

## Online Appendix

### Weighted Linear Discrete Choice

Christopher P. Chambers, Yusufcan Masatlioglu, Paulo Natenzon, and Collin Raymond.

#### 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 (2). We start by proving a useful lemma.

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

*Proof.*

$$\begin{aligned}
 \frac{\sum_{y \in S} u(y)m(y) - 1}{m(S)} &\geq u(x) \\
 \sum_{y \in S} u(y)m(y) - 1 &\geq u(x)m(S) \\
 m(x) \left( 1 - \sum_{y \in S} u(y)m(y) \right) &\leq -u(x)m(x)m(S) \\
 \left( \sum_{y \in S} u(y)m(y) - 1 \right) (m(S) - m(x)) &\leq m(S) \left( \sum_{y \in S} u(y)m(y) - u(x)m(x) - 1 \right) \\
 \left( \sum_{y \in S} u(y)m(y) - 1 \right) m(S \setminus x) &\leq m(S) \left( \sum_{y \in S \setminus x} u(y)m(y) - 1 \right) \\
 \Lambda(S) &\leq \Lambda(S \setminus x)
 \end{aligned}$$

□

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 Proposition 2.

A2

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

### A.3. Proof of Theorem 2.

*Proof.* Observe that for any  $S$ ,

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

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

To show the other way, take two different representations of  $\rho$ :  $(u, m)$  and  $(v, n)$ . For any arbitrary representation  $\rho$ ,  $(u, m)$ , 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  $(u(x) > \Lambda(X) \text{ for all } x)$ . Hence  $u'(x), v'(x) \neq 0$  for all  $x$ .

Moreover, since  $(u', m')$  and  $(v', n')$  represents  $\rho$ , 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.  $\square$

#### A.4. 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:  $c(x)d(x|S, T) = c(y)d(y|S, T)$ .  $\square$

We now recursively define  $\lambda$  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  $\lambda$  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  $\lambda$  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.**  $\lambda$  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  $\lambda$  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,

$$(5) \quad \Lambda(S) = c(x)d(x|T, S) + \Lambda(T) \text{ for all } x \in S \cap T$$

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

$$\begin{aligned} \Lambda(S) &= c(x)d(x|T, S) + \Lambda(T) \\ &= c(x)d(x|T, S) + c(x)d(x|\{x, x', y^*\}, T) + \lambda(\{x, x', y^*\}) \text{ by Eq (5)} \\ &= c(x)d(x|\{x, x', y^*\}, S) + \lambda(\{x, x', y^*\}) \\ &= c(x')d(x'|\{x, x', y^*\}, S) + \lambda(\{x, x', y^*\}) \text{ by Claim 1} \\ &= c(x')d(x'|T', S) + c(x')d(x'|\{x, x', y^*\}, T') + \lambda(\{x, x', y^*\}) \\ &= c(x')d(x'|T', S) + \lambda(T') \text{ by Eq (5)} \end{aligned}$$

This establishes that  $\lambda$  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  $\Lambda(S) - \Lambda(T) = c(x)d(x|B_x, S) - c(x)d(x|B_x, T) = c(x)d(x|T, S)$ .  $\square$

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  $\rho$ .  $\square$

### A.5. Proof of Proposition 3.

*Proof.* Let  $(u, m)$  be a WL representation of  $\rho$ . 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  $\rho$ , 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:  $\sum_{S': S \subseteq S'} (-1)^{|S' \setminus S|} \rho(x|S)$ . Non-negativity 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. □

## 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 (2).

Observe first that  $\Lambda$  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,  $\text{supp}_\rho(S) := \{x \in S \mid \rho(x|S) > 0\}$ . 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  $z \in S \setminus Q$   $u(z) \leq \Lambda(Q)$ .

**Proposition 5.** *For all  $u, m > 0$ , and  $S \subseteq X$ , there is a unique  $\emptyset \neq Q \subseteq S$  for which  $Q = \{x \in S \mid \Lambda(Q) < u(x)\}$ . Furthermore, by setting  $S_1 := S$  and defining recursively  $S_{k+1} := \{x \in S_k \mid \Lambda(S_k) < u(x)\}$ , 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  $\emptyset \neq T \subseteq S$ , 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  $S_{k+1} := \{x \in S_k \mid \Lambda(S_k) < u(x)\}$ . First, observe that  $M_S \subseteq S_{k+1}$ . Second, observe that if  $x \in S_k \setminus S_{k+1}$ , 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

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

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.  $\square$

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 (2) allows for zero probability choice, we show how to extend our characterization to this general case.<sup>20</sup> We first define a weaker version of our model allowing alternatives chosen with zero probability.

<sup>20</sup>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), and Matějka and McKay (2015) propose discrete choice models that accommodate zero probability choice.

**Definition 2.** A stochastic choice  $\rho$  function is a *weak WL* stochastic choice (WWLSC) if there exist a utility function  $u : X \rightarrow \mathbb{R}$  and a salience function  $m : X \rightarrow \mathbb{R}_{++}$  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  $Q(A) = \{x \in A \mid \Lambda(Q(A)) < u(x)\}$  and  $\Lambda(A) = (\sum_{y \in A} u(y)m(y) - 1) / \sum_{y \in A} m(y)$ .

It should be clear from the results above that  $\rho$  is a WWLSC with representation  $(u, m)$  if and only if  $\rho$  is the solution to Equation (2) with parameters  $(u, m)$ . Hence WWLSC captures the full range of solutions to Equation (2).

**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 online 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.

**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

$$(x_1, x_2, S_1, T_1), (x_2, x_3, S_2, T_2), \text{ and } (x_3, x_4, S_3, T_3)$$

such that  $x_4 = x_1$ ,  $x_i, x_{i+1} \in S_i \cap T_i$ ,  $\rho(x_1|S_1), \rho(x_2|T_1) > 0$  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  $\rho$  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  $S$ ,

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

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  $c(x)d(x|S, T) \leq c(y)d(y|S, T)$ . □

**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  $\lambda$  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  $\lambda$  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  $\lambda$  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  $\mathcal{A}_0 \cup \mathcal{A}_1 = \mathcal{D}$ ,  $\lambda$  is defined for the entire choice problems.

**Claim 5.**  $\lambda$  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  $\lambda$  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,

$$(6) \quad \Lambda(S) = c(x)d(x|T, S) + \Lambda(T) \text{ for all } x \in Q(S) \cap Q(T)$$

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  $x \in Q(S) \cap Q(T), x' \in Q(S) \cap Q(T')$ . Such alternatives exist since  $|Q(S)| \geq \min\{3, |S|\}$  for every set  $S$ . Then

$$\begin{aligned} \Lambda(S) &= c(x)d(x|T, S) + \Lambda(T) \\ &= c(x)d(x|T, S) + c(x)d(x|T_{xx'}, T) + \lambda(T_{xx'}) \text{ by Eq (6)} \\ &= c(x)d(x|T_{xx'}, S) + \lambda(T_{xx'}) \\ &= c(x')d(x'|T_{xx'}, S) + \lambda(T_{xx'}) \text{ by Claim 4} \\ &= c(x')d(x'|T', S) + c(x')d(x'|T_{xx'}, T') + \lambda(T_{xx'}) \\ &= c(x')d(x'|T', S) + \lambda(T') \text{ by Eq (6)} \end{aligned}$$

This establishes that  $\lambda$  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  $\Lambda(S) - \Lambda(T) = c(x)d(x|B_x, S) - c(x)d(x|B_x, T) = c(x)d(x|T, S)$ .  $\square$

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

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

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  $\min_{x \in S} u(x) > \Lambda(S)$ .  $\square$

**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  $(x, S)$ :

$$(7) \quad c(x)\rho(x|S) - u(x) - \gamma(S) - \mu(x|S) = 0.$$

Here,  $\mu$  is the Kuhn-Tucker multiplier on the non-negativity constraint for choice probabilities and  $\gamma$  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 (7) is a linear constraint. Here,  $\gamma$  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 (7). 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  $\rho$  has a WL representation with zero probabilities if and only if for any system of numbers  $\alpha(x|S) \in \mathbb{R}$  for which:*

- For every  $x \in X$ ,  $\sum_{S: x \in S} \alpha(x|S) = 0$  (cycle condition across sets)

- For every  $A \subseteq X$ ,  $\sum_{x \in A} \alpha(x|S) = 0$  (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 (7), let  $\beta(x|S)$  be the multiplier on the constraint  $\mu(x|S) \geq 0$  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  $S \subseteq X$ ,  $\sum_{x \in S} \alpha(x|S) = 0$
- 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  $\eta$  and  $\beta$ ,<sup>21</sup> 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  $S \subseteq X$ ,  $\sum_{x \in S} \alpha(x|S) = 0$
- 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.  $\square$

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  $\alpha(x_1|S_1)\rho(x_1|S_1) + \alpha(x_1|S_2)\rho(x_1|S_2) + \alpha(x_1|T_1)\rho(x_1|T_1) + \alpha(x_1|T_2)\rho(x_1|T_2) = 0$ .

<sup>21</sup>For example, we note that  $\eta(x) + \sum_{S: x \in S} \alpha(x|S) \rho(x|S) = 0$  implies  $\sum_{S: x \in S} \alpha(x|S) \rho(x|S) = -\eta(x) \leq 0$ .

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  $\alpha$ , 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  $\frac{d(x_1|S_1, T_1)}{d(x_2|S_1, T_1)} \frac{d(x_2|S_2, T_2)}{d(x_1|S_2, T_2)} = 1$ .  $\square$

## 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).<sup>22</sup> Here, we turn to discussing identification of our model in precisely this setting — with a fixed choice set, but observable product characteristics.<sup>23</sup> 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.

<sup>22</sup>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.

<sup>23</sup>Allen and Rehbeck (2019) proposes an axiomatic theory in such a framework.

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  $\beta$  such that  $u_i = \beta a_i$  for each  $i$ . Similarly, there exists a vector  $\alpha$  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 6.** *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  $\beta$  and  $\alpha$  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.  $\square$

For identification of the preference parameter vectors  $\beta$  and  $\alpha$ , 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.<sup>24</sup>

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  $\gamma$ . 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  $\alpha, \beta$  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

<sup>24</sup>Our approach is distinct from that typically used for discrete choice models, as outlined in 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).

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  $\alpha_1$  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. 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:

$$(8) \quad \rho(x, y) = \frac{u(x) - u(y) + \frac{1}{m(y)}}{\frac{1}{m(x)} + \frac{1}{m(y)}} = \frac{1}{2} + \frac{1}{2} \frac{U(x) - U(y)}{\frac{1}{m(x)} + \frac{1}{m(y)}}$$

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 (8) 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 (8), 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 (8) 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 (8) 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.

## APPENDIX E. DATA SIMULATION AND ESTIMATION

*Setting.* There are four choice alternatives  $X = \{1, 2, 3, 4\}$ . We choose  $n = 4$  choice alternatives 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).

*Data generating process.* Using simulated demand data allows us to assess the strengths and weaknesses of each model at counterfactual prediction over a wide domain of different demand systems. The domain of data generating processes we consider is the set of all possible demand systems that may arise from a heterogeneous population of standard rational consumers. That is, our domain is the entire RUM family. This is a natural domain to consider since it is often the de-facto assumed paradigm in discrete choice estimation. The RUM imposes no restrictions on the preference heterogeneity present in a population, other than that it is fixed across choice sets. Every parametric

model we use for estimation and prediction, including ours, is a special (restricted) case of this paradigm.

With  $n = 4$  options, a population of rational consumers is divided into  $n! = 24$  consumer types, that is, strict rankings over  $X$ . In Table 1, each type is represented as a column vector, where the most preferred (best) option is located at the top, and the least preferred (worst) option is located at the bottom.

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

TABLE 1. A population of consumers in the RUM is described by a probability  $\pi$  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  $\pi$  to types in the first row of the table; while  $P(3, \pi)$  is the probability given by  $\pi$  to types in the first column.

In Section 4 we defined the rank polarization for an option  $x$  in the RUM model  $\pi$  as the sum of the probabilities of types that rank  $x$  as their most preferred (best) option and as their least preferred (worst) option. We can easily compute the degree of polarization for any given instance of a RUM model. The rank polarization for option two,  $P(2, \pi)$ , is the probability given by  $\pi$  to types in the top row of Table 1, while  $P(3, \pi)$ , the rank polarization for option three, is the probability given by  $\pi$  to rankings in the left column.

By adjusting the weights given to the columns and rows of Table 1, we can generate a RUM  $\pi$  with any desired level of rank polarization for two options. It is not possible, however, to freely adjust the rank polarization for a third option. For instance, when  $P(2, \pi)$  and  $P(3, \pi)$  are close to one,  $\pi$  must give probability close to one to the first cell in Table 1. Since every type in that cell ranks options 1 and 4 in second or third place, 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.

We cover the domain of RUM models by simulating demand systems over the entire range of rank polarization levels for two options in a discrete grid

$$(P(2, \pi), P(3, \pi)) \in \left\{ \left( \frac{j}{100}, \frac{k}{100} \right) : 1 \leq j, k \leq 99 \right\}.$$

For each fixed level of  $P(2, \pi)$  and  $P(3, \pi)$ , we simulate 1,000 different demand systems drawn independently, as follows. First, for each individual cell in Table 1, we draw a RUM population using the uniform distribution over the simplex of probability distributions with support on the types from that cell. Then, we mix the RUMs drawn from different cells with appropriate weights to create a RUM  $\pi$  with the desired rank polarization levels. To mix the probabilities from each cell, we treat the events in the rows and columns of Table 1 as independent: for example, the weight given to the draw from the first column and first row is  $P(2, \pi) \times P(3, \pi)$ , and so on. Following this procedure, we generate a total of  $99^2 \times 1,000 = 9,801,000$  different RUM models.

*Leave-one-out prediction.* Our setup with  $n = 4$  choice options has six binary choice problems  $\{1, 2\}$ ,  $\{1, 3\}$ ,  $\{1, 4\}$ ,  $\{2, 3\}$ ,  $\{2, 4\}$ , and  $\{3, 4\}$ ; four ternary choice problems  $\{1, 2, 3\}$ ,  $\{1, 2, 4\}$ ,  $\{1, 3, 4\}$ , and  $\{2, 3, 4\}$ ; and one quaternary choice problem  $\{1, 2, 3, 4\}$ . For each simulated demand system, 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.

*Comparison models.* We compare the predictions of the WL model to three discrete choice workhorse models in the literature: the classic multinomial Logit, the Nested Logit and the Covariance Probit. These well-known models are described in any standard discrete choice estimation textbook, e.g., Train (2009). Figure 1 shows that when  $n = 4$ , Logit has three parameters to be estimated, WL has six, Probit has eight, and Nested Logit has thirteen parameters across seven different nest specifications. We review how each model is parameterized in detail in online Appendix F. All models including the WL are special cases of RUM and therefore misspecified. Logit is the most restrictive model and has well-known limitations. Probit, Nested Logit and WL address some of the limitations of classic logit in very different ways. There is no reason, ex-ante, for one model to perform better than the others in our prediction exercises.

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

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

where the summation is 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  $\rho$ , writing the log-likelihood for the finite sample, dividing by  $N$ , and taking the limit  $N \rightarrow \infty$ . This is also equivalent to minimizing the Kullback–Leibler divergence between the true data-generating choice rule and the parameterized model. This estimation procedure allows us to cleanly compare the prediction performance of different models without worrying about the finite-sample properties of any particular estimator.

*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. 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. For Mathematica replication code see Chambers et al., [2025](#).

## APPENDIX F. ESTIMATED MODELS

F.1. **WL.** To implement the maximum likelihood for the WL model on the computer, we first transform the constrained optimization problem into an unconstrained optimization problem using a straightforward change of variables:

$$m_i = e^{w_i} \text{ and } u_i m_i = \frac{e^{v_i}}{\sum_{j \in X} e^{v_j}}.$$

Then, we normalize one  $v_i = 0$  and one  $w_j = 0$  and estimate the remaining parameters using off-the-shelf unconstrained optimization routines. Under standard regularity assumptions (i.e. that  $v, w$  lie in a large compact subset of  $R^n$ ) the estimation is consistent by Theorem 2.1 of Newey and McFadden (1994).

F.2. **Logit.** Each option  $x$  has a value  $v(x)$  and choice probabilities are given by

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

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

F.3. **Nested Logit.** In the Nested Logit model the analyst first partitions the set of alternatives into  $K$  nests  $B_1, \dots, B_K$ . Each option  $x$  has a utility parameter  $v(x)$  and each nest  $B_k$  has an associated substitution parameter  $0 \leq \lambda_k \leq 1$ . Choice probabilities

for an option  $x \in B_k$  in the menu  $S$  are given by

$$\rho_{\text{NL}}(x, S, \{B_1, \dots, B_K\}, v, \lambda) = e^{v(x)/\lambda_k} \frac{\left(\sum_{y \in S \cap B_k} e^{v(y)/\lambda_k}\right)^{\lambda_k - 1}}{\sum_{\ell=1}^K \left(\sum_{y \in S \cap B_\ell} e^{v(y)/\lambda_\ell}\right)^{\lambda_\ell}}$$

The partition  $\{B_1, \dots, B_K\}$  of options into nests is chosen by the analyst. With four alternatives in the grand set  $X = \{1, 2, 3, 4\}$  the analyst chooses one of 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 nest specifications are automatically 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 substitution parameter for nest  $\{1, 2\}$  is equal to 1.

We estimate utility parameters and substitution parameters for each of the seven nest specifications, and in each prediction exercise we take the best fitting nest specification to make out-of-sample predictions. To simplify notation below, we write  $v_i$  for the utility value of option  $i$  and  $\lambda_{ijk}$  for the substitution parameter associated with nest  $\{i, j, k\}$ . We can write our estimated Nested Logit model, including the choice of nest specification, as

$$\begin{aligned} \rho(x|S) = & D_1 \times \rho_{\text{NL}}(x, S, \{\{1, 2, 3\}, \{4\}\}, v_1, v_2, v_3, \lambda_{123}) \\ & + D_2 \times \rho_{\text{NL}}(x, S, \{\{1, 2, 4\}, \{3\}\}, v_1, v_2, v_3, \lambda_{124}) \\ & + D_3 \times \rho_{\text{NL}}(x, S, \{\{1, 3, 4\}, \{2\}\}, v_1, v_2, v_3, \lambda_{134}) \\ & + D_4 \times \rho_{\text{NL}}(x, S, \{\{2, 3, 4\}, \{1\}\}, v_1, v_2, v_3, \lambda_{234}) \\ & + D_5 \times \rho_{\text{NL}}(x, S, \{\{1, 2\}, \{3, 4\}\}, v_1, v_2, v_3, \lambda_{12}, \lambda_{34}) \\ & + D_6 \times \rho_{\text{NL}}(x, S, \{\{1, 3\}, \{2, 4\}\}, v_1, v_2, v_3, \lambda_{13}, \lambda_{24}) \\ & + D_7 \times \rho_{\text{NL}}(x, S, \{\{1, 4\}, \{2, 3\}\}, v_1, v_2, v_3, \lambda_{14}, \lambda_{23}) \end{aligned}$$

where  $D_1, \dots, D_7$  are indicator (dummy) variables for nest specification, with the additional restriction that  $D_1 + \dots + D_7 = 1$ . Since one utility value can always be normalized, we set  $v_4 = 0$ . Hence, in addition to the dummy variables, we estimate the remaining 3 utility parameters and 10 substitution parameters:

$$v_1, v_2, v_3, \lambda_{123}, \lambda_{124}, \lambda_{134}, \lambda_{234}, \lambda_{12}, \lambda_{34}, \lambda_{13}, \lambda_{24}, \lambda_{14}, \lambda_{23}$$

More generally, with  $n$  alternatives, the Nested Logit model has  $n - 1$  utility parameters plus  $2^n - n - 1$  substitution parameters for a total of  $2^n - 3$  parameters. This is plotted in Figure 2 in the main text. We exclude the  $D_i$  indicator (dummy) variables of nest specification from the count of parameters.

**F.4. Covariance Probit.** The random utilities of each option  $U_1, U_2, \dots, U_4$  are assumed to have a joint Gaussian distribution with mean vector  $\mu$  and covariance matrix

$\Sigma$ . 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

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

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

$$\mathbb{E}[Z] = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ 0 \end{bmatrix} \text{ and } \text{Var}[Z] = \begin{bmatrix} 1 & \sigma_{12} & \sigma_{13} & 0 \\ \sigma_{12} & \sigma_2 & \sigma_{23} & 0 \\ \sigma_{13} & \sigma_{23} & \sigma_3 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

## APPENDIX G. ADDITIONAL PREDICTION ERROR METRICS

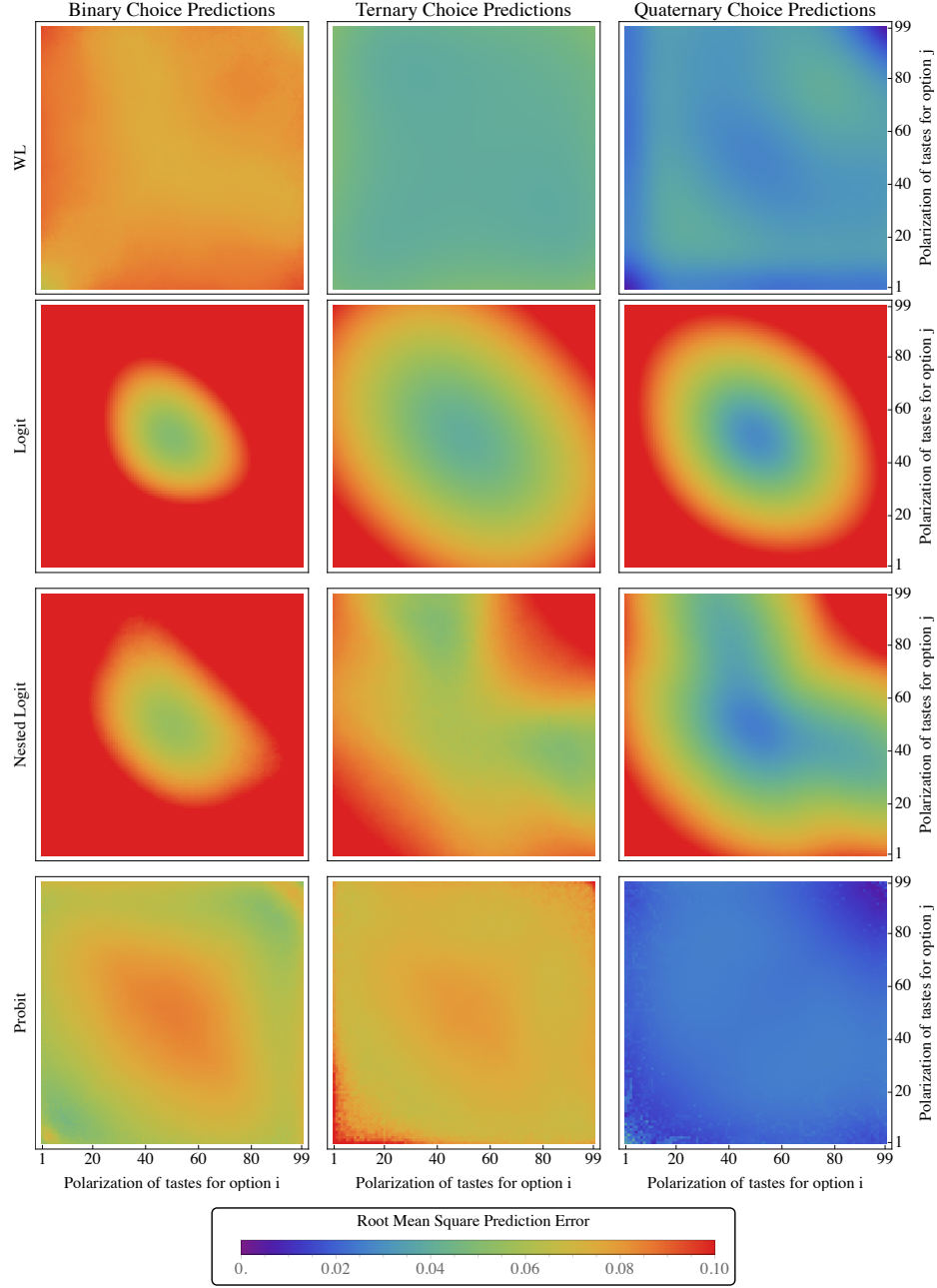


FIGURE 4. Root mean square prediction errors for the WL model and for Logit, Nested Logit and Probit. The axes display the fixed level of rank polarization for two different options.

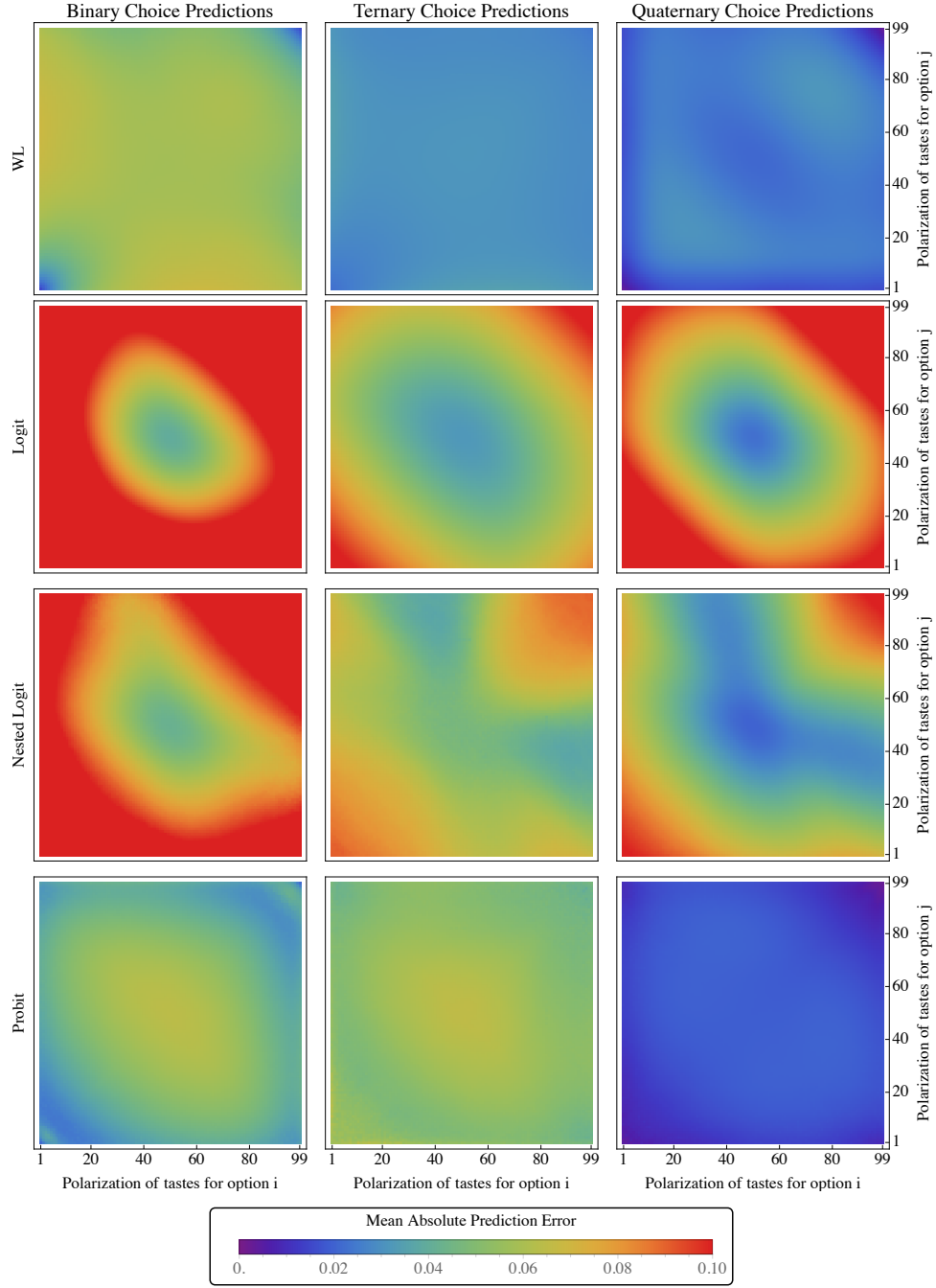


FIGURE 5. Mean absolute prediction error comparison for the WL model and for Logit, Nested Logit and Probit. The axes display the fixed level of rank polarization for two different options.