in determining choice is determined by the context in our model). We think that this corresponds most clearly to a notion of prominence, and relates to existing work on visual salience (in cognitive science), or brand salience (e.g., in marketing). Of course, work in both those fields also relates salience to other context-related factors like contrast. We assume that the salience of an object is treated as given by the decision-maker, although it may be manipulable by others, through, for example, advertising. We now turn to discuss how our particular model relates to stylized evidence on the importance of salience in both cognitive science and marketing literature.

First, we focus on the relatively recent literature examining visual salience. The initial stylized fact that our model captures is fairly intuitive: extensive evidence indicates that choice probabilities increase in salience. The best evidence comes from work on visual salience. Although the evidence in economics is relatively small, (Li and Camerer, 2022, Reutskaja et al., 2011), there extensive documentation in cognitive science (e.g., Bialkova and van Trijp, 2011, Milosavljevic et al., 2012, Orquin and Loose, 2013, Orquin and Lagerkvist, 2015, Peschel et al., 2019, Bhatnagar and Orquin, 2022). Moreover, the evidence indicates that salience is combined with subjective value (i.e. utility) in order to determine choice (Navalpakkam et al., 2012, Towal et al., 2013). Milosavljevic et al. (2012), find that the effects of salience on choice are particularly strong when preferences are weak, and salience can be as influential as preferences in determining choice probabilities; although other studies find subjective valuations matter more (Towal et al., 2013).

Second, studies have found salience increases choice probabilities in a way that is multiplicative with subjective values, rather than simply being additive with the utility value (Krajbich et al., 2010,Smith and Krajbich, 2019). In other words, subjective value (i.e. utility) and salience are complements, exactly what occurs in the WL model.

Third, in the WL model (although we focus on situations where all choice probabilities are positive) it is the case that when utility differences are large enough, the choice is determined by utility (so long as all items have positive salience). This is in line with findings in Li and Camerer, 2022 where if the value difference between items is large enough, salience seems to have a negligible impact on choice.

Fourth, the WL model predicts that as salience increases, choice probabilities increase, but that a relatively unattractive item will never be always chosen just because of a high m. It is the case that in literature increased salience does not lead to an item always being chosen, e.g., Li and Camerer, 2022. However, it is not clear that the literature is well powered to detect this.

Fifth, again though we focus on strictly positive probabilities, clearly as the salience of an item goes to 0 the probability that it is chosen goes to 0. This captures the adage in marketing that "unseen is unsold" (see Chandon et al., 2009 for a recent discussion of this).

Last, the WL model makes predictions about what happens to the impact of u and m as the choice set size changes. Unfortunately, most work studying the relationship between visual salience and choice has focused on binary choices, although a few have focused on choice sets of sizes 3 and 4 (Towal et al., 2013, Krajbich and Rangel, 2011, Gluth et al., 2018). The one study we know of looking at relatively larger choice set sizes, Thomas et al., 2021, looks at choice sets of size 9 to 36 and finds little role for salience in choice probabilities, in contrast to the work on binary choice sets, and in line with what our model predicts. Although not explicitly testing visual salience, Reutskaja et al., 2011 considers choice sets of size 4, 9, and 16 and finds that item location matters more for choice probabilities going from 4 to 9, it's importance then falls going from 9 to 16.

The literature on brand salience in marketing often leverages field data (with both the benefit of external validity and the cost of identification). Broadly speaking, this field's findings about the impact of salience accord with, and oftentimes are directly related to, the findings on visual salience in cognitive science.

The salience of a good or brand has been conceptualized as measuring the prominence of a brand in memory (Wedel and Pieters, 2000, Romaniuk and Sharp, 2003, Nedungadi, 1990, Zhang et al., 2021), visual, perceptual or cognitive awareness (Wedel and Pieters, 2000, Van der Lans et al., 2008, Janiszewski et al., 2013,Weingarten and

Hutchinson, 2017, Dai et al., 2020), how often it enters into consideration sets, or how loyal customers are to it (Romaniuk and Sharp, 2004, Ehrenberg et al., 1997). Extensive work has shown that, as in our model, higher salience implies a higher chance of being chosen (Sutherland and Galloway, 1981, Ehrenberg et al., 1997, Romaniuk and Sharp, 2003, Vieceli and Shaw, 2010). Not only that, but salience (as measured through awareness) can impact choice even in the face of utility differences such as quality and price (Macdonald and Sharp, 2000, Hoyer and Brown, 1990).

The marketing literature also suggests that firms can directly alter the salience of their brands via advertising (Sutherland and Galloway, 1981, Moran, 1990, Yi, 1990 Miller and Berry, 1998 Krajbich et al., 2024). We specifically explore the implications of this in Section 5, where we consider an equilibrium setting where firms can choose both the price and salience (via advertising) of their products. Of course, this is not to imply that our model captures all important details about salience, or even visual salience. Clearly, our model abstracts away from attribute-based notions of salience, as in Bordalo et al., 2013 and the following literature. Moreover, our model is silent about important interactions between salience, directed attention, and time, that models such as the attentional drift-diffusion model capture (Krajbich, 2019; Krajbich et al., 2010). More generally, many models of salience in cognitive science explicitly take into account choice process data, which our model neglects. Although this means our model is less rich, it means it is far more tractable for solving out for choice probabilities.

The cognitive science literature has developed a set of models that are meant to explain how salience and choice interact (e.g., the attention drift-diffusion model Krajbich et al., 2010). These models are process-focused and describe not only the choice probabilities but the process by which those probabilities are reached. However, they have primarily been applied to binary and trinary (Krajbich and Rangel, 2011) choice sets, and are difficult to solve in closed form. In contrast, our model, while neglecting to capture the process of choice, attends to many of the same stylized facts, and could be seen as a simple reduced-form alternative that can be used to understand salience.

2.3. **Two Special Cases.** Our formulation nests two important special cases. The first one is when the utility function is constant but salience differs across alternatives. If $u(x) = u(y)$ for all x, y, the second term in the representation disappears and the choice probabilities are solely driven by m:

$$\rho(x|S) = \frac{m(x)}{m(S)}$$

Clearly, this is the classical model of Luce (1959). Thus, if utility is constant, the ties in utility are broken by salience, and we obtain Luce's choice probabilities. In fact, this is the only way to obtain Luce choice probabilities in our framework: it is straightforward to verify that ρ is a WL where $u(x) = u(y)$ for all $x, y \in X$ if and only if it has a Luce representation.[4] This implies that our model provides an alternative characterization for the Luce model, which arises as the solution of minimizing quadratic cost:

$$\rho_{Luce}(x|S) \quad = \underset{\rho(.|S)\in\mathcal{S}}{\text{argmin}} \sum_{x\in S} \frac{\rho(x|S)^2}{2m(x)} \tag{3}$$

The other extreme case is where salience remains constant, $m(x) = \bar{m}$ for all x. In this case, the first term is simply $1/|S|$, each alternative attracts attention uniformly. Then, the weighted utility average becomes the ordinary average, $\bar{u}_m(S) = \bar{u}(S)$, and we have

$$\rho(x|S) = \frac{1}{|S|} + \bar{m}[u(x) - \bar{u}(S)]$$

This is equivalent to the basic linear demand system that features prominently in many models of monopolistic competition.[5] To see this more clearly, without loss we define $u(x) = \bar{u} - p(x)$. If we call $p(x)$ the price of x, then we can define demands as

$$\rho(x|S) = \frac{1}{|S|} + \bar{m}[\bar{p}(S) - p(x)]$$

Although this equation abstracts away from the possibility of demand which is zero, a few points are worth mentioning. The equation is intended to measure the market share for each firm x. Each firm gets a base share $\frac{1}{|S|}$, and the residual arises from the deviation of their price from the average market price. The level of this deviation is influenced by the parameter $\bar{m}$, which we suggest can be interpreted as a cardinal measure of market friction. This means that, by lowering the price, each firm

[4]Our model is also distinct from the nested logit where alternatives within a nest must satisfy Luce's IIA, although the idea of nests can be introduced in our framework. Kovach and Tserenjigmid (2022) provides several characterizations for the nested logit and its generalizations.

[5]Linear probability models are popular for their tractability in discrete choice estimation (e.g. section 4.2 in (Ben-Akiva & Lerman, 1985)) and appear in solution concepts with boundedly rational players in noncooperative games Rosenthal (1989), Voorneveld (2006).

can steal some of the market share, but that amount depends on the friction in the market. As $\bar{m}$ marginally increases, all firms with prices lower than the average receive a positive gain in market share. On the other hand, if $\bar{m}$ gets very small, consumers tend to purchase equally across all firms, ignoring price.

3. Characterization, Uniqueness and Identification

We now discuss the behavioral implications of our model. The first behavioral postulate is the often-invoked notion of positivity. Positivity says that every alternative is chosen with a positive probability.[6]

Axiom 1. [Positivity] $\rho(x|S) > 0$ for every $x \in S$ and $S \in \mathcal{D}$.

The next behavioral postulate is a well-known property in the stochastic choice literature. It states that when the competition gets fiercer among alternatives, choice probabilities strictly decrease.

Axiom 2. [Strict Regularity] $\rho(y|S) < \rho(y|S \setminus \{x\})$ for every $x \in S$ and $S \in \mathcal{D}$.

To state our next axiom, we define $d(x|S,T) := \rho(x|S) - \rho(x|T)$, where $S \neq T$ and $x \in S \cap T$. The quantity $d(x|S,T)$ is simply the change in the probability of choosing x as the choice set T changes to S.

Axiom 3. [Product Rule for Differences (PRD)] For any x, y, z and $S_i, T_i \in \mathcal{D}$,

$$d(x|S_1,T_1)d(y|S_2,T_2)d(z|S_3,T_3) = d(z|S_1,T_1)d(x|S_2,T_2)d(y|S_3,T_3)$$

whenever the expressions are well-defined.

In terms of interpretation, the property states that this product of probability differences depends only on the collection of elements with respect to which differences are taken. It does not, however, depend on how these elements are assigned to the given budgets.

The Luce model satisfies Axiom 3. In addition, the Luce model satisfies a product rule for choice probabilities:

$$\rho(x|S_1)\rho(y|S_2)\rho(z|S_3) = \rho(x|S_3)\rho(z|S_2)\rho(y|S_1)$$

[6]As discussed, this assumption cannot be rejected by any finite data set but we relax it in the Appendix.

whenever $x, y \in S_1$, $y, z \in S_2$ and $x, z \in S_3$. The left-hand side is the probability of a revealed preference choice cycle in the direction $x \to y \to z$ and the right-hand side is the probability of a choice cycle in the opposite direction.

It is easy to show that, if Axiom 3 holds in the full domain, then the axiom implies that an analogous condition holds for products of differences of length n, for any $n > 3$. Conversely, taking $y = z$, Axiom 3 implies a strictly weaker condition of length $n = 2$. When all differences are positive, this becomes

$$\frac{d(x|S_1, T_1)}{d(y|S_1, T_1)} = \frac{d(x|S_2, T_2)}{d(y|S_2, T_2)}$$

which is a version of Luce's independence of irrelevant alternatives (IIA) for differences. It says the ratio between differences for two options x, y only depends on x and y; it does not depend on which other alternatives were present initially in the set with x and y nor on which other alternatives were added or removed when the set changed.

We now state our characterization. This result does not require that we observe choices from the full domain, only the menus with sizes 2 and 3.

Theorem 1. *Suppose $\mathcal{D}$ contains all menus with size 2 and 3. Then a stochastic choice function ρ has a WL representation on $\mathcal{D}$ if and only if it satisfies Axioms 1-3.*

The idea of the proof for sufficiency is as follows. We first define the salience of each alternative by using the ratio of differences $d(x|S, T)/d(y|S, T)$ where S and T are menus with sizes 2 and 3. Then, instead of directly constructing the utility function, we define the "shadow values" for an optimization problem, for each set in the domain. This step helps us to define the utility function. We then show that the data can be represented by the WL model.

Theorem 1 provides two simple tests for our model. While Axiom 2 is innocuous, Axiom 3 is based on a principle similar in spirit to Luce's IIA. In our axiom, the ratio of *relative* levels is important rather than the absolute levels as in Luce's IIA.

3.1. **Uniqueness.** Our model enjoys strong uniqueness properties. If (u, m) represents ρ, then $(au + b, \frac{1}{a}m)$ also represents ρ for $a > 0$ and b. We also show that if (u, m) and (u', m') represent the same choice data, they are equivalent up to the same class of transformations. The utility function is unique up to an affine transformation, whereas the salience function is unique up to a scale transformation. The scale parameter of utility is the inverse of the scale parameter of salience.

Theorem 2 (Uniqueness)**.** *Let (u,m) be a WL representation of ρ. Then (u',m') is a WL representation of ρ if and only if $u' = au + b$ and $m' = \frac{1}{a}m$ for $a > 0$.*

3.2. **Identification and out-of-sample prediction.** We examine what can be inferred about the primitives of the model based on observed choices. This is important for understanding the underlying model and its predicted behavior, as well as for making out-of-sample predictions. We consider an analyst who observes stochastic choice data. The analyst posits that the data is generated by the weighted linear model. The analyst would like to answer what are the utility and the salience parameters for each alternative. We show in this section how this question can be answered within the framework of our model. Moreover, we would like to illustrate that we can make out of sample predictions (outside $\mathcal{D}$) given that our representation is unique.[7] This illustration will also help the reader to understand the proof of Theorem 1 better.

Consider four alternatives $X \equiv \{x,y,z,t\}$, and suppose that $\mathcal{D}$ consists of all sets containing at most three elements. Imagine the choice probabilities from binary and ternary sets satisfy the following conditions:

(1) Choices from pairs are equiprobable: for all $a, b \in X$, $\rho(a,b) = 0.50$,
(2) For any triple, if y, z, t are members of the triple, then they are chosen with equal probabilities,
(3) x is chosen with probability 0.30 from any triple.

These are our choice data on $\mathcal{D}$. Since $X \notin \mathcal{D}$, choices from the entire set X are not observed. Our goal is (i) to identify u, m values for each and (ii) to predict choice probabilities from X.

When two alternatives, a and b, are both chosen with positive probability from two sets, and the probability that they are chosen differs in both sets, then it becomes easy to identify the ratio of their salience parameters: $\frac{m(a)}{m(b)} = r_{S,T}(a,b)$ where both S and T including a and b. For example, we have

$$\frac{m(x)}{m(y)} = \frac{\rho(x|\{x,y\}) - \rho(x|\{x,y,z\})}{\rho(y|\{x,y\}) - \rho(y|\{x,y,z\})} = \frac{4}{3}$$

So, by symmetry among y, z, t, we may conclude directly that $m(y) = m(z) = m(t) = (3/4)m(x)$. We may normalize up to scale, so let us suppose that $m(x) = 4$, and that

[7]The parameters of the models are derived from the domain $\mathcal{D}$.

$m(y) = m(z) = m(t) = 3$. Once m is identified up to scale, we can use the equality

$$u(a) - u(b) = \frac{\rho(a|S)}{m(a)} - \frac{\rho(b|S)}{m(b)}$$

to identify u. Since we may normalize u up to translation, this allows us to choose $u(x) = 0$. In so doing, it becomes apparent that $u(y) = u(z) = u(t) = 1/24$. This is the full identification of our model. With these identifications in hand, we may directly conclude that $\rho(x|\{x, y, z, t\}) = 5/26$, whereas $\rho(y|\{x, y, z, t\}) = \rho(z|\{x, y, z, t\}) = \rho(t|\{x, y, z, t\}) = 7/26$, thus affording an out of sample prediction.

Even outside of this particular situation, our approach allows for a very transparent identification of the two parameters. The key function introduced above, $r_{S,T}(x, y)$, identifies m up to a scale factor: $r_{S,T}(x, y) = \frac{m(x)}{m(y)}$. Given our identified m's, we can then identify the ranking of u's by defining $u(x) - u(y) = \frac{1}{m(x)}\rho(x|S) - \frac{1}{m(y)}\rho(y|S)$.

Our identification result is based on choice set variations observed in the data. In the Appendix, we discuss how our model can be easily identified in choice settings with a fixed choice set (no choice set variation) but with observable attributes of outcomes. In particular, we can show that one can identify the parameters via solving a set of simple linear equations. This identification will improve the applicability of our model in different environments.

4. Relation to APU and RUM

4.1. **Relation to APU.** Our formulation as a solution to maximizing expected utility minus costs in (1) brings to mind the Additive Perturbed Utility (APU) model (Fudenberg et al., 2015). APU is a random choice model for which

$$\rho(x|S) \equiv \arg\max_{p \in \mathcal{S}} \sum_{x \in S} [u(x)p(x) - k(p(x))]$$

where k is some strictly convex and smooth function. Although this formulation is similar to our approach, the distinction is twofold:[8]

[8]In their working paper, Fudenberg et al. (2014) weaken the first condition, and consider the more general model:

$$\rho(x|S) \equiv \arg\max_{p \in \mathcal{S}} \sum_{x \in S} [u(x)p(x) - k(p(x), x)]$$

which nests our approach. Unlike our model, this model does not have a closed-form solution.

(1) In APU, the cost of probability is independent of the alternative in question, that is, k depends only on the probability of choosing each x but does not vary across alternatives.
(2) In APU, the cost function k need not be quadratic.

In terms of the generated choice behavior, APU and WL have a non-trivial intersection. For example, taking k to be quadratic in the APU is equivalent to assuming a constant salience function m in the WL. That is exactly the situation where we recover simple linear demands.

Another example of intersection is that APU and WL both nest the classic logit model. We note, however, that logit arises in two very different ways. APU becomes logit when the cost $k(p)$ is equal to a multiple of $p \log p$, and the choice probability of each option x is proportional to the exponential of its utility $e^{u(x)}$. In contrast, we saw that the WL produces logit choice probabilities if and only if utility is constant and ties are broken by salience, whereby the choice probability of each option x is proportional to its salience $m(x)$.

Despite having a non-trivial intersection, the APU and WL models are logically independent. In Section 5, below, we show that WL allows more flexible patterns of choice in binary comparisons. APU, like logit, satisfies a strong form of transitivity of binary comparisons, while the WL relaxes it to a moderate form of transitivity. On the other hand, Fudenberg et al. (2015) show the APU can violate the Block-Marschak inequalities that characterize the random utility model (Barbera & Pattanaik, 1986; Falmagne, 1978). We now show that the WL is a RUM; hence, despite being defined as a solution to the individual maximization problem (1), perhaps surprisingly, the WL also represents the choices of a heterogeneous population of standard rational individuals. In that sense, our model provides some additional flexibility without sacrificing the rationality assumed in the predominant paradigm of applied discrete choice estimation.

4.2. **Relation to RUM.** The most well-known generalization of the Luce model is the random utility model (RUM). In order to define the class of RUMs first let $\mathcal{R}$ be the set of all possible linear orders (rankings) on X and π be a probability distribution over rankings. $\pi(\succ)$ represents the probability of $\succ$ being realized as the preference. Given a set of available alternatives A, the probability of an alternative x being chosen is determined by the probability of a ranking for which x is at the top of A. Let $\mathcal{R}(a, A)$

be the set of rankings of X which rank a at the top of A, that is, $\mathcal{R}(a,A) := \{\succ \in \mathcal{R} : a \succ b \text{ for all } b \in A \setminus a\}$. The RUM stochastic choice associated with π is ρ_π defined by:

$$\rho_\pi(a|A) = \sum_{\succ \in \mathcal{R}(a,A)} \pi(\succ)$$

Our model is a RUM. Thus, although it is more general than Luce, it still fits within the most classic paradigm of random choice.

Proposition 2. *If ρ has a WL representation then it is a RUM.*

Let (u,m) be a WL representation of ρ. We now show how to construct a RUM representation for this ρ. First, we normalize u by subtracting $\Lambda(X)$, the shadow price for the grand set. Theorem 2 implies that $(\tilde{u}, m)$ is also a WL representation of ρ, where $\tilde{u} = u - \Lambda(X)$. With this normalization, we have $\rho(x|X) = \tilde{u}(x)m(x)$ for each x and therefore $\sum[\tilde{u}(x)m(x)] = 1$.

Next, to each ranking of alternatives, with the n alternatives enumerated as $x_1 \succ x_2 \succ \cdots \succ x_n$, our RUM representation assigns the following probability:

$$\pi(\succ) = \tilde{u}(x_1)m(x_1)\frac{m(x_2)}{m(X \setminus x_1)}\frac{m(x_3)}{m(X \setminus \{x_1, x_2\})}\cdots\frac{m(x_n)}{m(x_n)} \tag{4}$$

This construction mimics the construction of Theorem III.6 in Block and Marschak (1959). It is then standard to show that $\rho = \rho_\pi$.[9]

Rank polarization. Equation (4) helps us to understand our model in terms of the relative frequency of preference types in the population. Consider two types $\succ$ and $\succ'$, who only differ in terms of the first two alternatives, i.e., $x_1 \succ x_2$ and $x_2 \succ' x_1$, otherwise $\succ = \succ'$. The relative frequency of these two types is $\frac{\tilde{u}(x_1)[m(X \setminus x_2)]}{\tilde{u}(x_2)[m(X \setminus x_1)]}$. Now consider a similar comparison whereby $x_2 \succ x_3$ and $x_3 \succ' x_2$, otherwise $\succ = \succ'$. Then, the relative frequency of these types is $\frac{m(X \setminus x_1) - m(x_3)}{m(X \setminus x_1) - m(x_2)}$. More generally, whenever we are comparing individuals who have the same rankings except for a reversal between x and y and neither option is ranked first, then only m-ratios are involved in the relative odds. Hence, the u-ratio of items only occurs when we consider the relative odds of rankings that differ in their best item.

In the Luce model, these relative frequencies are functions of m alone. To see this, recall that Luce is a special case of the WL model with constant utility $u(x) = \bar{u}$. In

[9] This representation is not unique in general. See Turansick (2021) for details.

this special case, $\tilde{u} = \bar{u} - \Lambda(X) = 1/m(X)$, and the equation (4) becomes

$$\pi(\succ) = \frac{m(x_1)}{m(X)} \frac{m(x_2)}{m(X \setminus x_1)} \frac{m(x_3)}{m(X \setminus \{x_1, x_2\})} \cdots \frac{m(x_n)}{m(x_n)}$$

Hence, the additional parameter in the WL model generalizes the Luce case only in the relative probability of the top-ranked alternative.

This adds enough flexibility, however, to allow the WL model to capture the full range of extremeness of tastes in the RUM model. For a given RUM π, define the *rank polarization* of each option $x \in X$, denoted by $P(x, \pi)$, as the sum of the proportion of consumers that rank option x as their first (best) option and as their last (worst) option among all the options. Formally,

$$P(x, \pi) := \pi\{\succ \in \mathcal{R} : x \succ y \text{ for all } y \neq x\} + \pi\{\succ \in \mathcal{R} : y \succ x \text{ for all } y \neq x\}$$

Since Block and Marschak (1959), we know that the full distribution π from a RUM model cannot be identified from choice behavior. However, the probability that an option x is ranked in the k-th position is identified. In particular, whenever $\rho = \rho_\pi$ is a RUM, rank polarization is uniquely recovered from choice as follows:

$$P(x, \pi) = \rho(x|X) + \sum_{A : x \in A} (-1)^{|A \setminus \{x\}|} \rho(x|A)$$

In particular, while the RUM representation is not unique, we must have $P(x, \pi) = P(x, \pi')$ whenever $\rho_\pi = \rho_{\pi'}$. Hence, for a given RUM ρ the rank polarization measure is well defined; we could write $P(x, \rho)$ instead of $P(x, \pi)$ whenever ρ belongs to RUM.

By construction, the rank polarization of x may take any value in the range

$$\rho(x|X) \leq P(x, \pi) \leq 1 \tag{5}$$

where $\rho(x|X)$ is the market share of option x in X. For example, when $\rho(x|X) = 1/2$, one-half of the population ranks x as its first (best) option; and $P(x, \pi)$ can vary from $1/2$ to 1 according to how much probability π assigns to rankings where x is the worst (last) option. Equation (4) demonstrates that, by varying the tradeoff between the utility $u(x)$ and the salience $m(x)$ of an option x, the WL model can generate the full range of rank polarization in (5). This flexibility allows the WL to accommodate several empirical regularities (Section 5) and to perform surprisingly well in out of sample prediction (Section 6), despite having less parameters than other commonly used models.

Number of parameters. The RUM is the reigning paradigm for discrete choice estimation in applied work. However, in its most general formulation, the RUM can sometimes prove too unruly for practical use. With n alternatives, there are $n!$ possible rankings (or types) in a population. A probability distribution over types thus has $n! - 1$ free parameters, and these cannot be identified from stochastic choice. In applications, practitioners often employ special cases of RUM that (i) facilitate identification; (ii) provide more tractable and interpretable functional forms; and (iii) provide sharper out-of-sample predictions.

Figure 2 compares the number of free parameters in the WL model to the full RUM and some of the best-known special cases of RUM in the literature. Each one of these models attempts to provide a good balance between generality and practical use. The classic Logit model is the most extreme example of tractability and simplicity, with $n - 1$ free parameters. More general models like Nested Logit and Covariance Probit provide more flexibility at the expense of having more parameters, being less tractable, requiring more data for identification, and producing less sharp predictions out of sample.

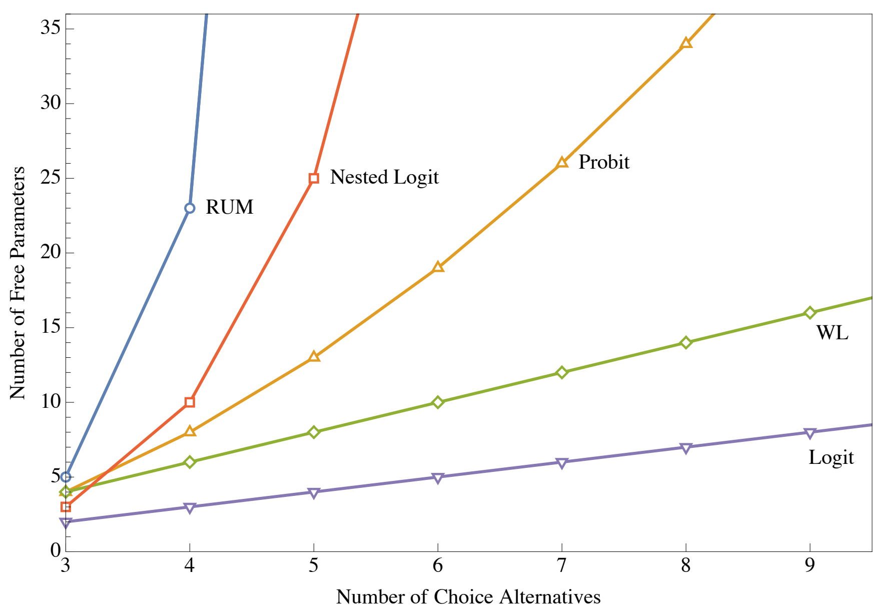

FIGURE 2. Comparing the WL to other special cases of RUM: number of free parameters as a function of the number of choice alternatives.

Compared to some of these alternatives, the WL model is closer in terms of simplicity and tractability to the classic logit model. In particular, WL, like logit, has a closed-form formula for choice probabilities, and its number of parameters increases linearly with the number of alternatives. In the next section, we show the WL adds enough flexibility to accommodate a wide range of relevant empirical phenomena; in Section 6, we also show it performs extremely well in out-of-sample prediction.

4.3. **Relation to Other Models.** Finally, we discuss the relationship with other models. First, we consider the rational inattention model (RI) of Matějka and McKay, 2015, which makes an important connection between the optimal choice of information and stochastic choice. While their state-dependent choice probabilities resemble the Luce formula, RI violates RUM. Moreover, as opposed to WL, RI does not have closed-form expression for state-independent choice probabilities and requires much richer data. The intersection between WL and RI is non-empty, however, as both models nest the classic logit model: RI becomes the classic logit in the special case of a symmetric prior, and WL nests logit in the special case of constant utility, as we have shown in Section 2.3.

Cattaneo et al., 2020 characterizes a class of stochastic choice rules referred to as the random attention model (RAM). RAM attributes randomness in choice to attention for fixed preferences. As in RI, RAM model is more general than RUM and it is not uniquely identified. Two parametric models on limited attention are introduced by Manzini and Mariotti, 2014 and Brady and Rehbeck, 2016. It is routine to show that WL are distinct from these two models.

Finally, the model of Ellis and Masatlioglu, 2021 provides a characterization for a general class of deterministic models of categorization. This model generalizes the salience model of Bordalo et al., 2013. These models are based on observable attributes and are deterministic. As far as we are aware, there is no stochastic version of such models.

5. Explaining Empirical Patterns

We now turn to demonstrating the explanatory power of our model for stylized empirical facts about random choice and market demand. We highlight the explanatory power of our model relative to not just the multinomial logit approach, but also many other widely used discrete choice models. In doing so, we will leverage the intuitions of rank polarization — showing how items with "inverse" relative rankings of u and m are key for the added explanatory power of our model.

5.1. **Binary Choice and Stochastic Transitivity.** Since Thurstone (1927), binary comparisons have been a key focus of research on random choice. If both alternatives are chosen with strictly positive probability, a simple calculation, plus the substitution $U(x) = 2u(x) - \frac{1}{m(x)}$, demonstrates that in the WL model:

$$\rho(x, y) = \frac{u(x) - u(y) + \frac{1}{m(y)}}{\frac{1}{m(x)} + \frac{1}{m(y)}} = \frac{1}{2} + \frac{1}{2}\frac{U(x) - U(y)}{\frac{1}{m(x)} + \frac{1}{m(y)}} \tag{6}$$

In the literature, stochastic binary choices are often taken as indicating the strength of preference among alternatives. Researchers have used various notions of stochastic transitivity to evaluate the explanatory power of random choice models relative to experimental data. Weak, moderate, and strong transitivity are the most important notions in the literature.

Weak transitivity is the mildest of these notions. It requires the relation $\succeq_\rho$, defined by $x \succeq_\rho y$ if and only if $\rho(x, y) \geq 1/2$, to be transitive. Equation (6) shows that $\rho(x, y) \geq 1/2$ if and only if $U(x) \geq U(y)$. Thus, the WL model satisfies weak transitivity and one can interpret U as the "implied" utility function in a setting with binary choice sets.

In fact, the WL model satisfies a more demanding notion of *moderate transitivity*: if $\rho(x, y) \geq 1/2$ and $\rho(y, z) \geq 1/2$, then $\rho(x, z) \geq \min\{\rho(x, y), \rho(y, z)\}$, where the last inequality is strict, unless $\rho(x, z) = \rho(x, y) = \rho(y, z)$. This again follows from equation (6), which shows the WL model belongs to the class of moderate utility models (Halff (1976)). He and Natenzon (2024) show this class is characterized by moderate transitivity and achieves a useful compromise between explanatory and predictive power.

However, the WL model does not conform to the most demanding notion of *strong transitivity*. This postulate requires $\rho(x, z) \geq \max\{\rho(x, y), \rho(y, z)\}$ when $\rho(x, y) \geq 1/2$ and $\rho(y, z) \geq 1/2$. The departure from strong transitivity distinguishes the binary choices in the WL model from the classic logit model, as well as the APU model (Fudenberg et al., 2015). Relaxing strong transitivity gives the WL model the flexibility to accommodate widespread and systematic violations of this postulate found in empirical settings (Mellers et al., 1992; Rieskamp et al., 2006).

Rieskamp et al. (2006) discusses how the most frequent violations of strong transitivity arise from settings where options present trade-offs across multiple dimensions.

Likewise, in the WL model these violations are captured by options that have opposite ordinal ranking in utility and salience. To illustrate, consider two options x and y that are equally likely to be chosen in a binary comparison: $\rho(x, y) = 1/2$. In this case, strong transitivity requires that x and y have the same probability of being chosen in a binary comparison against any third alternative z. From equation (6) we have $\rho(x, y) = 1/2$ in the WL model if and only if

$$2u(x) - 1/m(x) = U(x) = U(y) = 2u(y) - 1/m(y)$$

Hence, x may have higher a utility $u(x) > u(y)$ exactly offset by a lower salience $m(x) < m(y)$ in order to generate a fifty-fifty choice against y. This allows x to perform differently from y against a third option z: while the numerators $U(x) - U(z) = U(y) - U(z)$ in equation (6) are the same, the denominators $1/m(x) + 1/m(z) > 1/m(y) + 1/m(z)$ are different. In particular, the more salient option y has a smaller denominator and a more extreme choice probability against z than the less salient option x. In other words, increased salience makes the decision-maker more sensitive to the same U difference.

5.2. **Substitution Patterns.** A key question confronting models of market demand is their predictions regarding cross-outcome substitution patterns. We next turn to describing substitution patterns in our model, looking at both utility (including price) and salience changes.

Consider what happens to $\rho(x|S)$ as $u(y)$ increases. Not unexpectedly, the demand for x will decrease. Moreover, it decreases in a linear fashion, that is proportion to the ratio of $m(x)m(y)$. In other words, the more salient either y or x, the more impactful a change in $u(y)$ on $\rho(x|S)$ is. This should not be surprising — changes in the utility of more salient items impact the demand for other items more, and a salient item (which is often thought about with other objects) also is impacted more by shifts in the value of competing items. Moreover, we have symmetry as

$$\frac{\partial \rho(x|S)}{\partial u(y)} = \frac{\partial \rho(y|S)}{\partial u(x)} = -\frac{m(x)m(y)}{m(S)}$$

analogous to the Slutsky substitution patterns.[10] We do not have symmetric responses for cross-m changes:

$$\frac{\partial \rho(x|S)}{\partial m(y)} \quad = \quad -\frac{m(x)}{m(y)m(S)}\rho(y|S)$$

Again, this should be relatively intuitive. The effect of changing y's salience should depend on the choice probability of y, $\rho(y|S)$, since items chosen with higher probability benefit more from increased salience.

These comparative statics extend in a straightforward manner to analyze the substitution patterns arising from changes in an attribute of an option x, such as price, that enter utility or salience. It is well known that, in the multinomial logit model, the cross-price elasticity of item x for item y does not depend on x, a condition that is at odds with both intuition and reality: we would expect that cross-price effects to be much more variable. Like other alternative models (such as nested logit) our model allows for more flexible patterns than logit.

To see this, suppose that price enters utility linearly $u(x) = \tilde{u}(x) - p(x)$ and let $m(x)$, $\tilde{u}(x)$ and the set of products, S, be fixed. Then, the cross-price elasticity of x for y is

$$\epsilon_{x,y} = \frac{m(x)}{\rho(x|S)} \frac{p(y)m(y)}{m(S)}$$

Thus, the WL model allows for any given pair of items x and y to have a distinct cross-price elasticity. The cross-price elasticities of x to y are increasing in the price of y as well as the salience of both x and y, but decreasing in the existing market share of x.

That said, our model still imposes restrictions on relationships between pairs of cross-price elasticities. In particular the ratio of $\epsilon_{x,y}$ to $\epsilon_{x,z}$ is equal to the ratio $\frac{p(y)m(y)}{p(z)m(z)}$, and so is independent of x. We believe there is a natural interpretation of this condition. Suppose y is more salient than z. In this case, when the price of y shifts, it is more noticeable (than if the price of z shifts) and so causes greater changes in the demand for x. Of course, because y is more salient than z, it also means that changes in the price of y will also impact the demand for w more than z. Thus, increased saliency leads to an increase in cross-price substitution patterns. Similarly, the ratio of $\epsilon_{x,y}$ to $\epsilon_{z,y}$ is independent of y (with a similar interpretation). Moreover, the ratio of elasticities is

[10]Positive semidefiniteness follows as well, because this matrix of substitution patterns is obviously diagonally dominant, with positive diagonal elements.

independent of choice sets (S) — in other words, the relative impact of y on x compared to z on x does not depend on the availability of other products.

Although we believe such patterns are realistic, they are also restrictive. However, we can use existing approaches to allow for additional flexibility (and realism) in our substitution patterns. Random coefficients models, such as mixed logit, generate more realistic patterns of substitution than logit by incorporating additional information provided by a vector of observable characteristics for each object $x = (x_1 \ldots, x_n)$. They usually specify a convenient utility function, such as $u(x) = e^{\beta' x}$, where β are randomly distributed coefficients. This formulation yields "local" substitution patterns, in the sense that changes in an attribute x_i of an object x will impact more intensely the market share of objects that are "close" to x in the space of observable characteristics. Given its tractable closed-form expression for choice probabilities, the WL model can be deployed in a setting with observable attributes and random coefficients in the same way, with the added flexibility of including a salience function, in addition to utility. The identification of the WL model with attributes is straightforward (see Appendix C).

5.3. **Market Share Persistence, Replicas, and the Introduction of New Products.** We now turn to explain how our model can more naturally capture the reaction of market shares to the introduction of new products compared to a wide variety of discrete choice models where the support of the error terms is continuous and unbounded above (including multinomial logit).

We first describe an immediate implication of our model in the presence of changing choice sets: larger choice sets benefit (in terms of choice probabilities) higher utility items. In Section 4, we defined rank polarization $P(x, \pi)$ as the sum of the proportions of consumers that rank x as their first (best) and last (worst) option in the RUM model π. Rank polarization has a clear empirical counterpart: more extreme tastes for x translate into smaller variations in the market share for option x as the set of options changes.

To see this, consider a set with three options $X = \{x, y, z\}$ and let ρ be the behavior of a population of consumers described by a RUM π. Then, the rank polarization for x can be obtained from market share data simply by

$$P(x, \pi) = 1 - [\rho(x, y) - \rho(x|\{x, y, z\})] - [\rho(x, z) - \rho(x|\{x, y, z\})]$$

The first term in brackets is the change in x's market share when z is removed from X; it is precisely the proportion of consumers that rank $z \succ x \succ y$. Likewise, the second term in brackets is the proportion of consumers that rank $y \succ x \succ z$. Hence, the equation above shows a direct relationship between the extremeness of tastes for x and the size of the market share that x gains (loses) when other options are removed (introduced).

Now consider a population where one-half of the consumers rank option x first, and the other half ranks x last. Notice that this implies from doubletons, x is always chosen one-half the time, e.g., $\rho(x, y) = \frac{1}{2}$. However, since nobody ranks option x between y and z, it retains the same market share throughout, that is,

$$\rho(x, y) = \rho(x|\{x, y, z\}) = \rho(x, z) = \frac{1}{2}$$

We can extend this intuition to larger choice sets: let X be any finite set of options and fix any value of $\rho(x, X)$, the market share of x in the grand set X. Let (u, m) be a WL representation of ρ. By Theorem 2, we may subtract a constant from the utility function u, if necessary, so that $\Lambda(X) = 0$ and $\rho(x, X) = u(x)m(x)$. We will adjust x so that it has, in the limit, the same market share for all choice sets. Multiply the initial utility $u(x)$ by a scaling factor $k > 1$ and divide the initial salience $m(x)$ by the same factor. As $k \to \infty$, $\rho(x|S)$ converges to $u(x)m(x)$ for all S as desired.

Intuitively what we've done in order to help x have a constant market share is simultaneously increase the u and decrease the m. In other words, as choice sets gets larger, the market share attached to x does not decline. This intuition extends more broadly, in a weaker form: the next Proposition shows that if x has a higher utility than y, it retains more market share in larger sets than y.[11]

Proposition 3. *Suppose $u(x) \geq u(y)$. Then $\rho(x|S) \geq \rho(y|S)$ implies $\rho(x|S \cup T) \geq \rho(y|S \cup T)$.*

Logit and other models based on a single-dimensional utility measure cannot generate nor approximate persistent market share behavior. In the case of both logit and the APU model, for example, the binary choices $\rho(x, y) = \rho(y, z) = \frac{1}{2}$ immediately imply that each option must be chosen with probability 1/3 from the large set (this

[11] One can extend the proposition to construct examples where a lower utility product enjoys a higher market share in a smaller set ($u(x) > u(y)$ and $\rho(y|S) > \rho(x|S)$), but enlarging the product variety favors the higher utility product ($\rho(x|S \cup T) \geq \rho(y|S \cup T)$).

is similar to the restriction on Debreu's red bus, blue bus example). These models' difficulty in accommodating persistent market shares is pervasive, easily seen to hold with a larger number of options, and with any market share values that are not exactly zero or one.

One particular way of increasing the choice set is by adding replicas of existing products. A replica only differs from the original product in terms of seemingly unimportant attributes, such as in the famous "red bus, blue bus" example (Debreu, 1960). Assume two distinct products, a C (car) and BB (blue bus), receive equal market share in a binary comparison. Introducing a third option RB (red bus) that closely resembles BB does not alter the likelihood of choosing C. One can show that, in our model, if we have $u(C) \geq u(BB) = u(RB)$, then the choice probability of the car could be any number between 0.33 and 0.5 after introducing RB. While the lower limit is the same as the prediction of the Luce model in this setting, the higher limit is in line with Debreu's argument that the likelihood of choosing C should stay the same.

In the previous example, we added a single low utility replica.[12] We could continue the process, adding many low utility replicas. As the next proposition shows, in contrast to many other discrete choice models, adding a plentitude of lower utility options does not drive the market share of a higher utility item to zero in the WL model.

Proposition 4. *For all x, if $\rho(x|S) > 0$, and denoting T_n as S with n replicas of x, then for all $y \in S$, there exists an $\epsilon_{y,x} > 0$ such that $\lim_{n\to\infty} \rho(y|T_n) > \epsilon_{y,x}$ if and only if $u(y) > u(x)$.*

In particular, if $u(y) > u(x)$, then in the limit $\rho(y|T_n)$ will go to $m(y)[u(y) - u(x)]$. As Benkard and Bajari (2001) note, almost all discrete choice models in markets with large numbers of items predict that demand for any one item must converge to 0.[13] In contrast, the WL model can also allow for non-negligible market shares even as the number of products grows infinitely large. Moreover, this can happen even when each of the additional products attracts strictly positive probability. Thus, the WL model accomodates the notion of dominant and inferior products.

[12] Throughout, we are defining that an option x' is a replica of an option x when it has the same salience and utility as x. This implies x' is a replica of x in the behavioral sense of Faro (2023).

[13] They show this is true, (focusing on the case of where there is an outside good present) in the models where, in addition to mild technical conditions, the support of the error terms is continuous and unbounded above. This not only nests the multinomial logit approach but almost all other widely used approaches, such as nested logit and random coefficients.

A distinct concern with many well-known discrete choice problems is that in large markets the share that any one item has must be strictly positive: so long as the net utility of any item is positive, it must be chosen with positive probability.[14] Thus, the introduction of a new product cannot drive an existing product out of the market. Although this is a well-known property of the multinomial logit model, Benkard and Bajari (2001) shows that under very mild assumptions (always satisfied in applications of discrete choice models), this issue extends to any model where the conditional error distributions have unbounded upper support and a continuous upper tail.

The WL model can allow for new entrants to drive some, but not all, incumbent items out of a market.[15] Observe that even if $u(x) - p(x) > 0$, if it is small enough we can construct choice sets T and S where $T \subset S$ and where $u(x) - p(x) - \Lambda(S) < 0 < u(x) - p(x) - \Lambda(T)$. Thus x will be chosen with positive probability in T but with 0 probability in S, despite it generating strictly positive net utility, conditional on purchase. At the same time, there could also be a $y \in T$ such that y is chosen with positive probability in both T and S.

5.4. **Markups and Advertising.** We now embed the WL model into an equilibrium model of price-setting and advertising and show it can naturally generate intuitive results around markups, advertising, and the relationship between them (see Rosen, 1974 for a seminal paper looking at price-setting and demand when products have multi-dimensional attributes). As discussed in Section 2, evidence from marketing is typically interpreted to mean that firms can directly alter the salience of their brands via advertising (Sutherland and Galloway, 1981, Moran, 1990, Yi, 1990 Miller and Berry, 1998). We show that the WL model allows for both tractable closed form equilibrium solutions, which match stylized facts from the literature.

We assume each firm i offers a single item, and that the item has an underlying exogenous quality $\tilde{u}_i$. The firms can compete by setting prices, p_i, and the total utility a consumer gets from item i is then $u_i = \tilde{u}_i - p_i$. However, we also allow the firms to compete on another margin: they can adjust the salience of items. In line with our interpretation of m, we suppose that increasing m_i reduces the mental friction

[14]Many specifications of the multinomial logit model require that net utilities (i.e. gross utility less price) are always positive (e.g. if the probability of choosing x is $\frac{e^{v(x)}}{\sum_y e^{v(y)}}$).

[15]However, our approach still puts some structure on zero probability choice — e.g., if firm x has 0 demand facing a set of competitors S, then adding additional competitors can never increase its demand above 0.

required to purchase i. This can occur because, e.g., advertising raises the salience of an item.[16] The flexibility embodied in our model via m allows us to easily capture these novel considerations, like competition on salience, outside of the typical one-parameter random utility models.[17]

In order to endogenize entry in the market, we allow for n firms (we refer to the set of firms as N). We normalize the size of the market to 1. All firms face a marginal cost of production k. We will think of m_i as being proportional to the amount of advertising — so that higher m_i corresponds to more advertising. We will assume that the cost of advertising m_i is $\gamma(m_i) = gm_i^2$, where g is the marginal cost of increasing advertising.

Given these assumptions, and letting $\Lambda = \frac{\sum_j m_j(\tilde{u}_j - p_j) - 1}{\sum_j m_j}$, the profit function for the firm is

$$(p_i - k)\left(m_i(\tilde{u}_i - p_i - \Lambda)\right) - gm_i^2 - hu_i^2$$

We will focus on symmetric equilibria, and so will suppose that exogenous variables are the same across all firms (i.e. that k and $\tilde{u}$ are the same for all firms), and that all firms play the same strategies (so that $p_i = p$ and $m_i = m$ for all i in equilibrium). We initially suppose that n firms exist in the market, but will later consider what happens when there is entry and exit. We also suppose k is small enough so that an equilibrium with positive firms profits exists for a fixed n.[18]

To solve, we take the first order conditions for p_i, $m_i(\tilde{u}_i - p_i - \Lambda) + M(p_i - k)m_i(-1 - \Lambda_p(N)) = 0$, and m_i, $(p_i - k)((\tilde{u}_i - p_i - \Lambda) - m_i\Lambda_m) + 2gm_i = 0$. Given the symmetry assumption $\Lambda = \tilde{u} - p - \frac{1}{mn}$ and $\Lambda_p = -\frac{1}{n}$, $\Lambda_m = -\frac{1}{m^2n^2}$. Substituting back into the first-order conditions, we obtain the following proposition.[19]

Proposition 5. *With n firms the symmetric equilibrium has the following solution:* $p_i = k + \frac{1}{m_i(n-1)}$, $m_i = \frac{1}{(2gn^2)^{\frac{1}{3}}}$.

---

[16]As Bagwell (2007), points out, advertising has been seen by economists as having three approaches: persuasive, informative, and complementary. Our model is closest in spirit with the informative approach to advertising, where advertising helps raise "awareness" of a product.

[17]In fact, our model can also be extended to allow for endogenous choice of quality, $\tilde{u}$ as well, while still being able to generate closed form solutions for the equilibrium.

[18]Therefore, in a symmetric equilibrium all products will be purchased with positive probability.

[19]The equilibrium p_i also characterizes the solution where m_i is exogenous.

Our model has several interesting, and empirically relevant, implications. First, suppose that the number of firms, n, is exogenous. As n gets large, markups $p_i - k$, must go to 0; which is in contrast to the equilibrium in the multinomial logit model.[20] Thus, the WL captures the intuition that in markets with a large number of firms, no firm has market power.

Second, we can also use the results to understand firm entry. Suppose that firms face a fixed cost of entry f. Then firm profits are $\frac{g^{\frac{1}{3}}(n+1)}{2^{\frac{2}{3}}(n-1)n^{\frac{4}{3}}}$. As the cost of advertising grows (i.e. g increases), we see a larger number of firms (n increases). The increase in g and n jointly cause m_i to fall for all firms; implying that advertising falls. Because m_i falls and n increases, it is not immediately obvious what happens to markups. But algebra (using the fact m_i depends on n and g, and that the 0 profit condition allows us to write g as a function of m), shows that $m_i(n-1)$ is falling in the equilibrium number of firms, implying price, and markups, increase.

Therefore, in equilibrium, we observe price and markups being positively correlated with the equilibrium number of firms n, but negatively correlated with the degree of advertising (m_i). Intuitively, high advertising costs mean firms cannot use advertising as effectively to gain market share by increasing their salience. This in turn means that firms compete less with each other on price. This means profits increase, and more firms enter.

This result, a correlation between low prices and high levels of advertising, is reminiscent of results that emerge in completely different contexts in Robert and Stahl (1993) and Bagwell and Ramey (1994) (who find that advertising is greater when prices are lower), and which have been extensively investigated in the literature (see Bagwell (2007) for a relatively recent survey). Our result is also in line with the empirical evidence in Syverson (2019), which indicates that reductions in market frictions (i.e., increases in m_i) prompt customers to shift towards larger, lower-cost sellers, creating higher market concentration but lower markups.

The microfoundations of our model also allow us to understand the degree to which advertising changes welfare (as in a large literature beginning with Butters (1977)). In our setting, advertising, by reducing the mental frictions involved in thinking about, and purchasing, a particular item, can generate consumer surplus (related mechanisms are developed in other papers, e.g., by Grossman and Shapiro (1984) in a horizontally

[20]As Benkard and Bajari (2001) point out, this issue applies more generally to GEV and random coefficients logit models, although probit models can avoid these implications.

differentiated market). Moreover, in equilibrium, because increases in advertising are associated with reductions in advertising costs, they are also correlated with reductions in markups, generating an additional channel of welfare gains. Thus, our model can directly link the amount of advertising in the market to consumer surplus via changes both in price and the mental costs associated with choice.

6. Predicting Choice

In this section, we turn to showing when, and how, the WL model does well at predicting choice. We compare our models to other widely used approaches and aim to provide two insights. First, our model has transparent identification, is computationally tractable, and has fewer degrees of freedom relative to some well-known alternative specifications. Second, our model performs well at out-of-sample prediction relative to many alternatives. We show this in two ways. First, it outperforms any of the alternative approaches we consider when a subset of the outcomes are dominant (i.e., they tend to maintain market share in larger markets) compared to others. Although our model may not perform as well as other models such as correlated probit in other situations because of its computational simplicity and far fewer number of parameters, it still may be preferred.

6.1. **Setting.** The set of choice alternatives is $X = \{1, 2, 3, 4\}$. We choose $n = 4$ choice options because it is the smallest number that allows enough choice menu variation to identify the parameters of all the models that we consider, and also allows us to carry out prediction exercises with both increases in the number of options (entry) and decreases in the number of options (exit/mergers).

6.2. **Data generating process.** We simulate choice data from heterogeneous populations of standard rational consumers. Each consumer in each population is described by strict ranking $\succ$ over the choice options. With $n = 4$ options, a population is divided into $n! = 24$ possible types, that is, strict rankings over X. As before, $\mathcal{R}$ denotes the set of types. A population is a probability π over $\mathcal{R}$, identified with an element of the 23-dimensional simplex $\Delta(\mathcal{R})$.

We slice $\Delta(\mathcal{R})$ according to the rank polarization of the tastes in the population, which we defined in Section 4. Table 1 categorizes the 24 types in $\mathcal{R}$ according to rank polarization for options 2 and 3. Each preference is represented as a column vector where the best option is located at the top and the worst option is located

	$P(3,\pi)$	$1-P(3,\pi)$
$P(2,\pi)$	2 2 3 3 1 4 1 4 4 1 4 1 3 3 2 2	2 2 2 2 1 1 4 4 1 4 3 3 3 4 3 1 3 3 1 4 4 3 1 3 4 1 4 1 2 2 2 2
$1-P(2,\pi)$	3 3 3 3 1 1 4 4 1 4 2 2 2 4 2 1 2 2 1 4 4 2 1 2 4 1 4 1 3 3 3 3	1 1 4 4 2 3 2 3 3 2 3 2 4 4 1 1

TABLE 1. A population of consumers is a probability π over the 24 preference types displayed. Each preference ranking is shown with the best option at the top and the worst option at the bottom. $P(2,\pi)$, the rank polarization of tastes for option 2, is the probability given by π to types in the first row of the table; while $P(3,\pi)$ is the probability given by π to types in the first column.

at the bottom. The rank polarization of tastes for the second option, $P(2,\pi)$, is the probability given by π to rankings in the top row, while $P(3,\pi)$ is the probability given by π to rankings in the left column.

We cover the entire range of possible levels of rank polarization $P(2,\pi)$ and $P(3,\pi)$ in a discrete grid $\{1/100,\ldots,99/100\}^2$. To draw a random population conditional on a fixed level of $P(2,\pi)$ and $P(3,\pi)$, we mix independent uniform draws over the rankings from each cell in Table 1. We treat the events that options 2 and 3 are polarizing as independent: the weight given to the draw from the uniform distribution over the rankings in the first column and first row is $P(2,\pi)\times P(3,\pi)$, and so on.

Note that we can freely vary the rank polarization of tastes for up to two options across $\Delta(\mathcal{R})$, but not three. For example, when $P(2,\pi)$ and $P(3,\pi)$ are close to one, π gives probability close to one to the first cell in the Table, and that implies both $P(1,\pi)$ and $P(4,\pi)$ must be close to zero. Intuitively, we can only have two "poles" when tastes are highly polarized.

6.3. **Leave-one-out prediction.** Our setup with $n=4$ choice options has six binary choice problems $\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\}$, four ternary choice problems $\{1,2,3\},\{1,2,4\},\{1,3,4\},\{2,3,4\}$ and one quaternary choice problem $\{1,2,3,4\}$. We perform three leave-one-out prediction exercises: fit each model to binary and ternary to predict quaternary choice data; fit each model to binary and quaternary to predict ternary choice data; and fit each model to ternary and quaternary to predict binary choice data.

6.4. **Comparison models.** We compare the predictions of the WL model to three well-known and often used discrete choice workhorse models in the literature: the classic multinomial Logit, the Nested Logit and the Covariance Probit, described in any standard discrete choice estimation textbook, e.g., Train (2009). Figure 2 shows when $n = 4$ Logit has three parameters to be estimated, WL has six, Probit has eight, and Nested Logit has ten. We review how each model is parameterized in Appendix D.

6.5. **Estimation.** Each draw from the data generating process described above produces a stochastic choice function ρ over X. We estimate the parameters θ of each model p by maximizing the log-likelihood:

$$\max_{\theta} \sum_{i,A} \rho(i, A) \ln p(i, A, \theta),$$

which we sum over each option i and each in-sample menu A. The objective function above is equivalent to drawing N data-point choices from each in-sample menu A according to ρ, writing the log-likelihood for the finite sample, dividing by N, and taking the limit $N \to \infty$. This is also equivalent to minimizing the Kullback–Leibler divergence between the true data-generating choice rule and the parameterized model. This allows us to cleanly compare the flexibility and prediction power of different models without interference from the finite-sample properties of any particular estimator.

6.6. **Computation.** We estimated the parameters of each model in the 3 leave-one-out exercises, for 9,801 polarization levels, and with 1,000 iid draws for each polarization level. Taking advantage of the symmetry between the demand from a population with polarization levels $(P(2, \pi), P(3, \pi))$ and the demand from a population with the reverse levels $(P(3, \pi), P(2, \pi))$, we reduced the number of estimations to $3 \times 4,950 \times 1,000 = 14,850,000$ for each model. Maximum likelihood was obtained with off-the-shelf optimization routines in Mathematica software. Estimating the parameters for WL, Logit and Nested Logit took roughly a week on a (vintage 2023) desktop computer. Estimating the Covariance Probit, however, took several weeks running in parallel in Washington University's RIS cluster computer and Amazon's Elastic Cloud Computing cluster. Replication code is available as an online appendix.

6.7. **Results.** For each fixed rank polarization level $P(i, \pi)$ and $P(j, \pi)$ for two options (which we label i and j throughout), we compare the empirical distribution of market share prediction errors for each model. We use several standard prediction metrics to compare the accuracy of the predictions: the median absolute error (Figures 5 and 6), mean absolute error (Appendix E), root mean square error (Appendix E) and, in addition, we simply compare the proportion of datasets in which the WL model makes a smaller prediction error than each alternative model (Figures 3 and 4).

The top row in Figure 3 shows how often the WL model makes a more accurate prediction than Logit. The left panel shows results for predicting binary choice, the middle panel shows ternary choice, and the right panel shows quaternary choice prediction. WL makes more accurate predictions (denoted by a light blue or dark blue color) for almost all rank polarization levels in the three prediction exercises. Logit only becomes equally likely or slightly more likely to make a better prediction (denoted by a white or light yellow color) in a small region towards the center of the graph, when rank polarization levels are fixed close to 1/2. The advantages of the WL over logit are the most pronounced for predicting quaternary choice.

The middle row of Figure 3 shows that the Nested Logit model performs only slightly better than Logit against the WL model. The improvement is most noticeable for quaternary choice, when the rank polarization for one option is fixed close to 1/2.

The bottom row of Figure 3 compares the WL with the covariance Probit model. WL does better at ternary choice prediction (middle panel) while probit does better at quaternary choice prediction (right panel). For binary choice, WL outperforms probit for extreme levels of polarization (close to zero or close to one for both options) and when polarization has intermediate levels for both options.

Figure 4 shows the same metric as Figure 3 along the diagonal $P(i, \pi) = P(j, \pi)$ in which options i, j have the same level of rank polarization. This allows stacking the performance for binary, ternary, and quaternary choice predictions into a single panel, facilitating their comparison. The WL performs the best for polarization levels close to zero and close to one, where it outperforms all the other models in binary and ternary choice prediction, and is only outperformed by the probit model in quaternary choice predictions.

Figure 5 depicts the median absolute prediction error in each prediction exercise for the WL (top row), Logit (second row), Nested Logit (third row) and Probit (last row). For the Logit model, the left panel (binary choice prediction) and right panel (quaternary choice prediction) show significant areas of red, denoting a median error of 0.10 or larger in market share prediction. Nested Logit again shows a moderate improvement over Logit. WL and Probit have more modest errors across the three prediction exercises and across almost all levels of polarization.

To more clearly see the comparison across models, Figure 5 plots median absolute prediction errors across the diagonal $P(i, \pi) = P(j, \pi)$ in which options i, j have the same level of rank polarization. Logit and Nested Logit only display the lowest errors in binary choice prediction, and for a narrow range of polarization levels around 1/2. In all other cases, either WL or Probit have smaller prediction errors.

WL has the lowest median absolute prediction errors throughout in the case of predicting ternary choice data, shown in the middle panel of Figure 5. Probit has the lowest prediction errors throughout for quaternary choice (bottom panel), though WL is a very close contender in that case. It is noteworthy that for polarization levels of exactly 1/2 all four models have the same median absolute errors.

In sum, these results show that rank polarization is an important measure to track the prediction performance of the WL model versus the alternatives. When polarization is close to the middle value of 1/2, all models become close in prediction performance including the simplest classic Logit model. Away from this case, we have two sets of results. First, classic logit becomes heavily disadvantaged once polarization moves away from 1/2. Nested Logit improves prediction over Logit somewhat, but the WL and the Probit model clearly perform in a class of their own. Second, the WL model predicts best, relative to probit, in specific instances: i) predicting binary choice with extreme degrees of rank polarization; and ii) predicting binary choice with rank polarization levels close to 1/2; and iii) predicting market shares for intermediate choice set size (i.e. ternary sets). In other situations, Probit does better, though the WL comes close.

While the WL does not clearly dominate the Probit model, and performs slightly worse in some cases, it could prove a better model to use given its smaller number of parameters, closed form choice probabilities, much lower computational burden, and easy interpretation.

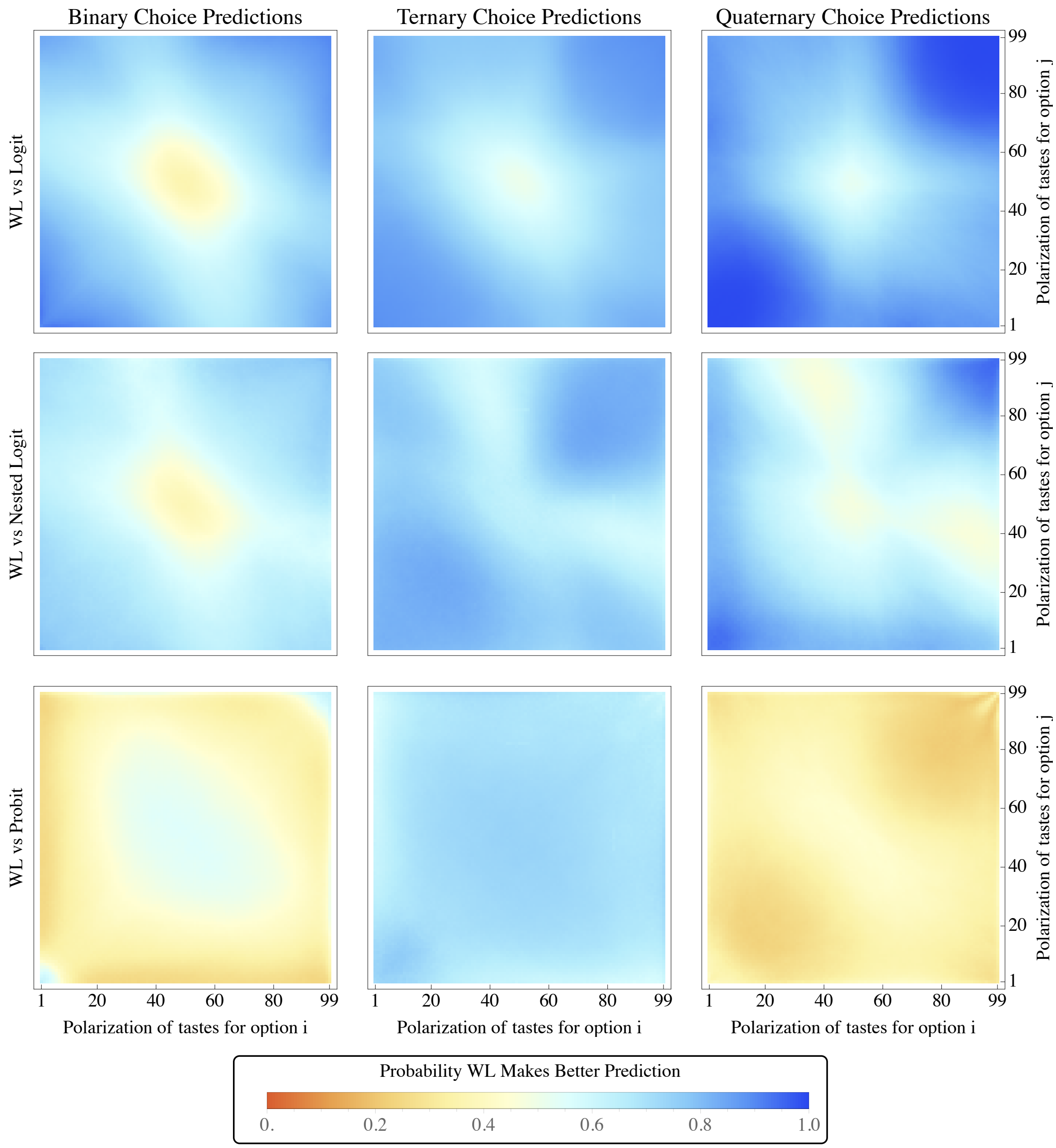

FIGURE 3. Leave-one-out prediction comparison between the WL model and Logit, Nested Logit and Probit. The axes display the fixed level of rank polarization for two different options. The color scale represents the proportion of the simulated demand systems in which the WL model makes a closer market share prediction than the competing model.

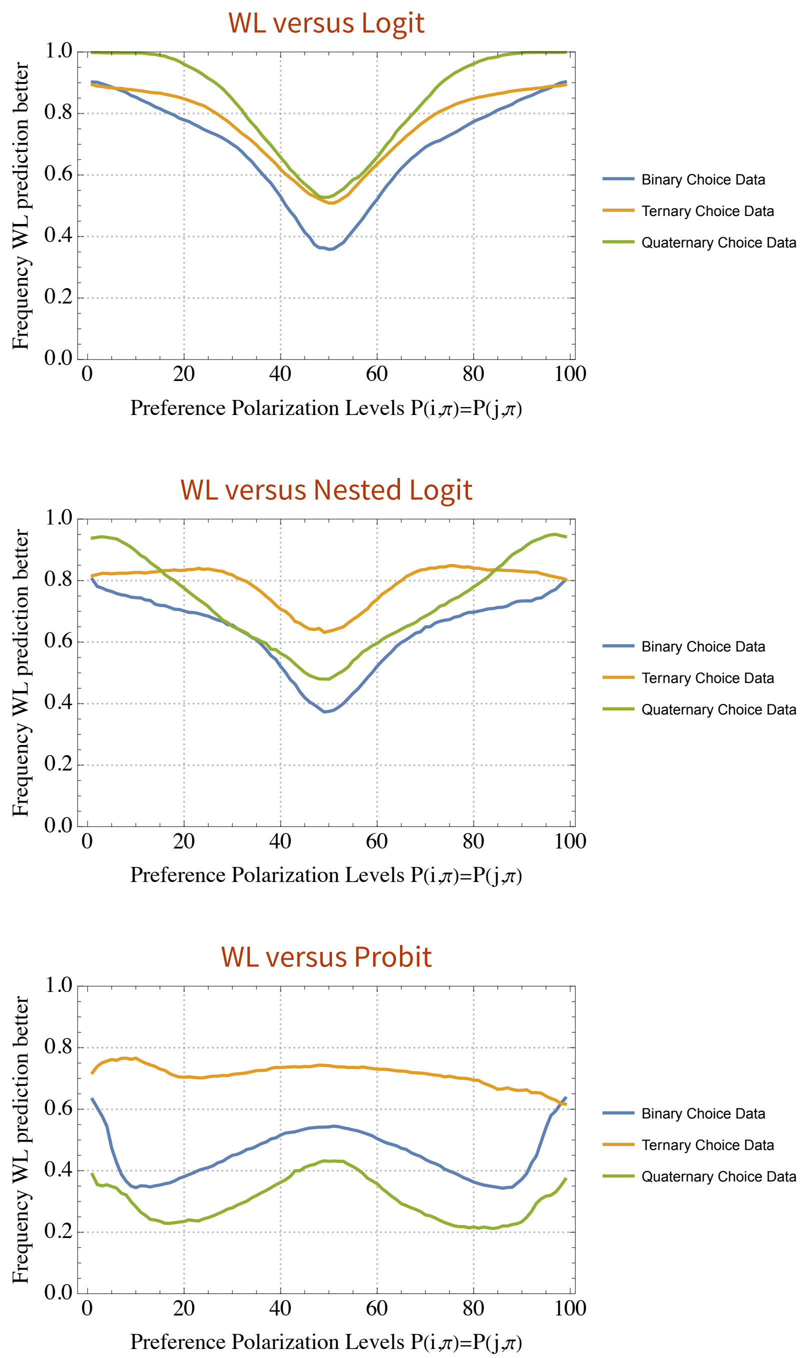

FIGURE 4. Leave-one-out prediction comparison between the WL model and Logit, Nested Logit and Probit. The horizontal axis displays the fixed level of rank polarization set equal for two options. The vertical axis displays the proportion of the simulated demand systems in which the WL model makes a closer market share prediction than the competing model.

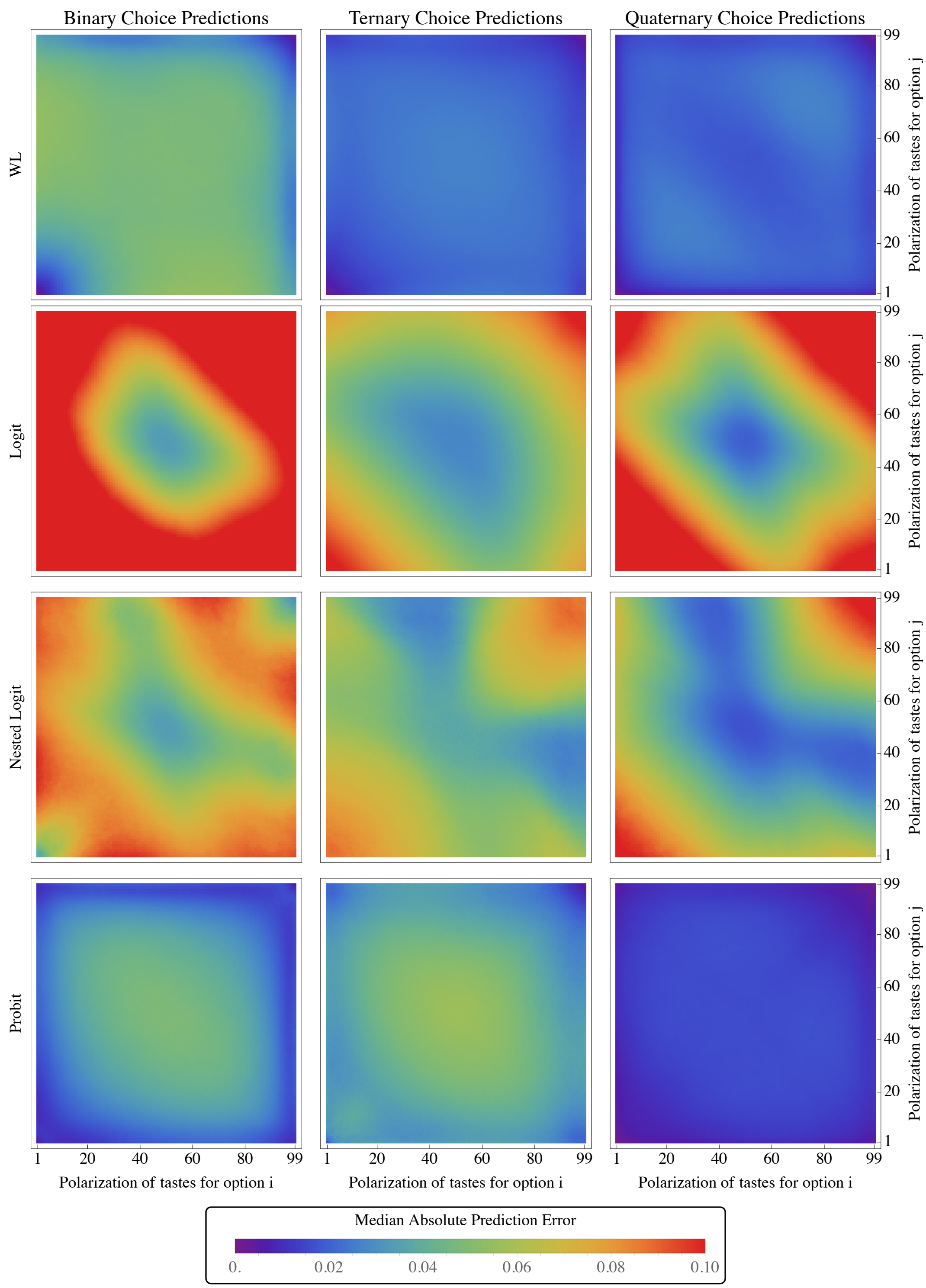

Figure 5. Median absolute prediction error in leave-one-out prediction for WL, Logit, Nested Logit and Probit. The axes display the fixed level of rank polarization for two different options.

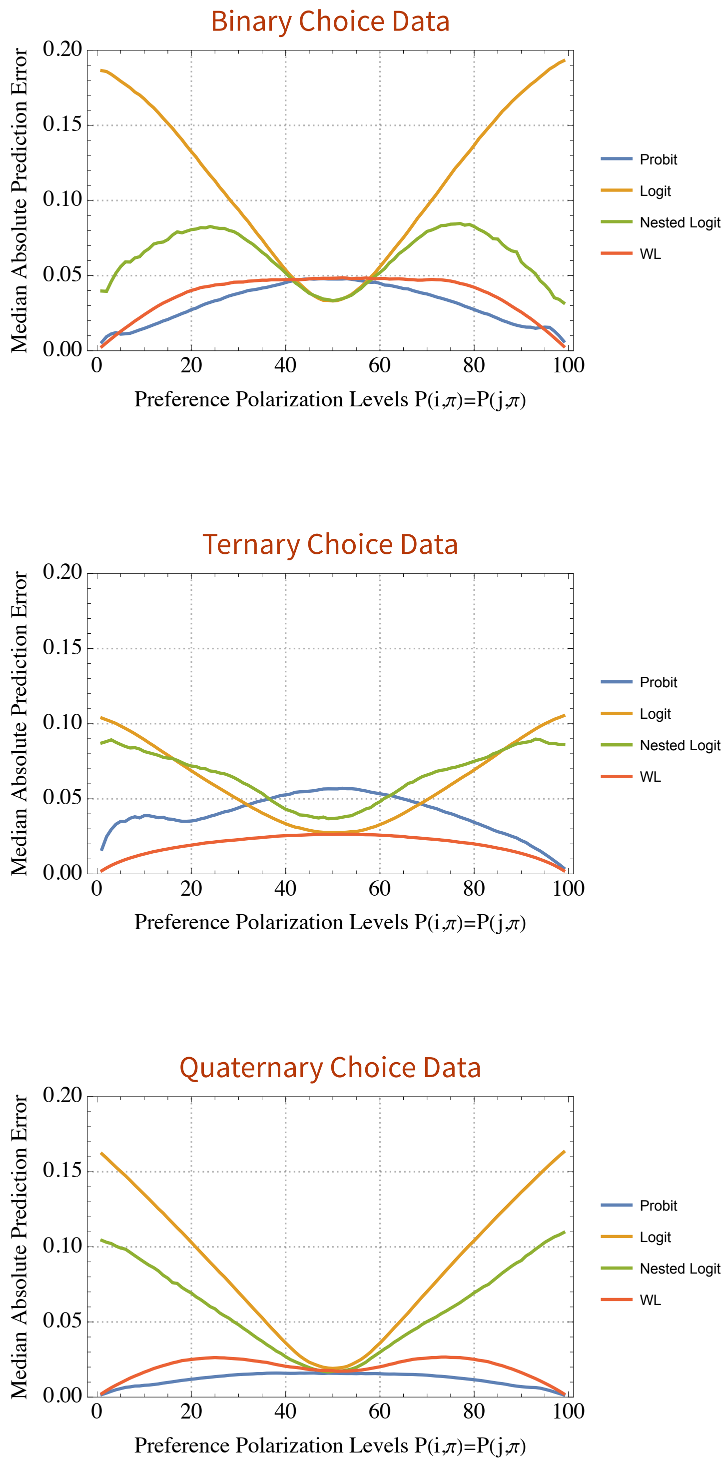

FIGURE 6. Median absolute prediction error in leave-one-out prediction for WL, Logit, Nested Logit and Probit. The horizontal axis displays the fixed and equal level of rank polarization for two options.

7. Conclusion

This paper has introduced a new model of stochastic choice: the weighted linear model of discrete choice. The choice probabilities of any given product depend on two dimensions, the utility of an item, and the salience of an item. The model sits at the intersection of the classic models of random utility and models of deliberate stochastic choice.

The weighted linear model represents a generalization of the classic model of Luce, and, as such, provides more explanatory power. It is closely related to well-known models in which demand is linear in the utility differences between products. The importance of salience for choice has been recognized by researchers (not just in economics but also in marketing, cognitive science, and psychology). Our model presents one way of incorporating such considerations. The WL also overcomes many potentially concerning implications of other random/discrete choice approaches used to capture consumer behavior in market settings. At the same time, it is quite tractable. The model lends itself naturally to describing consumers who experience some friction in being able to choose the best item (whether they be physical or attentional frictions). When used in models of strategic firm interactions, it can generate intuitive closed-form solutions that shed light on advertising, quality choice, and the number of firms serving a market.

Our hope is that the flexibility of our model, its intuitive approach to choice as depending on both utility and salience, and its tractability can help economists better understand market interactions between firms and consumers. In particular, the fact that our model allows for intuitive empirical patterns, such as flexible patterns of cross-price substitution patterns, or the existence of dominant market shares in large markets, can lead to new insights in many markets where these kinds of behavior need to be captured. We also think that empirical work geared towards understanding which kind of product attributes affect utility versus salience (e.g., does advertising increase the perceived utility of an item versus changing the cost of choosing it) could help shed useful insights into the structure of consumer choice.

References

Ahumada, A., & Ulku, L. (2018). Luce rule with limited consideration. *Mathematical Social Sciences*, *93*, 52–56.

Allen, R., & Rehbeck, J. (2019). *Revealed stochastic choice with attributes* [working paper]. working paper.

Armstrong, M., & Vickers, J. (2015). Which demand systems can be generated by discrete choice? *Journal of Economic Theory*, *158*, 293–307.

Bagwell, K. (2007). The economic analysis of advertising (M. Armstrong & R. H. Porter, Eds.). In M. Armstrong & R. H. Porter (Eds.), *Handbook of industrial organization*, Elsevier.

Bagwell, K., & Ramey, G. (1994). Coordination economies, advertising, and search behavior in retail markets. *The American Economic Review*, *84*(3), 498–517.

Barbera, S., & Pattanaik, P. K. (1986). Falmagne and the Rationalizability of Stochastic Choices in Terms of Random Orderings. *Econometrica*, *54*(3), 707.

Ben-Akiva, M., & Lerman, S. R. (1985). *Discrete choice analysis: Theory and application to travel demand* (Vol. 2). The MIT Press google schola.

Benkard, C. L., & Bajari, P. (2001). *Discrete choice models as structural models of demand: Some economic implications of common approaches* (tech. rep.).

Berry, S. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 242–262.

Berry, S., & Pakes, A. (2007). The pure characteristics demand model. *International Economic Review*, *48*(4), 1193–1225.

Bhatnagar, R., & Orquin, J. L. (2022). A meta-analysis on the effect of visual attention on choice. *Journal of Experimental Psychology: General*.

Bialkova, S., & van Trijp, H. C. (2011). An efficient methodology for assessing attention to and effect of nutrition information displayed front-of-pack. *Food Quality and Preference*, *22*(6), 592–601.

Block, H. D., & Marschak, J. (1959). *Random orderings and stochastic theories of response* (tech. rep.). Cowles Foundation, Yale University.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2013). Salience and consumer choice. *Journal of Political Economy*, *121*(5), 803–843.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2022). Salience. *Annual Review of Economics*, *14*, 521–544.

Brady, R. L., & Rehbeck, J. (2016). Menu-dependent stochastic feasibility. *Econometrica*, *84*(3), 1203–1223.

Butters, G. R. (1977). Equilibrium distributions of sales and advertising prices. *The Review of Economic Studies*, *44*(3), 465–491.

Cattaneo, M. D., Ma, X., Masatlioglu, Y., & Suleymanov, E. (2020). A Random Attention Model. *Journal of Political Economy*, *128*(7), 2796–2836.

Cerreia-Vioglio, S., Dillenberger, D., Ortoleva, P., & Riella, G. (2019). Deliberately stochastic. *American Economic Review*, *109*(7), 2425–45.

Chandon, P., Hutchinson, J. W., Bradlow, E. T., & Young, S. H. (2009). Does in-store marketing work? effects of the number and position of shelf facings on brand attention and evaluation at the point of purchase. *Journal of marketing*, *73*(6), 1–17.

Chetty, R., Looney, A., & Kroft, K. (2009). Salience and taxation: Theory and evidence. *American economic review*, *99*(4), 1145–1177.

Chew, S. H., Epstein, L. G., & Segal, U. (1991). Mixture symmetry and quadratic utility. *Econometrica*, *59*(1), 139–163.

Chew, S. H., Epstein, L. G., & Segal, U. (1994). The projective independence axiom. *Economic Theory*, *4*(2), 189–215.

Choné, P., & Linnemer, L. (2020). Linear demand systems for differentiated goods: Overview and user's guide. *International Journal of Industrial Organization*, 102663.

Dai, J., Cone, J., & Moher, J. (2020). Perceptual salience influences food choices independently of health and taste preferences. *Cognitive research: principles and implications*, *5*(1), 1–13.

Debreu, G. (1960). Review of individual choice behavior by rd luce. *American Economic Review*, *50*(1), 186–188.

Dixit, A. K., & Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *The American Economic Review*, *67*(3), 297–308.

Echenique, F., & Saito, K. (2018). General luce model. *Economic Theory*, 1–16.

Ehrenberg, A., Barnard, N., & Scriven, J. (1997). Differentiation or salience. *Journal of Advertising Research*, *37*(6), 7–15.

Ellis, A., & Masatlioglu, Y. (2021). Choice with Endogenous Categorization. *The Review of Economic Studies*.

Falmagne, J.-C. (1978). A representation theorem for finite random scale systems. *Journal of Mathematical Psychology*, *18*(1), 52–72.

Faro, J. H. (2023). The luce model with replicas. *Journal of Economic Theory*, *208*, 105596.

Fechner, G. T. (1860). *Elemente der psychophysik* (Vol. 2). Breitkopf u. Härtel.

Fudenberg, D., Iijima, R., & Strzalecki, T. (2014). *Stochastic choice and revealed perturbed utility* [working paper]. working paper.

Fudenberg, D., Iijima, R., & Strzalecki, T. (2015). Stochastic choice and revealed perturbed utility. *Econometrica*, *83*(6), 2371–2409.

Gluth, S., Spektor, M. S., & Rieskamp, J. (2018). Value-based attentional capture affects multi-alternative decision making. *Elife*, *7*, e39659.

Grossman, G. M., & Shapiro, C. (1984). Informative advertising with differentiated products. *The Review of Economic Studies*, *51*(1), 63–81.

Halff, H. M. (1976). Choice theories for differentially comparable alternatives. *Journal of Mathematical Psychology*, *14*(3), 244–246.

He, J., & Natenzon, P. (2024). Moderate utility. *American Economic Review: Insights*.

Horan, S. (2021). Stochastic semi-orders. *Journal of Economic Theory*, *192*, 105171.

Hoyer, W. D., & Brown, S. P. (1990). Effects of brand awareness on choice for a common, repeat-purchase product. *Journal of consumer research*, *17*(2), 141–148.

Jaffe, S., & Kominers, S. D. (2012). Discrete choice cannot generate demand that is additively separable in own price. *Economics Letters*, *116*(1), 129–132.

Jaffe, S., & Weyl, E. G. (2010). Linear demand systems are inconsistent with discrete choice. *The BE Journal of Theoretical Economics*.

Janiszewski, C., Kuo, A., & Tavassoli, N. T. (2013). The influence of selective attention and inattention to products on subsequent choice. *Journal of Consumer Research*, *39*(6), 1258–1274.

Kovach, M., & Tserenjigmid, G. (2022). Behavioral foundations of nested stochastic choice and nested logit. *Journal of Political Economy*, *130*(9), 2411–2461.

Krajbich, I. (2019). Accounting for attention in sequential sampling models of decision making. *Current opinion in psychology*, *29*, 6–11.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, *13*(10), 1292–1298.

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857.

Krajbich, I., Yang, X., Retzler, C., Ratcliff, R., & Philiastides, M. G. (2024). Attention to brand labels affects, and is affected by, evaluations of product attractiveness. *Frontiers in Behavioral Economics*, *2*, 1274815.

Li, X., & Camerer, C. F. (2022). Predictable effects of visual salience in experimental decisions and games. *The Quarterly Journal of Economics*, *137*(3), 1849–1900.

Luce, R. D. (1959). *Individual choice behavior*. Wiley, New York.

Macdonald, E. K., & Sharp, B. M. (2000). Brand awareness effects on consumer decision making for a common, repeat purchase product: A replication. *Journal of business research*, *48*(1), 5–15.

Machina, M. J. (1985). Stochastic choice functions generated from deterministic preferences over lotteries. *The Economic Journal*, *95*(379), 575–594.

Mangasarian, O. L. (1994). *Nonlinear programming*. SIAM.

Manzini, P., & Mariotti, M. (2014). Stochastic choice and consideration sets. *Econometrica*, *82*(3), 1153–1176.

Matějka, F., & McKay, A. (2015). Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model. *American Economic Review*, *105*(1), 272–298.

Mellers, B. A., Chang, S.-j., Birnbaum, M. H., & Ordonez, L. D. (1992). Preferences, prices, and ratings in risky decision making. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(2), 347.

Miller, S., & Berry, L. (1998). Brand salience versus brand image: Two theories of advertising effectiveness. *Journal of Advertising Research*, *38*(5), 77–78.

Milosavljevic, M., Navalpakkam, V., Koch, C., & Rangel, A. (2012). Relative visual saliency differences induce sizable bias in consumer choice. *Journal of Consumer Psychology*, *22*(1), 67–74.

Moran, W. T. (1990). Brand presence and the perceptual frame. *Journal of Advertising Research*.

Navalpakkam, V., Kumar, R., Li, L., & Sivakumar, D. (2012). Attention and selection in online choice tasks, In *International conference on user modeling, adaptation, and personalization*.

Nedungadi, P. (1990). Recall and consumer consideration sets: Influencing choice without altering brand evaluations. *Journal of consumer research*, *17*(3), 263–276.

Orquin, J. L., & Lagerkvist, C. J. (2015). Effects of salience are both short-and long-lived. *Acta psychologica*, *160*, 69–76.

Orquin, J. L., & Loose, S. M. (2013). Attention and choice: A review on eye movements in decision making. *Acta psychologica*, *144*(1), 190–206.

Peschel, A. O., Orquin, J. L., & Loose, S. M. (2019). Increasing consumers' attention capture and food choice through bottom-up effects. *Appetite*, *132*, 1–7.

Reutskaja, E., Nagel, R., Camerer, C. F., & Rangel, A. (2011). Search dynamics in consumer choice under time pressure: An eye-tracking study. *American Economic Review*, *101*(2), 900–926.

Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. (2006). Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature*, *44*(3), 631–661.

Robert, J., & Stahl, D. O. (1993). Informative price advertising in a sequential search model. *Econometrica*, *61*(3), 657–686.

Romaniuk, J., & Sharp, B. (2003). Brand salience and customer defection in subscription markets. *Journal of Marketing Management*, *19*(1-2), 25–44.

Romaniuk, J., & Sharp, B. (2004). Conceptualizing and measuring brand salience. *Marketing theory*, *4*(4), 327–342.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of political economy*, *82*(1), 34–55.

Rosenthal, R. W. (1989). A bounded-rationality approach to the study of noncooperative games. *International Journal of Game Theory*, *18*, 273–292.

Shubik, M., & Levitan, R. (1980). Market structure and innovation. *New York: John Wileyand Sons*.

Singh, N., & Vives, X. (1984). Price and quantity competition in a differentiated duopoly. *The Rand journal of economics*, 546–554.

Smith, S. M., & Krajbich, I. (2019). Gaze amplifies value in decision making. *Psychological science*, *30*(1), 116–128.

Spence, M. (1976). Product selection, fixed costs, and monopolistic competition. *The Review of Economic Studies*, *43*(2), 217–235.

Sutherland, M., & Galloway, J. (1981). Role of advertising: Persuasion or agenda setting. *Journal of Advertising Research*, *21*(5), 25–29.

Syverson, C. (2019). Macroeconomics and market power: Context, implications, and open questions. *Journal of Economic Perspectives*, *33*(3), 23–43.

Thomas, A. W., Molter, F., & Krajbich, I. (2021). Uncovering the computational mechanisms underlying many-alternative choice. *Elife*, *10*, e57012.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, *34*(4), 273.

Towal, R. B., Mormann, M., & Koch, C. (2013). Simultaneous modeling of visual saliency and value computation improves predictions of economic choice. *Proceedings of the National Academy of Sciences*, *110*(40), E3858–E3867.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge University Press.

Turansick, C. (2021). *Identification in the random utility model* [working paper]. working paper.

Van der Lans, R., Pieters, R., & Wedel, M. (2008). Research note—competitive brand salience. *Marketing Science*, *27*(5), 922–931.

Vieceli, J., & Shaw, R. N. (2010). Brand salience for fast-moving consumer goods: An empirically based model. *Journal of marketing management*, *26*(13-14), 1218–1238.

Voorneveld, M. (2006). Probabilistic choice in games: Properties of rosenthal's t-solutions. *International Journal of Game Theory*, *34*, 105–121.

Wedel, M., & Pieters, R. (2000). Eye fixations on advertisements and memory for brands: A model and findings. *Marketing science*, *19*(4), 297–312.

Wegge, L. L. (1968). The demand curves from a quadratic utility indicator. *The Review of Economic Studies*, *35*(2), 209–224.

Weingarten, E., & Hutchinson, J. (2017). Perceptual and cognitive salience and their effects on product valuations. *ACR North American Advances*.

Yi, Y. (1990). The effects of contextual priming in print advertisements. *Journal of Consumer Research*, *17*(2), 215–222.

Zhang, Z., Wang, S., Good, M., Hristova, S., Kayser, A. S., & Hsu, M. (2021). Retrieval-constrained valuation: Toward prediction of open-ended decisions. *Proceedings of the National Academy of Sciences*, *118*(20), e2022685118.

Appendix A. Proofs

A.1. Proof of Proposition 1.

Proof. Our goal is to establish a method whereby, given u, m, and $S \subseteq X$, we can find the condition to guarantee all alternatives chosen with positive probability in the solution to the problem (1). We start by proving a useful lemma.

Lemma 1. *$\Lambda(S) \gtreqless u(x)$ if and only if $\Lambda(S \setminus x) \gtreqless \Lambda(S)$.*

Proof.

$$\frac{\sum_{y\in S} u(y)m(y) - 1}{m(S)} \gtreqless u(x)$$

$$\sum_{y\in S} u(y)m(y) - 1 \gtreqless u(x)m(S)$$

$$m(x)\left(1 - \sum_{y\in S} u(y)m(y)\right) \lesseqgtr -u(x)m(x)m(S)$$

$$\left(\sum_{y\in S} u(y)m(y) - 1\right)(m(S) - m(x)) \lesseqgtr m(S)\left(\sum_{y\in S} u(y)m(y) - u(x)m(x) - 1\right)$$

$$\left(\sum_{y\in S} u(y)m(y) - 1\right) m(S\setminus x) \lesseqgtr m(S)\left(\sum_{y\in S\setminus x} u(y)m(y) - 1\right)$$

$$\Lambda(S) \lesseqgtr \Lambda(S\setminus x)$$

□

Since $\rho(a|X) = m(x)(u(x) - \Lambda(X))$, we have $u(x) > \Lambda(X)$ for all x given $\rho(x|X) > 0$. To show the other direction, Lemma 1 shows that if $u(x)$ is greater than the Lagrange multiplier for any set S, then the elimination of x from the set lowers the Lagrange multiplier for $S \setminus x$. Hence if we guarantee an interior solution for the grand set ($u(x) > \Lambda(X)$ for all x), we have an interior solution for all sets. □

A.2. Proof of Theorem 2.

Proof. Observe that for any S,

$$\operatorname*{argmax}_{p\in\Delta(S)} \sum_{x\in S} \left(\rho(x)(au(x)+b) - \frac{a}{2m(x)}\rho(x)^2\right) = \operatorname*{argmax}_{p\in\Delta(S)} a[\sum_{x\in S} \left(\rho(x)u(x) - \frac{1}{2m(x)}\rho(x)^2\right)] + b,$$

the result follows as strictly monotone transformations of objective functions preserve maxima.

To show the other way, take two different representations of ρ: (u,m) and (v,n). For any arbitrary representation ρ, (u,m), we can find $a_{(u,m)} > 0$ and $b_{(u,m)}$ such that $u' = a_{(u,m)}u + b_{(u,m)}$ and $m' = \frac{m}{a_{(u,m)}}$ such that

$$\sum_{x\in X}(u'(x)m'(x)) = 1 \text{ and } \sum_{x\in X} m'(x) = 1$$

For such representations, we must have

$$\rho(x|X) = \frac{m'(x)}{m'(X)} + m'(x)[u'(x) - \bar{u}'_{m'}(X)] = m'(x) + m'(x)[u'(x) - 1] = m'(x)u'(x)$$

Hence, $u'(x)m'(x) = v'(x)n'(x)$ for all x, call this condition $(*)$.

Note that $\rho(x|X) > 0$ for all x by the condition $(u(x) > \Lambda(X)$ for all $x)$. Hence $u'(x), v'(x) \neq 0$ for all x.

Moreover, since (u',m') and (v',n') represents ρ, we must have

$$\frac{m'(x)}{m'(S)} + m'(x)[u'(x) - \bar{u}'_{m'}(S)] = \frac{n'(x)}{n'(S)} + n'(x)[v'(x) - \bar{v}'_{n'}(S)]$$

Since $(*)$,

$$\frac{m'(x)}{m'(S)} - m'(x)\bar{u}'_{m'}(S) = \frac{n'(x)}{n'(S)} - n'(x)\bar{v}'_{n'}(S)$$

$$\frac{m'(x)}{m'(S)}[1 - \sum_{x\in S} u'(x)m'(x)] = \frac{n'(x)}{n'(S)}[1 - \sum_{x\in S} v'(x)n'(x)]$$

Since $u', v' \neq 0$ and $m', n' > 0$, then $\sum_{x\in X\setminus y}(u'(x)m'(x)) \neq 1$. Let S_z denote $X \setminus z$. Then we must have

$$\frac{m'(x)}{m'(S_y)} = \frac{n'(x)}{n'(S_y)} \text{ for all } x \in S_y$$

Therefore,

$$\frac{m'(x)}{n'(x)} = \frac{1 - m'(y)}{1 - n'(y)}$$

which implies

$$m'(x) - n'(x) = m'(x)n'(y) - n'(x)m'(y)$$

Similarly, we have

$$\frac{m'(y)}{n'(y)} = \frac{1 - m'(x)}{1 - n'(x)}$$

which yields

$$m'(y) - n'(y) = m'(x)n'(y) - n'(x)m'(y)$$

Therefore, we have $m'(x) - n'(x) = m'(y) - n'(y)$ for all x, y. If $m'(x) - n'(x) \neq 0$ then we get a contradiction, so $m' = n'$. This proves that $u' = v'$. This implies that $m(x) = \frac{a_{(u,m)}}{a_{(v,n)}} n(x)$ and $u(x) = \frac{a_{(v,n)}}{a_{(u,m)}} v(x) + \frac{b_{(v,n)} - b_{(u,m)}}{a_{(u,m)}}$. By letting $a = \frac{a_{(v,n)}}{a_{(u,m)}}$ and $b = \frac{b_{(v,n)} - b_{(u,m)}}{a_{(u,m)}}$, we get the desired result. □

A.3. **Proof of Theorem 1.**

Proof. Let $\mathcal{D}$ be the domain of stochastic choice functions containing all menus with size 2 and 3. Necessity of the axioms is straightforward.

For sufficiency, select any $y^* \in X$ and define binary sets $B_z := \{z, y^*\}$ and ternary sets $T_{zz'} := \{z, z', y^*\}$, which belong to our domain. We use these sets to define

$$m(x) := \frac{d(x|B_x, T_{xz})}{d(y^*|B_x, T_{xz})} = \frac{d(x|\{x, y^*\}, \{x, y^*, z\})}{d(y^*|\{x, y^*\}, \{x, y^*, z\})}$$

for some z different from x and y^*. First note that Axiom 2 implies that both the denominator and the numerator are strictly positive. Hence, the ratio is well-defined and positive. Axiom 3 implies that $d(x|S, T)d(y|S, T') = d(y|S, T)d(x|S, T')$ for all T and T' such that $x, y \in S \cap T \cap T'$, hence the choice of z does not alter the ratio. Therefore, these observations guarantee that m is well-defined and strictly positive for any x. Notice that $m(y^*) = 1$. Since m is non-zero, we can define $c(x) := \frac{1}{m(x)}$.

Claim 1. *For all $S \neq T \in \mathcal{D}$ and $x, y \in S \cap T$, $c(x)d(x|S, T) = c(y)d(y|S, T)$.*

Proof. Consider distinct $S, T \in \mathcal{D}$ and $x, y \in S \cap T$ distinct from y^*. By Axiom 3, $d(x|S, T)d(y|B_y, T_{yz})d(y^*|B_x, T_{xy}) = d(y|S, T)d(y^*|B_y, T_{yz})d(x|B_x, T_{xy})$. Hence, by Axiom 2, we get the desired result: $c(x)d(x|S, T) = c(y)d(y|S, T)$. □

We now recursively define λ for every element in $\mathcal{D}$ by the following formula

$$\Lambda(S) := c(x)d(x|T, S) + \Lambda(T)$$

Given S, to define $\Lambda(S)$, we first must find a set T such that $S \cap T \neq \emptyset$. Hence, in the first step, we define λ for all S with y^* as a member. Denote this set by $\mathcal{A}_0 := \{S \in \mathcal{D} \mid y^* \in S \neq \emptyset\}$. Then $S \in \mathcal{A}_0$, $\Lambda(S)$ is defined as $1 - \rho(y^*|S)$. In the second step, we define λ for the rest of the subsets, denoted by $\mathcal{A}_1 := \{S \in \mathcal{D} \mid y^* \notin S \neq \emptyset\}$. Take $S \in \mathcal{A}_1$ and find $T \in \mathcal{A}_0$ such that $S \cap T \neq \emptyset$, and define $\Lambda(S) := c(x)d(x|T, S) + \Lambda(T)$ for some x in $S \cap T$.

Claim 2. *λ is well-defined and $\Lambda(S) + c(x)\rho(x|S) = \Lambda(T) + c(x)\rho(x|T)$ for all $x \in S \cap T$.*

Proof. We need to show that λ is well-defined (i.e., $\Lambda(S)$ is independent of the choice of x and T). By Claim 1, $\Lambda(S)$ is independent of the choice of x for a given T. Now we establish independence of T. Take two overlapping sets S and T. If $S, T \in \mathcal{A}_0$, by definition we get $\Lambda(S) = c(y^*)d(y^*|T,S) + \Lambda(T)$ (note that $c(y^*) = 1$). By Claim 1,

(7) $$\Lambda(S) = c(x)d(x|T,S) + \Lambda(T) \text{ for all } x \in S \cap T$$

This establishes that the equation holds on $\mathcal{A}_0$. Now take $S \in \mathcal{A}_1$ and assume that $T, T' \in \mathcal{A}_0$ such that $x \in S \cap T, x' \in S \cap T'$. Then

$$\begin{aligned}
\Lambda(S) &= c(x)d(x|T,S) + \Lambda(T) \\
&= c(x)d(x|T,S) + c(x)d(x|\{x,x',y^*\},T) + \lambda(\{x,x',y^*\}) \text{ by Eq (7)} \\
&= c(x)d(x|\{x,x',y^*\},S) + \lambda(\{x,x',y^*\}) \\
&= c(x')d(x'|\{x,x',y^*\},S) + \lambda(\{x,x',y^*\}) \text{ by Claim 1} \\
&= c(x')d(x'|T',S) + c(x')d(x'|\{x,x',y^*\},T') + \lambda(\{x,x',y^*\}) \\
&= c(x')d(x'|T',S) + \lambda(T') \text{ by Eq (7)}
\end{aligned}$$

This establishes that λ is well-defined on $\mathcal{A}_0 \cup \mathcal{A}_1$. Finally, we must show that $\Lambda(S) = c(x)d(x|T,S) + \Lambda(T)$ holds for $S, T \in \mathcal{A}_1$. Let x be an alternative in $S \cap T$. Since $B_x \in \mathcal{A}_0$, we have $\Lambda(S) = c(x)d(x|B_x,S) + \lambda(B_x)$ and $\Lambda(T) = c(x)d(x|B_x,T) + \lambda(B_x)$. Subtracting these two equations yields $\Lambda(S) - \Lambda(T) = c(x)d(x|B_x,S) - c(x)d(x|B_x,T) = c(x)d(x|T,S)$. □

Now define $u(x) := c(x)\rho(x|S) + \Lambda(S)$. u is well-defined by Claim 2. Note that $\Lambda(S) < u(x)$ for all $x \in S \in \mathcal{D}$ since $c(x)\rho(x|S) > 0$. Then we have $\rho(x|S) = m(x)[u(x) - \Lambda(S)]$ where $\Lambda(S) = \bar{u}_m(S) - \frac{1}{m(S)}$. Therefore, (u,m) is a WL representation of ρ. □

A.4. **Proof of Proposition 2.**

Proof. Let (u,m) be a WL representation of ρ. First, we normalize u by subtracting the shadow price for the grand set. Theorem 2 implies that $(\tilde{u},m)$ is also a WL representation of ρ, where $\tilde{u} = u - \Lambda(X)$. This normalization implies that $\sum[\tilde{u}(x)m(x)]$ is equal to 1 and $\rho(x|X) = \tilde{u}(x)m(x)$, which we define as $w(x)$. Then we have

$$\rho(x|S) = w(x) + (1 - w(S))\frac{m(x)}{m(S)}.$$

Let us write the BM polynomial for this model: $\sum_{S':S\subseteq S'}(-1)^{|S'\setminus S|}\rho(x|S)$. Nonnegativity is obvious for $S = X$. Assume $S \neq X$, then the BM polynomial for (x,S) is

$$\sum_{S':S\subseteq S'}(-1)^{|S'\setminus S|}\frac{w(X\setminus S')}{m(S')}m(x).$$

Now let us rewrite this, indexing by elements not in S. Then we get

$$\sum_{z\notin S} w(z)\left(\sum_{S\subseteq S'\subseteq X\setminus\{z\}} \frac{m(x)}{m(S')}(-1)^{|S'\setminus S|}\right)$$

Observe the term in parentheses for a given z:

$$\sum_{S\subseteq S'\subseteq X\setminus\{z\}} \frac{m(x)}{m(S')}(-1)^{|S'\setminus S|}$$

Observe that this expression is exactly the BM-polynomial for (x, S), given a Luce rule defined on $X\setminus\{z\}$ with weights m. Therefore, this expression is non-negative given that Luce has RUM representation. Since $w(z) > 0$ for all $z \in X$, we obtain

$$\sum_{S':S\subseteq S'} \frac{w(X\setminus S')}{m(S')} m(x)(-1)^{|S'\setminus S|} \geq 0$$

Thus the WL model is RUM. □

A.5. **Proof of Proposition ??.**

Proof. Because $z = \frac{\tilde{u}_1-\tilde{u}_2+\frac{1}{m_2}}{\frac{1}{m_1}+\frac{1}{m_2}}$, then $\frac{1}{m_1} = \frac{(1-z)\frac{1}{m_2}+\tilde{u}_1-\tilde{u}_2}{z}$.

We can substitute this back into our original equilibrium pricing equations, giving $p_1 = \frac{\frac{1}{m_2}+\tilde{u}_1-\tilde{u}_2-\frac{z}{m_2}+3kz+\tilde{u}_1 z-\tilde{u}_2\sigma_1}{3z}$ and $p_2 = \frac{\frac{2}{m_2}+2\tilde{u}_1-2\tilde{u}_2-\frac{z}{m_2}+3kz-\tilde{u}_1 z+\tilde{u}_2 z}{3z}$. The derivative of p_1 and p_2 with respect to $\tilde{u}_1$ is then $\frac{1+z}{3z}$ and $\frac{2-z}{3z}$ respectively. Notice that these are both positive since $z \in [0, 1]$. Moreover, the difference between them is $\frac{2z-1}{3z}$ which is positive if and only if $z \geq 0.5$.

Profits, as a function of z for Firms 1 and 2 are $\frac{(\frac{1}{m_2}+\tilde{u}_1-\tilde{u}_2)(1+z)^2}{9z}$ and $\frac{(\frac{1}{m_2}+\tilde{u}_1-\tilde{u}_2)(z-2)^2}{9z}$ respectively. The difference in profits $\pi_1 - \pi_2$ is then $\frac{(\frac{1}{m_2}+\tilde{u}_i-\tilde{u}_2)(-1+2z)}{3z}$

Notice that, not surprisingly, firm 1 has higher profits than firm 2 only if z is greater than 1. More interestingly, notice that the derivative of this with respect to $\tilde{u}_1$ is $\frac{2z-1}{3z}$, from which the result follows. □

A.6. **Proof of Proposition 3.**

Proof. We prove this claim by contradiction. Assume $\rho(x|S) \geq \rho(y|S) > 0$ and $\rho(x|S\cup T) < \rho(y|S\cup T)$. Then we have

$$\frac{u(x)-\Lambda(S)}{u(y)-\Lambda(S)} \geq \frac{m(y)}{m(x)} > \frac{u(x)-\Lambda(S\cup T)}{u(y)-\Lambda(S\cup T)}$$

Then

$$u(y)(\Lambda(S\cup T)-\Lambda(S)) > u(x)(\Lambda(S\cup T)-\Lambda(S))$$

Since Λ is increasing, then $\Lambda(S \cup T) - \Lambda(S) > 0$, which implies that $u(y) > u(x)$, a contradiction. □

A.7. **Proof of Proposition 4.**

Proof. Since $\rho(x|S) > 0$, we have $u(x) > \Lambda(S)$. Since $\Lambda(T_n)$ approaches to $u(x)$ from below, as n goes to infinity, for all $y \in S$, $\rho(y|T_n) = m(y)[u(y) - \Lambda(T_n)] > m(y)[u(y) - u(x)] > 0$ if and only if $u(y) > u(x)$. □

Appendix B. Allowing Zero Probabilities

Our goal is to establish a method whereby, given u, m, and $S \subseteq X$, we can recover the alternatives chosen with positive probability in the solution to the problem (1).

Observe first that Λ can be explicitly calculated by summing across the alternatives that are chosen with positive probabilities. We first denote the set of alternatives chosen with positive probability in S by $\text{supp}_\rho(S)$. That is, $\text{supp}_\rho(S) := \{x \in S|\ \rho(x|S) > 0\}$. Summing across elements of $\text{supp}_\rho(S)$, we must have $\Lambda(\text{supp}_\rho(S)) = \frac{\sum_{y \in \text{supp}_\rho(S)} u(y)m(y) - 1}{\sum_{y \in \text{supp}_\rho(S)} m(y)}$.

We now describe a simple algorithm that outputs the support for a given set of parameters u and m and budget S. The following result describes the procedure. Intuitively, what we do is for any given set S, we find a subset Q such that if elements of Q are the only ones chosen with positive probability, then given the implied $\Lambda(Q)$, for all $y \in Q$, $u(y) \geq \Lambda(Q)$, and for all $z \in S \backslash Q$ $u(z) \leq \Lambda(Q)$.

Proposition 6. *For all u, $m > 0$, and $S \subseteq X$, there is a unique $\varnothing \neq Q \subseteq S$ for which $Q = \{x \in S|\ \Lambda(Q) < u(x)\}$. Furthermore, by setting $S_1 := S$ and defining recursively $S_{k+1} := \{x \in S_k|\ \Lambda(S_k) < u(x)\}$, there is a finite K^* for which $Q = S_{K^*}$, so that for all $k \geq K^*$, $Q = S_k$.*

Proof. Define $M_S \equiv \arg\max_{x \in S} u(x)$. Observe that for any $\varnothing \neq T \subseteq S$, and any $x \in M_S$, we have $\Lambda(T) < u(x)$. Initialize $S_1 := S$, and observe $M_S \subseteq S_1$. Given $M_S \subseteq S_k$, define $S_{k+1} := \{x \in S_k|\ \Lambda(S_k) < u(x)\}$. First, observe that $M_S \subseteq S_{k+1}$. Second, observe that if $x \in S_k \backslash S_{k+1}$, it follows that $u(x) \leq \Lambda(S_k)$, from which we obtain (using repeated applications of Lemma 1) that $\Lambda(S_k) \leq \Lambda(S_{k+1})$. Define $Q(S) := \bigcap_k S_k$; clearly $M_S \subseteq Q(S)$. We claim that $x \in Q(S)$ if and only if $u(x) > \Lambda(Q(S))$. By finiteness, there is K^* for which $Q(S) = S_{K^*} = S_k$ for all $k \geq K^*$. Suppose that $u(x) > \Lambda(Q(S))$; then $u(x) > \Lambda(S_{K^*})$ and so $x \in S_{K^*+1} = Q(S)$. Conversely, suppose that $x \in Q(S)$; then $x \in S_{K^*+1}$, so that $u(x) > \Lambda(S_{K^*}) = \Lambda(Q(S))$.

Next, we claim that Q is unique. Assume by means of contradiction that there exist two distinct subsets of S, say T_1 and T_2, such that

$$T_1 = \{x \in S|\ \Lambda(T_1) < u(x)\} \text{ and } T_2 = \{x \in S|\ \Lambda(T_2) < u(x)\}.$$

Without any loss of generality, assume $x \in T_1 \setminus T_2$. This implies that $\Lambda(T_1) < u(x) \leq \Lambda(T_2)$. Hence, T_2 is a proper subset of T_1. Since $u(z) > \Lambda(T_1)$ for all $z \in T_1 \setminus T_2$, by repeated applications of Lemma 1, $\Lambda(T_2) \leq \Lambda(T_1)$, a contradiction. □

Accordingly, let us define $Q(S)$ to be the unique $Q \subseteq S$ for which $Q = \{x \in S : \Lambda(Q) < u(x)\}$. As demonstrated, $Q(S)$ can be explicitly constructed from the primitives u, m and S via an iterative algorithm. Obviously, this algorithm must terminate in at most $|S| - 1$ steps.

B.1. **Weak WL Model.** Our characterization provided by Theorem 1 is based on the positivity assumption. Since our general formulation in equation (1) allows for zero probability choice, we show how to extend our characterization to this general case.[21] We first define a weaker version of our model allowing alternatives chosen with zero probability.

Definition 2. A stochastic choice ρ function is a *weak WL* stochastic choice (WWLSC) if there exist a utility function $u : X \to \mathbb{R}$ and a salience function $m : X \to \mathbb{R}_{++}$ such that for all $S \subset X$

$$\rho(x|S) = \begin{cases} m(x)u(x) - m(x)\Lambda(Q(S)) & x \in Q(S) \\ 0 & x \notin Q(S) \end{cases}$$

where $Q(A)$ is the unique subset of A satisfying $Q(A) = \{x \in A |\ \Lambda(Q(A)) < u(x)\}$ and $\Lambda(A) = \left(\sum_{y\in A} u(y)m(y) - 1\right) / \sum_{y\in A} m(y)$.

It should be clear from the results above that ρ is a WWLSC with representation (u, m) if and only if ρ is the solution to Equation (1) with parameters (u, m). Hence WWLSC captures the full range of solutions to Equation (1).

B.2. **Characterization.** We now provide two characterizations when zero probabilities are allowed. These characterizations differ in terms of how much they relax positivity. In the first one, we assume that positivity holds for menus with size 2 and 3. This will capture the example given at the end of Section 3. In Appendix B.3, we provide another characterization which entirely drops the positivity requirement. This characterization is based on linear programming duality. We provide both characterizations because, while the second is more general, the axioms used in the first characterization have a simpler behavioral interpretation, and as such, are useful for gaining intuition for the model.

For the first characterization, we modify our original axioms. The next axiom requires that the positivity holds for pairs and triples. In addition, the axiom also states that the strict regularity holds for these sets. Hence, the next axiom is a weakening of both Axiom 1 and Axiom 2.

Axiom 1*. For all binary and ternary set S, $\rho(x|S) > 0$ and $\rho(x|S) < \rho(x|S \setminus y)$.

The next axiom requires that the choice probabilities are not affected by removing alternative chosen with zero probability. If there is an alternative that is chosen with zero probability, removing it should not change the choice probabilities of the remaining items. This axiom is novel.

[21] In classical demand theory, demand with nonnegativity constraints for quadratic preferences is studied in Wegge (1968). In decision theory, Ahumada and Ulku (2018), Echenique and Saito (2018), Horan (2021), Matějka and McKay (2015) propose discrete choice models that accommodate zero probability choice.

Axiom Z. For all $S, S \setminus x \in \mathcal{D}$, if $\rho(x|S) = 0$ and $z \in S \setminus x$ then $\rho(z|S) = \rho(z|S \setminus x)$.

The next axiom is a version of Axiom 3. There are two differences. First, without positivity, we explicitly assume that some choice probabilities are positive. Second, the implication of the axiom is weaker now. The equality of Axiom 3 is replaced by an inequality. Other than these difference the intuition of the axiom stays the same: the ratio of relative levels of choice are important rather than the absolute levels.

Axiom 3*. For any list of three quadruples $((x_1, x_2, S_1, T_1), (x_2, x_3, S_2, T_2), (x_3, x_4, S_3, T_3))$ such that $x_4 = x_1$, $x_i, x_{i+1} \in S_i \cap T_i$ and $\rho(x_1|S_1), \rho(x_2|T_1) > 0$ and $\rho(x_i|A_i), \rho(x_{i+1}|A_i) > 0$ for all $i \in \{2, 3\}$ and $A_i \in \{S_i, T_i\}$,

$$d(x_1|S_1, T_1)d(x_2|S_2, T_2)d(x_3|S_3, T_3) \leq d(x_2|S_1, T_1)d(x_3|S_2, T_2)d(x_1|S_3, T_3)$$

Our first characterization in is as follows.

Theorem 3. *Suppose $\mathcal{D}$ contains all menus with size 2 and 3. Then a stochastic choice function ρ has a weak WL representation (u, c) on $\mathcal{D}$ such that $\Lambda(S) < min_{x \in S} u(x)$ for all $|S| \leq 3$ if and only if it satisfies Axiom 1*, 3* and Axiom Z.*

This theorem also enjoys the same uniqueness results observed in Theorem 1.

Proof. Necessity of the axioms is straightforward. We now illustrate sufficiency.

Since the domain of stochastic choice functions contains all menus with size 2 and 3 and positivity holds for these sets, we can define c the same way as we did in the proof of Theorem 1. That is,

$$c(x) := \frac{d(y^*|B_x, T_{xz})}{d(x|B_x, T_{xz})} = \frac{d(y^*|\{x, y^*\}, \{x, y^*, z\})}{d(x|\{x, y^*\}, \{x, y^*, z\})}$$

for some z different from x and y^*. We showed that c is well-defined and strictly positive for any x. Notice that $c(y^*) = 1$.

Now define the support of choice data for each set S,

$$Q(S) := \{x \in S |\ \rho(x|S) > 0\}$$

By Axiom Z, if $x \notin Q(S)$ then $Q(S \setminus x) = Q(S)$. Hence, $|Q(S)| \geq 3$ for every set S with $|S| \geq 3$ by Axiom 1*.

Claim 3. *If $\rho(x|S) > 0$ and $\rho(y|T) > 0$ then*

$$c(x)[\rho(x|S) - \rho(x|T)] \leq c(y)[\rho(y|S) - \rho(y|T)]$$

Proof. Consider distinct $S, T \in \mathcal{D}$ and $x, y \in S \cap T$. Axiom 1* implies that all choice probabilities in binary and ternary sets are different from zero. Since $\rho(x|S)$ and $\rho(y|T)$ are positive, Axiom 3* yields

$$d(x|S, T)d(y|B_y, T_{yz})d(y^*|B_x, T_{xy}) \leq d(y|S, T)d(y^*|B_y, T_{yz})d(x|B_x, T_{xy})$$

This implies $c(x)d(x|S,T) \le c(y)d(y|S,T)$. ☐

Claim 4. *If $x, y \in Q(S) \cap Q(T)$ then*

$$c(x)[\rho(x|S) - \rho(x|T)] = c(y)[\rho(y|S) - \rho(y|T)]$$

Proof. Applying Claim 3 twice yields $c(x)d(x|S,T) = c(y)d(y|S,T)$. ☐

We now recursively define λ for every element in $\mathcal{D}$ by the following formula

$$\Lambda(S) := c(x)d(x|T,S) + \Lambda(T)$$

Given S, to define $\Lambda(S)$, we first must find a set T such that $Q(S) \cap Q(T) \neq \emptyset$. Hence, in the first step, we define λ for all S with y^* as a member and $\rho(y^*|S) > 0$. Denote this set by $\mathcal{A}_0 := \{S \in \mathcal{D} \mid y^* \in Q(S)\}$. Then for all $S \in \mathcal{A}_0$, $\Lambda(S)$ is defined as $1 - \rho(y^*|S)$. In the second step, we define λ for the set of subsets, denoted by $\mathcal{A}_1 := \{S \in \mathcal{D} \mid y^* \notin Q(S)\}$. Take $S \in \mathcal{A}_1$ and for all $T \in \mathcal{A}_0$ such that $Q(S) \cap Q(T) \neq \emptyset$, and define $\Lambda(S) := c(x)d(x|T,S) + \Lambda(T)$ for some x in $Q(S) \cap Q(T)$. Existence of such T is trivial since $B_x \in \mathcal{A}_0$ whenever $x \in Q(S)$. Since $\mathcal{A}_0 \cup \mathcal{A}_1 = \mathcal{D}$, λ is defined for the entire choice problems.

Claim 5. *λ is well-defined and $\Lambda(S) + c(x)\rho(x|S) = \Lambda(T) + c(x)\rho(x|T)$ for all $x \in Q(S) \cap Q(T)$.*

Proof. We need to show that λ is well-defined (i.e., $\Lambda(S)$ is independent of the choice of x and T). By Claim 4, $\Lambda(S)$ is independent of the choice of x for a given T. Now we establish independence of T. Take two sets S, T such that $Q(S) \cap Q(T) \neq \emptyset$. If $S, T \in \mathcal{A}_0$, by definition we get $\Lambda(S) = c(y^*)d(y^*|T,S) + \Lambda(T)$ since $\Lambda(S) = 1 - \rho(y^*|S)$, $\Lambda(T) = 1 - \rho(y^*|T)$ and $c(y^*) = 1$. By Claim 4,

$$\Lambda(S) = c(x)d(x|T,S) + \Lambda(T) \text{ for all } x \in Q(S) \cap Q(T) \tag{8}$$

This establishes that the equation holds on $\mathcal{A}_0$. Now take $S \in \mathcal{A}_1$ and assume that $T, T' \in \mathcal{A}_0$ such that $x \in Q(S) \cap Q(T)$, $x' \in Q(S) \cap Q(T')$. Such alternatives exist since $|Q(S)| \ge \min\{3, |S|\}$ for every set S. Then

$$\begin{aligned}
\Lambda(S) &= c(x)d(x|T,S) + \Lambda(T) \\
&= c(x)d(x|T,S) + c(x)d(x|T_{xx'},T) + \lambda(T_{xx'}) \text{ by Eq (8)} \\
&= c(x)d(x|T_{xx'},S) + \lambda(T_{xx'}) \\
&= c(x')d(x'|T_{xx'},S) + \lambda(T_{xx'}) \text{ by Claim 4} \\
&= c(x')d(x'|T',S) + c(x')d(x'|T_{xx'},T') + \lambda(T_{xx'}) \\
&= c(x')d(x'|T',S) + \lambda(T') \text{ by Eq (8)}
\end{aligned}$$

This establishes that λ is well-defined on $\mathcal{A}_0 \cup \mathcal{A}_1$. Finally, we must show that $\Lambda(S) = c(x)d(x|T,S) + \Lambda(T)$ holds for $S, T \in \mathcal{A}_1$. Let x be an alternative in $Q(S) \cap Q(T)$.

Since $B_x \in \mathcal{A}_0$, we have $\Lambda(S) = c(x)d(x|B_x, S) + \lambda(B_x)$ and $\Lambda(T) = c(x)d(x|B_x, T) + \lambda(B_x)$. Subtracting these two equations yields $\Lambda(S) - \Lambda(T) = c(x)d(x|B_x, S) - c(x)d(x|B_x, T) = c(x)d(x|T, S)$. □

By Axiom Z, $Q(S) = Q(Q(S))$. In other words, the choice probabilities in S and $Q(S)$ are the same by Axiom Z. This means that $d(x|S, Q(S)) = 0$ for all x in $Q(S)$. This then gives us that $\Lambda(S) = \lambda(Q(S))$ for all S. Now define $u(x) := c(x)\rho(x|S) + \Lambda(S)$ for some S such that $x \in Q(S)$. u is well-defined by Claim 5. Note that $\Lambda(S) = \lambda(Q(S)) < u(x)$ for all $x \in Q(S)$ since $c(x)\rho(x|S) > 0$. Hence, for $x \in Q(S)$, the representation holds for those alternatives. Now assume $x \notin Q(S)$. That is, $\rho(x|S) = 0$. We need to show that $u(x) \leq \lambda(Q(S))$. Take $y \in Q(S)$. Then by definition, we have $\Lambda(S) = c(y)d(y|\{x, y\}, S) + \lambda(\{x, y\})$ and $u(x) = c(x)\rho(x|\{x, y\}) + \lambda(\{x, y\})$. Then we have

$$\Lambda(S) - u(x) = c(y)d(y|\{x, y\}, S) - c(x)\rho(x|\{x, y\})$$
$$\Lambda(S) - u(x) \geq c(x)d(x|\{x, y\}, S) - c(x)\rho(x|\{x, y\}) \text{ by Claim 3}$$
$$\Lambda(S) - u(x) \geq -c(x)\rho(x|S) = 0$$

Since $\Lambda(S) = \lambda(Q(S))$, the representation holds.

Finally, for all S such that $|S| \leq 3$, by Axiom 1^*, we have $Q(S) = S$ and $min_{x \in S} u(x) > \Lambda(S)$.

□

B.3. **A Characterization without Positivity.** Axiom 1^* still imposes a weak version of positivity. In this subsection we do not impose any positivity requirement. The intuition behind our result is that if we allow for arbitrary zero choice probabilities, the characterization holds if and only if there exists (i) a $c(x) > 0$, (ii) a $u(x)$, (iii) a $\mu(x|S) \geq 0$ only if $\rho(x|S) = 0$, and otherwise $\mu(x|S) = 0$, and (iv) $\gamma(S)$ so that for all (x, S):

$$c(x)\rho(x|S) - u(x) - \gamma(S) - \mu(x|S) = 0. \tag{9}$$

Here, μ is the Kuhn-Tucker multiplier on the non-negativity constraint for choice probabilities and γ is the constraint on probabilities summing to one (we have one constraint for each S).

Observe that, with knowledge of $\rho(x|S)$ for each x, S, equation (9) is a linear constraint. Here, γ is obviously $-\lambda$, but writing it in positive form makes the characterization slightly easier to state and helps us clearly distinguish the case allowing for zero probabilities compared to when probabilities are non-zero.

We will use the notation $\alpha(x|S)$ to refer to a multiplier on the constraint in equation (9). The unknowns are: (i) $c(x)$ for each x, $u(x)$ for each x, (ii) $\gamma(S)$ for each S, and (iii) $\mu(x|S)$ for each (x, S) with $x \in S$. Furthermore, we have the restrictions that (i) $\mu(x|S) = 0$ if $\rho(x|S) > 0$, and otherwise, $\mu(x|S) \geq 0$, and (ii) that $c(x) > 0$. These form a system of homogeneous linear inequalities.

The following is an application of Motzkin's Theorem of the Alternative.

Theorem 4. *The stochastic choice ρ has a WL representation with zero probabilities if and only if for any system of numbers $\alpha(x|S) \in \Re$ for which:*

- *For every $x \in X$, $\sum_{S:x\in S} \alpha(x|S) = 0$ (cycle condition across sets)*
- *For every $A \subseteq X$, $\sum_{x:x\in S} \alpha(x|S) = 0$ (cycle condition across alternatives)*
- *If $\rho(x|S) = 0$, then $\alpha(x|S) \geq 0$*
- *For every $x \in X$, $\sum_{S:x\in S} \alpha(x, S)\rho(x|S) \leq 0$*

it follows that for every $x \in X$, $\sum_{S:x\in S} \alpha(x|S)\rho(x|S) = 0$.

Proof. We apply Motzkin's Theorem of the Alternative (see Mangasarian (1994)). Let $\alpha(x|S)$ be the multiplier on the constraint specified by equation (9), let $\beta(x|S)$ be the multiplier on the constraint on $\mu(x|S) \geq 0$ when $\rho(x|S) = 0$, and let $\eta(x)$ be the multiplier on the constraint that $c(x) > 0$. Observe that there is no WL representation with zeroes if and only if there is $\alpha(x|S)$ for each x, S with $x \in S$, $\beta(x|S) \geq 0$ for each x, S where $x \in S$ and $\rho(x|S) = 0$, and finally $\eta(x) \geq 0$ for each x and where there exists x^* for which $\eta(x^*) > 0$, for which:

- For every $x \in X$, $\sum_{S:x\in S} \alpha(x|S) = 0$
- For every $S \subseteq X$, $\sum_{x:x\in S} \alpha(x|S) = 0$
- For every (x, S) with $x \in S$ and $\rho(x|S) = 0$, $-\alpha(x|S) + \beta(x|S) = 0$
- For every x, $\eta(x) + \sum_{S:x\in S} \alpha(x|S)\rho(x|S) = 0$.

By eliminating the multipliers η and β,[22] we get that the preceding is equivalent to the existence of $\alpha(x|S)$ for which

- For every $x \in X$, $\sum_{S:x\in S} \alpha(x|S) = 0$
- For every $S \subseteq X$, $\sum_{x:x\in S} \alpha(x|S) = 0$
- If $\rho(x|S) = 0$, then $\alpha(x|S) \geq 0$
- For every $x \in X$, $\sum_{S:x\in S} \alpha(x|S)\rho(x|S) \leq 0$
- There exists $x^* \in X$ for which $\sum_{S:x\in S} \alpha(x|S)\rho(x|S) < 0$.

Observe that the last of these properties is exactly what is ruled out by the conclusion of the statement in Theorem 4. Consequently, the satisfaction of this system must be equivalent to a violation of the statement listed in Theorem 4. ☐

[22]For example, we note that $\eta(x)+\sum_{S:x\in S} \alpha(x|S)\rho(x|S) = 0$ implies $\sum_{S:x\in S} \alpha(x|S)\rho(x|S) = -\eta(x) \leq 0$.

To see how this result relates to Theorem 1, we will show how it implies that $\frac{d(x_1|S_1,T_1)}{d(x_2|S_1,T_1)}\frac{d(x_2|S_2,T_2)}{d(x_1|S_2,T_2)} = 1$; the related condition on multiplicative cycles of triples or cycles of larger length follows similarly.

To this end, first define an auxiliary function $d(x|S,T) = \rho(x|S) - \rho(x|T)$, where $S \neq T$ and $x \in S \cap T$. Let us assume that $d(x|S,T) \neq 0$ for all relevant sets. Then take $\alpha(x_1|S_1) = 1 = -\alpha(x_1|T_1)$ and $\alpha(x_1|S_2) = -\frac{d(x_1|S_1,T_1)}{d(x_1|S_2,T_2)} = -\alpha(x_1|T_2)$, and for each set E, $\alpha(x_2|E) = -\alpha(x_1|E)$. All remaining coefficients are zero.

Observe that the constraints listed in Theorem 4 are satisfied, and in particular that $\alpha(x_1|S_1)\rho(x_1|S_1) + \alpha(x_1|S_2)\rho(x_1|S_2) + \alpha(x_1|T_1)\rho(x_1|T_1) + \alpha(x_1|T_2)\rho(x_1|T_2) = 0$.

Now, we claim that $\alpha(x_2|S_1)\rho(x_2|S_1) + \alpha(x_2|S_2)\rho(x_2|S_2) + \alpha(x_2|T_1)\rho(x_2|T_1) + \alpha(x_2|T_2)\rho(x_2|T_2) = 0$. To this end, observe that Theorem 4 implies that $\alpha(x_2|S_1)\rho(x_2|S_1) + \alpha(x_2|S_2)\rho(x_2|S_2) + \alpha(x_2|T_1)\rho(x_2|T_1) + \alpha(x_2|T_2)\rho(x_2|T_2) \geq 0$. If we had $\alpha(x_2|S_1)\rho(x_2|S_1) + \alpha(x_2|S_2)\rho(x_2|S_2) + \alpha(x_2|T_1)\rho(x_2|T_1) + \alpha(x_2|T_2)\rho(x_2|T_2) > 0$, then by choosing the system with coefficients $-\alpha$ instead of α, we would obtain a contradiction.

In particular now observe that the fact that $\alpha(x_2|S_1)\rho(x_2|S_1) + \alpha(x_2|S_2)\rho(x_2|S_2) + \alpha(x_2|T_1)\rho(x_2|T_1) + \alpha(x_2|T_2)\rho(x_2|T_2) = 0$ implies:

$$-\rho(x_2|S_1) + \frac{d(x_1|S_1,T_1)}{d(x_1|S_2,T_2)}\rho(x_2|S_2) + \rho(x_2|T_1) - \frac{d(x_1|S_1,T_1)}{d(x_1|S_2,T_2)}\rho(x_2|T_2) = 0.$$

This implies $d(x_2|S_2,T_2)\frac{d(x_1|S_1,T_1)}{d(x_1|S_2,T_2)} = d(x_2|S_1,T_1)$. Conclude $\frac{d(x_1|S_1,T_1)}{d(x_2|S_1,T_1)}\frac{d(x_2|S_2,T_2)}{d(x_1|S_2,T_2)} = 1$. □

Appendix C. Identification with Attributes

Our previous results on identification relied on variation in choice sets. This raises questions of how to identify the parameters in our model where the choice set is fixed, but where product attributes are observable (a standard situation in many empirical applications beginning with Rosen, 1974).[23] Here, we turn to discussing identification of our model in precisely this setting — with a fixed choice set, but observable product characteristics.[24] We demonstrate that even in this setting we can leverage the micro-foundations we provided earlier to generate simple conditions that allow for transparent identification using a set of linear equations.

Of course, as the previous results should make clear, given a single set S and choice probabilities $\rho(i|S)$ for all $x \in S$ we can always construct a WL model that rationalizes the data. In other words, with no additional information our model is not falsified by observing any single choice set. Similarly, we cannot uniquely identify u and m with such data.

However, if we assume (as is typical) that both u and m are functions of observable attributes, then identification proceeds in a clear manner. In particular, suppose that there is a set of N observable attributes, with a_i denoting the vector of attributes for product i. a_i includes not only things that affect product quality, but also things like price, advertising, etc.

Typically utility is assumed to be a linear function of attributes. We maintain the same assumption here and extend it to m. Thus, we assume that there exists a vector β such that $u_i = \beta a_i$ for each i. Similarly there exists a vector α such that $1/m_i = \alpha a_i$ for each i. Thus, we assume that attributes affect utility in the same way for all products, and similarly for salience. The only difference between products is the value that each attribute takes on. Given this assumption, under relatively mild conditions our model is identified using standard linear equations.

Proposition 7. *Suppose that $u_i = \beta a_i$ and $1/m_i = \alpha a_i$ where a_i is a $N \times 1$ vector. Suppose that we have at least $2N$ linearly independent observations of $(\rho(i)a_i - \rho(j)a_j, a_i - a_j)$ for $i, j \in S$. Then β and α are identified from choices in S up to positive scalar multiplication.*

Proof. We know that within a choice set it must be the case that: $\rho(i)/m_i - u_i = -\Lambda(S) = \rho(j)m_j - u_j$ or $\rho(i)/m_i - u_i = \rho(j)m_j - u_j$. This means $\rho(i)\alpha a_i - \beta a_i = \rho(j)\alpha a_j - \beta a_j$ or $\alpha[\rho(i)a_i - \rho(j)a_j] = \beta[a_i - a_j]$. Denote $PA(i,j) = \rho(i)a_i - \rho(j)a_j$ and $A(i,j)$ as $a_i - a_j$. Suppose for all pairs $PA(i,j)$ and $A(i,j)$ are linearly independent. With at least $2N$ pairs the model is then identified. □

[23]An additional consideration in typical applications is that there may be endogenous, unobserved, characteristics to products. Given that there is extensive discussion of this issue in the literature, we abstract away from it, and suppose that all attributes are observable to the researcher.

[24]Allen and Rehbeck (2019) proposes an axiomatic theory in such a framework.

For identification of the preference parameter vectors β and α, corresponding to the weights $1/m$ and u put on each attribute, one simply finds the solutions to the system of equations: $\beta[\rho(i)a_i - \rho(j)a_j] = \alpha[a_i - a_j]$ (where there is one equation for each pair of outcomes). The equations are linear in the attributes, making this a computationally simple problem.[25]

Notice that we obtain one equation for each pair of outcomes. Thus, fixing a number of attributes m, as long as there are (i) enough products and (ii) enough linearly independent combinations of attributes across products, the model is identified. In particular, if the choice set has $|S|$ products we will have $\frac{|S|(|S|-1)}{2}$ pairwise comparisons. If all pairs of comparison are linearly independent then all we need for identification is that $\frac{|S|(|S|-1)}{2} \geq 2m$.

Recall that u, m is unique up to transformations of the type $\kappa u + \gamma, \frac{1}{\kappa}m$ where $\kappa > 0$ and for any γ. Observe that because we assume that $u_i = \beta a_i$, we impose that $\gamma = 0$. Thus, the representation is unique up to transformations of the form $\kappa u, \frac{1}{\kappa}m$. Notice that this occurs if and only if α, β are unique up to transformation $\kappa\alpha, \kappa\beta$.

In order to highlight our approach, we will consider a stylized example. Suppose we have products with two attributes, and there are four products (the minimum needed for identification). Moreover we observe choices from the grand choice set (again, necessary for identification), and $\rho(1|\{1,2,3,4\}) = 0.273$, $\rho(2|\{1,2,3,4\}) = 0.197$, $\rho(3|\{1,2,3,4\}) = 0.121$, and $\rho(4|\{1,2,3,4\}) = 0.409$. Moreover, suppose $a_1 = (4,2)$; $a_2 = (4,1)$; $a_3 = (2,8)$; $a_4 = (4,8)$. Then we have 6 pairwise comparisons. One can show that each of the six entries in the $(\rho(i)a_i - \rho(j)a_j, a_i - a_j)$ are linearly independent of all the others. Thus, our conditions are met. Using any subset of 4 of the pairwise comparisons delivers the result that, using α_1 as the "numeraire" coefficient, $\beta_1 = 3\alpha_1$, $\beta_2 = \alpha_1$ and $\alpha_2 = 2\alpha_1$. In this world, we can do out of sample predictions, which include introducing a new product or improving existing product. For example, if product 1 is improved in terms of attribute 1 (i.e., $a_{11} \geq 6$), product 2 will be driven out of the market.

Appendix D. Estimated models

D.1. **Logit.** Each option x has a value $v(x)$ and choice probabilities are given by

$$\rho(x|S) = \frac{e^{v(x)}}{\sum_{y \in S} e^{v(y)}}$$

[25]Our approach is distinct from that typically used for discrete choice models, as outlined in S. Berry (1994). In fact, Armstrong and Vickers (2015), building on the work of Jaffe and Weyl (2010) and Jaffe and Kominers (2012), show that although linear demands (which our model is an example of) can be consistent with a model of discrete choice, they fail some standard assumptions about the distribution of the utility shock, assumptions which are used to ensure identification in standard discrete choice approaches (i.e. continuity and full support assumptions on the error term).

we estimate v values normalizing the value of one option $v(x) = 0$.

D.2. **Nested Logit.** Alternatives are partitioned into K nests $B_1, \ldots, B_K$. Each option x has a value $v(x)$ and each nest B_k has an associated parameter $0 \leq \lambda_k \leq 1$. Choice probabilities for an option $x \in B_k$ in the menu S are given by

$$\rho(x|S) = e^{v(x)/\lambda_k} \frac{\left(\sum_{y \in S \cap B_k} e^{v(y)/\lambda_k}\right)^{\lambda_k - 1}}{\sum_{\ell=1}^{K} \left(\sum_{y \in S \cap B_\ell} e^{v(y)/\lambda_\ell}\right)^{\lambda_\ell}}$$

and setting $\lambda_\ell = 1$ for all nests yields the classic logit model as a special case. Note that we can always normalize one of the values to $v(x) = 0$.

The partition of options into nests is a degree of freedom chosen by the analyst. With four alternatives $X = \{1, 2, 3, 4\}$ the analyst can employ seven different nest specifications, namely, $\{\{1, 2, 3\}, \{4\}\}$, $\{\{1, 2, 4\}, \{3\}\}$, $\{\{1, 3, 4\}, \{2\}\}$, $\{\{2, 3, 4\}, \{1\}\}$, $\{\{1, 2\}, \{3, 4\}\}$, $\{\{1, 3\}, \{2, 4\}\}$, and $\{\{1, 4\}, \{2, 3\}\}$. Additional partitions are subsumed by one of the seven nest specifications above. For example, the nest specification $\{\{1\}, \{2\}, \{3, 4\}\}$ is a special case of $\{\{1, 2\}, \{3, 4\}\}$, when the parameter $\lambda_{\{1,2\}} = 1$. We estimate v's and λ's separately for each of the seven nest specifications, and in each prediction exercise we take the best fitting nest specification to make predictions.

D.3. **Covariance Probit.** The random utilities of each option $U_1, U_2, \ldots, U_4$ are assumed to have a joint Gaussian distribution with mean vector μ and covariance matrix Σ. Choice probabilities maximize random utility

$$\rho(x|S) = \mathbb{P}\{U_i \geq U_j \text{ for all } j \in S\}$$

which is an integral that has no closed-form but can be numerically calculated. For identification, we do a normalization

$$Z_i = \frac{U_i - U_4}{\sqrt{\mathrm{Var}(U_1 - U_4)}}$$

and note the Gaussian variables $Z_1, \ldots, Z_4$ represent the same choice behavior, with eight parameters to estimate:

$$\tilde{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ 0 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} 1 & \sigma_{12} & \sigma_{13} & 0 \\ \sigma_{12} & \sigma_2 & \sigma_{23} & 0 \\ \sigma_{13} & \sigma_{23} & \sigma_3 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Appendix E. Additional simulation prediction metrics

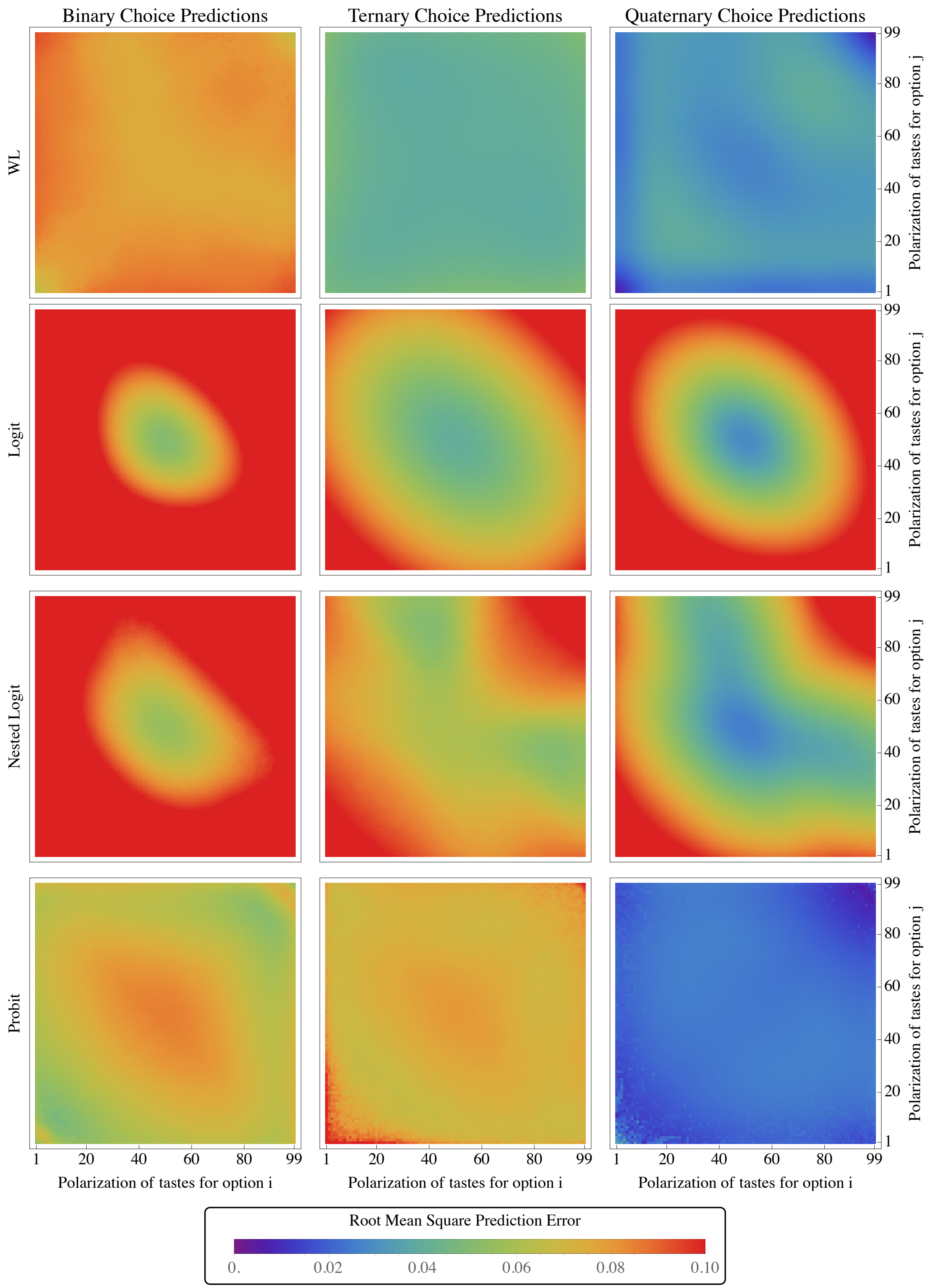

Figure 7. Root mean square prediction errors for the WL model and for Logit, Nested Logit and Probit. The axes display the fixed level of rank polarization for two different options.

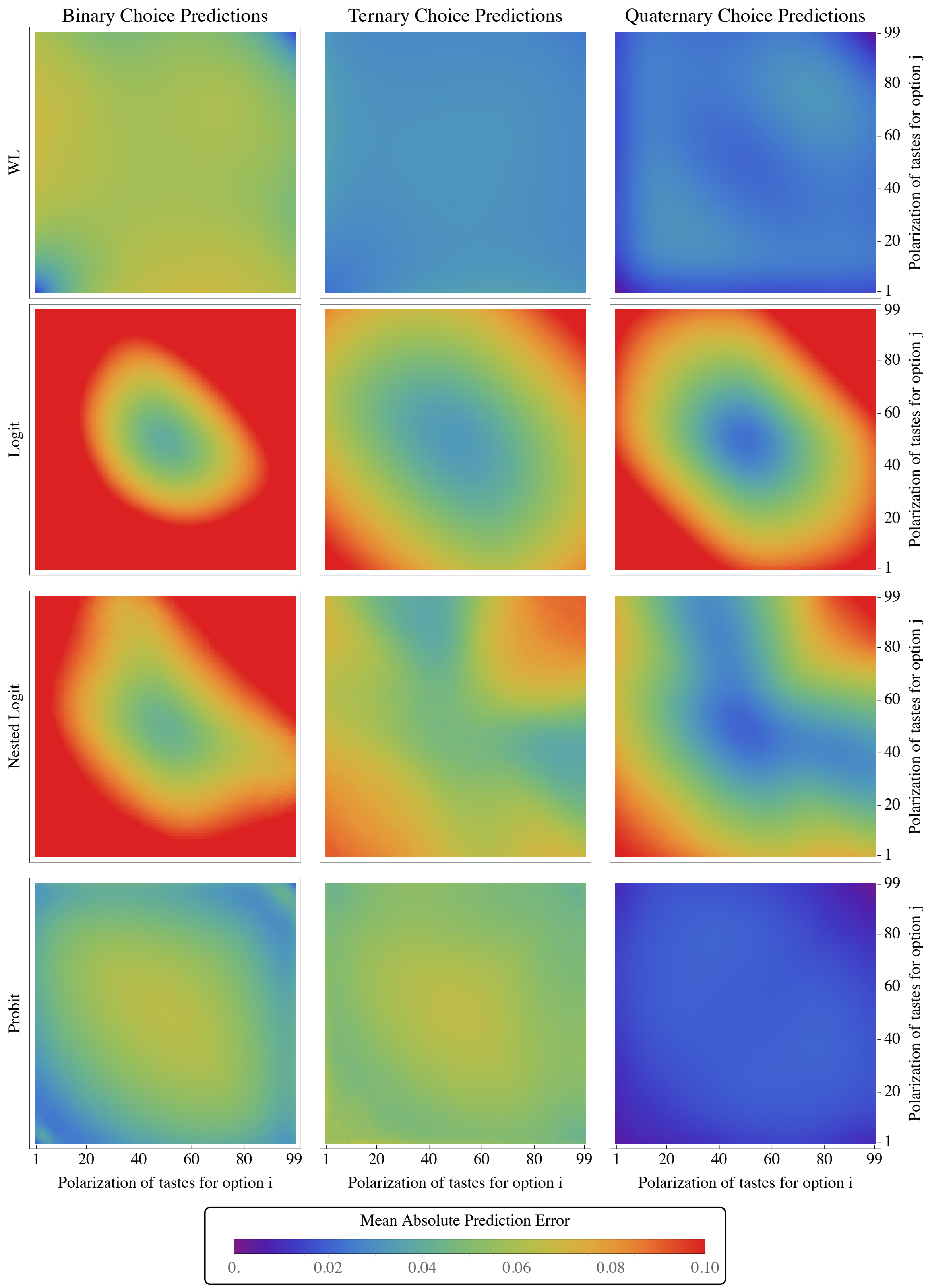

FIGURE 8. Mean absolute prediction error comparison for the WL model and for Logit, Nested Logit and Probit. The axes display the fixed level of rank polarization for two different options.