Supplemental Appendix to “A Random Attention Model”[*]

Matias D. Cattaneo[†] Xinwei Ma[‡] Yusufcan Masatlioglu[§] Elchin Suleymanov[¶]

August 28, 2019

Abstract

This Supplemental Appendix contains additional discussion of related literature, other numerical and methodological results, omitted theoretical proofs, and simulation evidence showcasing the finite sample performance of our proposed econometric procedures.

Contents

[*] Financial support from the National Science Foundation through grant SES-1628883 is gratefully acknowledged.

[†] Department of Operations Research and Financial Engineering, Princeton University.

[‡] Department of Economics, University of California at San Diego.

[§] Department of Economics, University of Maryland.

[¶] Department of Economics, Krannert School of Management, Purdue University.

SA.1 Related Literature

Manzini and Mariotti (2014) and Brady and Rehbeck (2016) are the two closest related papers to our work. Similar to ours, both models consider data of a single individual. Both provide parametric random attention models, which are described in the third and fourth examples in Section 2.1 of the main paper, respectively. Since their attention rules are monotonic, these models are two interesting special cases of our random attention model. To provide an accurate comparison, we need to introduce an outside/default option, which is required by both models. Thus, we first extend RAM to accommodate an outside option and then offer a detailed comparison between our work and both of these papers.

Let $a^* \notin X$ be the default option. In the model with a default option, we will allow an empty consideration set. Hence, now $\mu(\cdot|S)$ is defined over all subsets of S including the empty set. The default option is always available and can be interpreted as choosing nothing whenever the consideration set is empty. Let $X^* = X \cup \{a^*\}$ and $S^* = S \cup \{a^*\}$ for all $S \in \mathcal{X}$. We require that the choice rule satisfy $\sum_{a\in S^*} \pi(a|S) = 1$ and $\pi(a|S) \geq 0$ for all $a \in S^*$. We say that a choice rule π has a random attention representation with a default option if there exists a preference ordering $\succ$ on X and a monotonic attention rule μ such that for each $a \in S$, $\pi(a|S) = \sum_{T\subset S} \mathbb{1}(a \text{ is } \succ\text{-best in } T) \cdot \mu(T|S)$. Thus $\pi(a^*|S) = \mu(\emptyset|S)$.

An implication of Assumption 1 in the main paper is that for all S, $\mu(\emptyset|S) \leq \mu(\emptyset|S - a)$. In terms of choice probabilities, this implies that the default option satisfies regularity. In fact, it is easy to see that regularity on default and acyclicity of P are necessary and sufficient for π to have a RAM representation with a default option.

Remark SA.1. A choice rule π has a RAM representation with a default option if and only if it satisfies acyclicity of P and regularity on default. ⌟

In Manzini and Mariotti (2014), a choice rule has the representation

$$\pi(a|S) = \gamma(a) \prod_{b\in S: b\succ a} (1 - \gamma(b))$$

where $\gamma(a)$ is the fixed consideration probability of alternative a. See Horan (2018a) for an axiomatic characterization of this model when there is no default option. Demirkan and Kimya (2018) study a generalization of this model where γ depends on the menu. Independent consideration model is a special case of RAM with a default option if we set

$$\mu(T|S) = \prod_{a\in T} \gamma(a) \prod_{b\in S-T} (1-\gamma(b)).$$

In Brady and Rehbeck (2016), a choice rule has the representation

$$\pi(a|S) = \frac{\sum_{T\subset S} \mathbb{1}(a \in T \text{ is } \succ\text{-maximal in T}) \cdot w(T)}{\sum_{T\subset S} w(T)},$$

where $w(T)$ is the weight of consideration set T. This model is also a special case of RAM with an outside option if we set

$$\mu(T|S) = \frac{w(T)}{\sum_{T'\subset S} w(T')}.$$

Another closely related paper on random consideration sets is Aguiar (2015). In fact, acyclicity of P appears as an axiom in his representation. Hence, his model is also a special case of RAM.

Our paper is also related to the recent literature which attempts to combine heterogeneous preferences with random attention. The most closely related papers in this literature are Abaluck and Adams (2017), Barseghyan, Coughlin, Molinari, and Teitelbaum (2018), and Dardanoni, Manzini, Mariotti, and Tyson (2018). In a general setup, Abaluck and Adams (2017) show that, by exploiting asymmetries in cross-partial derivatives, consideration set probabilities and utility can be separately identified from observed choices when there is rich exogenous variation in observed covariates. Barseghyan et al. (2018) provide partial identification results when exogenous variation in observed covariates is more restricted. As opposed to our paper, both of these papers consider heterogeneous preferences and do not assume observable variation in menus decision makers face. On the other hand, we adopt abstract choice theory setup and our results do not rely on the existence of observed covariates. Lastly, similar to previous papers, Dardanoni et al. (2018) study choices from a fixed menu of alternatives. They consider aggregate choice where individuals might differ both in terms of their consideration capacities and preferences. Their model is complementary to ours as the attention rule each individual utilizes in their model can be considered a special case of RAM.

We also compare our model with other random choice models even though they are not in the framework of random consideration sets. Gul, Natenzon, and Pesendorfer (2014) consider the following model. There is a collection of attributes, denoted by Z. Each attribute $z \in Z$ has a weight $v(z)$, and each alternative $a \in X$ has intensity $i(a, z)$ in attribute z. The decision maker first randomly picks an attribute using Luce rule given the weights of all attributes. Then the decision maker picks an alternative using Luce rule given the intensities of all alternatives in attribute z. Gul, Natenzon, and Pesendorfer (2014) show that any attribute rule is a random utility model. Since RUM is a subset of RAM, any choice behavior that can be explained by an attribute rule can also be explained by RAM.

Echenique and Saito (2019) consider a general Luce model (GLM) where the decision maker uses Luce rule to choose among alternatives in her (deterministic) consideration set instead of the whole choice set. See Ahumada and Ulku (2018) for a characterization of GLM and Horan (2018b) for a related model. We show that RAM and GLM are distinct in terms of observed choices. Recall Example 2 in the main paper, which cannot be explained by RAM. This example can be explained by GLM by assuming that λ_i is the utility weight of the alternative i, and that whenever an alternative is not chosen that is because the alternative does not belong to the consideration set. Now consider a choice rule where all alternatives are always chosen with positive probability but Luce's IIA is not satisfied. We can construct such an example where acyclicity of P holds, and hence the choice rule has a RAM representation. However, this choice behavior cannot be explained by GLM as GLM reduces to Luce rule when all alternatives are chosen with positive probability in all menus. Echenique and Saito (2019) also consider two interesting special cases of GLM: (i) in two-stage Luce model, consideration sets are induced by an asymmetric, transitive binary relation, (ii) in threshold Luce model, an alternative belongs to the consideration set only if its utility is not too small compared to the utility of the maximal element in the choice set. It can be shown that RAM and two-stage Luce model are distinct, and threshold Luce model is a subset of both models.

Echenique, Saito, and Tserenjigmid (2018) propose a model (PALM) which uses violations of Luce's IIA to reveal perception priority of alternatives. For an example of stochastic choice data which can be explained by RAM but not PALM, consider any data where the outside option is never chosen. When the outside option is never chosen, PALM reduces to Luce rule. However, RAM allows for violations of Luce's IIA in the absence of an outside option. On the other hand, RAM with an outside option satisfies regularity on default, but PALM does not necessarily satisfy this property. Hence, RAM with an outside option does not nest PALM.

Fudenberg, Iijima, and Strzalecki (2015) consider a model (Additive Perturbed Utility-APU) where agents randomize as making deterministic choices can be costly. In their model, choices satisfy regularity. Since any choice rule that satisfies regularity has a RAM representation, RAM includes APU.

Aguiar, Boccardi, and Dean (2016) consider a satisficing model where the decision maker searches till she finds an alternative above the satisficing utility level. If there is no alternative above the satisficing utility level, the decision maker picks the best available alternative. They focus on two special cases of this model: (i) Full Support Satisficing Model (FSSM) where in any menu each alternative has a positive probability of being searched first, and (ii) Fixed Distribution Satisficing Model (FDSM). They show that FDSM is a subset of RUM, and hence it is also a subset of RAM. On the other hand, FSSM has no restrictions on observed choices if all alternatives are always chosen with positive probability. Hence, there exist choice rules that can be explained by FSSM but not RAM. Lastly, consider the choice data $\pi(a|\{a,b,c\}) = \pi(a|\{a,b\}) = \pi(a|\{a,c\}) = 1$ and $\pi(b|\{b,c\}) = 1/2$. FSSM cannot explain this choice behavior even though regularity is satisfied. Hence, FSSM and RAM are distinct, and FDSM is a subset of both.

SA.2 Other Examples of RAM

Here we provide more examples of random consideration sets that satisfy our key monotonicity assumption (Assumption 1 in the main paper).

1. (Full Attention) The decision maker considers everything with probability one: $\mu(T|S) = \mathbb{1}(T = S)$.

2. (Top N; Salant and Rubinstein, 2008) The decision maker faces a list of alternatives created by some ordering. She pays attention to the first N elements among available alternatives (e.g., first page of Google search results). If the number of available alternatives is less than N, she pays attention to the entire set. Formally, let $S(k,R)$ denote the set of first k elements in S according ordering R provided that $k \leq |S|$. ($S(|S|,R)$ is equal to S.) In our framework: $\mu(T|S) = \mathbb{1}(T = S(\min\{|S|,N\},R))$.

3. (Satisficing Consideration Set) The decision maker observes alternatives sequentially from a predetermined list. The order of alternatives is unknown to the decision maker in the beginning of the search and uncovers them during the search process. The decision maker stops searching upon finding a satisfactory alternative (Simon, 1955). If there is no such alternative, she searches the entire budget set. Formally, given the list L, $RS_L(S)$ denotes the range of search (the consideration set) when the budget set is S. In our framework: $\mu(T|S) = \mathbb{1}(T = RS_L(S))$.

4. (At Most k Alternatives) The decision maker considers at most k alternatives for any decision problem. If there are more alternatives than k, she considers only subsets including exactly k alternatives with equal probability. If there are less alternatives than k, she considers everything. In our framework:
$$\mu(T|S) = \begin{cases} 1 & \text{if } |S| \leq k \text{ and } T = S \\ \binom{|S|}{k}^{-1} & \text{if } |S| > k \text{ and } |T| = k \\ 0 & \text{otherwise} \end{cases}$$

5. (Uniform Consideration) The decision maker considers any subset of the feasible set with equal probabilities. That is, for all $T \subset S$, $\mu(T|S) = 1/(2^{|S|} - 1)$.

6. (Fixed Correlated Consideration; Barberà and Grodal, 2011; Aguiar, 2017) The decision maker pays attention to each alternative with a fixed probability but the consideration of alternatives is potentially correlated. Formally, let ω be a probability distribution over $\mathcal{X}$. Then each alternative $a \in X$ is considered with a fixed probability $\sum_{T \in \mathcal{X}} \mathbb{1}(a \in T) \cdot \omega(T)$ for all $S \ni a$. In our framework:
$$\mu(T|S) = \sum_{T' \in \mathcal{X}} \mathbb{1}(T' \cap S = T) \cdot \omega(T').$$

7. (Ordered Logit Attention) This is another generalization of the logit attention example. The decision maker ranks all subsets in terms of their attention priority, and she only considers subsets which are maximal with respect to that ordering. When there are several best subsets, the decision maker considers each of them with certain frequency as in the Logit Attention example. Thus, consideration sets are constructed in the spirit of standard maximization paradigm. Formally, let $\trianglerighteq$ be a complete and transitive priority order over subsets $\mathcal{X}$. $S \trianglerighteq T$ reads as "S has a higher attention priority than T". The case when S and T have the same attention priority is denoted by $S \bowtie T$. Formally,
$$\mu(T|S) = \begin{cases} \frac{w_T}{\sum_{T' \subset S} \mathbb{1}(T' \bowtie T) \cdot w_{T'}} & \text{if } T \text{ is } \trianglerighteq\text{-best in } S \\ 0 & \text{otherwise} \end{cases}$$

8. (Elimination by Aspects, General) The example is similar to the one given in the main text,

except that when the decision maker picks an irrelevant aspect, she selects a subset at random drawn from the uniform distribution. Formally,

$$\mu(T|S) = \sum_{B_i \cap S = T} \omega(i) + \frac{1}{2^{|S|}-1} \sum_{B_k \cap S = \emptyset} \omega(k).$$

The probability that the decision maker selects an irrelevant aspect is $\sum_{k: B_k \cap S = \emptyset} \omega(k)$. In this case, T is randomly chosen, which is reflected by the number $\frac{1}{2^{|S|}-1}$. We can generate similar examples. For instance, if the initial screening is not successful (choosing an irrelevant aspect), the decision maker may consider all the alternatives. Formally,

$$\mu(T|S) = \begin{cases} \sum_{B_i \cap S = T} \omega(i) & \text{if } T \neq S \\ \sum_{B_k \cap S = S, \emptyset} \omega(k) & \text{if } T = S \end{cases}$$

9. (Stochastic Satisficing) Suppose the satisficer faces multiple lists. The probability that the decision maker faces list L is denoted by $p(L)$. As opposed to Example 3, consideration sets are stochastic. Formally,

   $$\mu(T|S) = \sum_{L: T = RS_L(S)} p(L)$$

   where $RS_L(S)$ is the range of search when the budget set is S and L is the list.

10. ($1/N$ Rule) The decision maker utilizes N different sources of recommendations with equal probabilities. Given a fixed source s, she considers only top k_s alternatives according to ordering R_s that source s is using, which could be different from her preference. For example, she utilizes either Google or Yahoo with equal probability. Once she decides which one to look at, she pays attention to only the products appearing in the first page of the corresponding search result. Formally,

    $$\mu(T|S) = \frac{\big|\{s \| T = S(\min\{k_s, |S|\}, R_s)\}\big|}{N}$$

    where $S(k, R)$ denotes the first k alternatives in S according to R.

11. (Random Product Network; Masatlioglu and Suleymanov, 2017) Consider a decision maker faced with a product network $\mathcal{N}$. If $(a, b) \in \mathcal{N}$, then the decision maker who considers a is recommended alternative b (or alternatively, b is linked to a). The decision maker starts search from a random starting point. Given a realized starting point in the product network, the decision maker considers all alternatives which are directly or indirectly linked to that starting point. Formally, let η be a

probability distribution on X. Then

$$\mu(T|S) = \sum_{a\in S} \mathbb{1}(T = N_a(S)) \cdot \frac{\eta(a)}{\sum_{b\in S}\eta(b)}$$

where $N_a(S)$ denotes all alternatives which are directly or indirectly linked to a in S.

12. (Path Dependent Consideration; Suleymanov, 2018) This example is similar to the one above, except that now the decision maker starts searching from a fixed starting point $a^* \notin X$ (the default option) and takes a random path on a network. Let $X^* = X \cup \{a^*\}$, and $\mathcal{P}_{a^*}$ stands for all possible paths in X^* with the initial node a^*. When the choice set is X, the decision maker takes the path $\rho \in \mathcal{P}_{a^*}$ with probability $\gamma(\rho)$. Given $\rho \in \mathcal{P}_{a^*}$ and $S \in \mathcal{X}$, $\rho_S \in \mathcal{P}_{a^*}$ is the subpath of ρ in S with the same initial node a^*. For any ρ, let $V(\rho)$ be the vertices of the path ρ excluding the initial node. Then

$$\mu(T|S) = \sum_{\rho\in\mathcal{P}_{a^*}} \mathbb{1}(T = V(\rho_S)) \cdot \gamma(\rho).$$

This model is a subset of RAM with a default option.

SA.3 Limited Data

We discuss how our results can be adapted to handle limited data, that is, settings where not all possible choice probabilities or choice problems are observed. To investigate the implications of random attention models with limited data, assume that we observe choices from the set of choice problems $\mathcal{S}$. Let π_{obs} denote the observed choice behavior. We say that π_{obs} is consistent with the random attention model if there exists π defined on the entire domain $\mathcal{X}$ such that $\pi(a|S) = \pi_{\text{obs}}(a|S)$ for all $S \in \mathcal{S}$ and π is a RAM. We call such π an extension of π_{obs}. As de Clippel and Rozen (2014) point out, it is possible that π_{obs} satisfies the acyclicity of P even though it is inconsistent with RAM. Here we provide an example with stochastic choices in which the same problem occurs.

Example SA.1. Let $\mathcal{S} = \{\{a,b,c,d\},\{b,c,d\},\{a,c\}\}$. Consider the following choice rule.

$\pi_{\text{obs}}(\cdot|S)$	$S = \{a,b,c,d\}$	$\{b,c,d\}$	$\{a,c\}$
a	1/4		1/5
b	1/4	1/5	
c	1/4	3/5	4/5
d	1/4	1/5	

In this example, we observe choices only from 3 choice problems instead of 15 potential choice problems.

We show that these observations are sufficient to conclude that π_{obs} is not consistent with RAM. Suppose there exists a RAM π that extends π_{obs}. Then the observation $\pi(a|\{a,b,c,d\}) > \pi(a|\{a,c\})$ tells us that either $a\mathsf{P}b$ or $a\mathsf{P}d$. To see this, notice that at least one of $\pi(a|\{a,b,c,d\}) > \pi(a|\{a,c,d\})$ and $\pi(a|\{a,c,d\}) > \pi(a|\{a,c\})$ must hold or we get a contradiction. On the other hand, from $\pi(b|\{a,b,c,d\}) > \pi(b|\{b,c,d\})$ and $\pi(d|\{a,b,c,d\}) > \pi(d|\{b,c,d\})$ we learn that $d\mathsf{P}a$ and $b\mathsf{P}a$. Hence, even though acyclicity of P is satisfied on the limited domain, there does not exist an extension of π_{obs} that is RAM. $\lrcorner$

From the example above we can note the following: if $\pi_{\text{obs}}(a|S) > \pi_{\text{obs}}(a|S-A)$, then it must be that $a\mathsf{P}b$ for some $b \in A$. Now suppose $\pi_{\text{obs}}(a|S) + \pi_{\text{obs}}(b|S') > 1$ and $\{a,b\} \subset S \cap S'$ where $a \neq b$. Then if π is an extension of π_{obs} it has to be the case that either $\pi(a|S) > \pi(a|S \cap S')$ or $\pi(b|S) > \pi(b|S \cap S')$. Hence, either $a\mathsf{P}c$ for some $c \in S - S'$ or $b\mathsf{P}d$ for some $d \in S' - S$. This is exactly the probabilistic analog of the condition in de Clippel and Rozen (2014). However, this condition is not enough in probabilistic domain. The next example illustrates this point.

Example SA.2. Consider the following choice rule.

$\pi_{\text{obs}}(\cdot\|S)$	$S=\{a,b,c,d\}$	$\{a,b,c,e\}$	$\{a,b,d\}$	$\{a,c,d\}$	$\{b,c,e\}$
a	1/4	2/3	1/2	1/2	
b	1/4	1/6	1/2		5/6
c	1/4	0		1/2	1/12
d	1/4		0	0	
e		1/6			1/12

First, $\pi_{\text{obs}}(d|\{a,b,d\}) < \pi_{\text{obs}}(d|\{a,b,c,d\})$ and $\pi_{\text{obs}}(d|\{a,c,d\}) < \pi_{\text{obs}}(d|\{a,b,c,d\})$ imply that $d\mathsf{P}c$ and $d\mathsf{P}b$. Furthermore, $\pi_{\text{obs}}(e|\{b,c,e\}) < \pi_{\text{obs}}(e|\{a,b,c,e\})$ implies that $e\mathsf{P}a$. Now consider the set $\{a,b,c\}$ and notice that $\pi_{\text{obs}}(a|\{a,b,c,e\}) + \pi_{\text{obs}}(b|\{a,b,c,d\}) + \pi_{\text{obs}}(c|\{a,b,c,d\}) > 1$. Thus if we had observations on the choice problem $\{a,b,c\}$ one of the following would have been true: (i) the probability that a is chosen decreases when e is removed from the choice problem $\{a,b,c,e\}$, (ii) the probability that b is chosen decreases when d is removed from the choice problem $\{a,b,c,d\}$, or (iii) the probability that c is chosen decreases when d is removed from the choice problem $\{a,b,c,d\}$. Hence, one of the following must be true: $a\mathsf{P}e$, $b\mathsf{P}d$, or $c\mathsf{P}d$. Since we have a contradiction in all cases, π_{obs} is inconsistent with RAM. $\lrcorner$

We generalize the intuition from this example. Suppose there exists a collection of pairs $(a_i, S_i)_{i=1}^m$ such that $\{a_1,\dots,a_m\} \subset \bigcap_{i=1}^m S_i$ and $\sum_{i=1}^m \pi_{\text{obs}}(a_i, S_i) > 1$ where a_i are all distinct. Now in the choice problem $\bigcap_{i=1}^m S_i$ the probability that a_i is chosen must decrease for at least one $i \in \{1,\dots,m\}$. From here we can conclude that $a_i\mathsf{P}b_i$ for some $b_i \in S_i - (S_1 \cap \cdots \cap S_{i-1} \cap S_{i+1} \cap \cdots \cap S_m)$ for some $i \in \{1,\dots,m\}$. Hence, the existence of an acyclic P that satisfies this condition is necessary for π_{obs} to be consistent with RAM. Theorem SA.1 shows that it is also sufficient.

Theorem SA.1. A choice rule π_{obs} is consistent with RAM if and only if there exists an acyclic binary relation P on X which satisfies the following: for any collection $(a_i, S_i)_{i=1}^m$ with distinct a_i such that $\{a_1, \dots, a_m\} \subset \bigcap_{i=1}^m S_i$ and $\sum_{i=1}^m \pi_{\text{obs}}(a_i, S_i) > 1$, $a_i \mathsf{P} b_i$ for some $b_i \in S_i - (S_1 \cap \dots \cap S_{i-1} \cap S_{i+1} \cap \dots \cap S_m)$ for some $i \in \{1, \dots, m\}$.

Proof. Let $\succ$ be a transitive completion of P. We reorder the alternatives in X such that $a_{1,\succ} \succ \dots \succ a_{K,\succ}$. We define μ as follows. For any $S \in \mathcal{S}$,

$$\mu(T|S) = \begin{cases} \pi_{\text{obs}}(a_{k,\succ}|S) & \text{if } \exists\ k \ s.t.\ T = L_{k,\succ} \cap S \\ 0 & \text{otherwise} \end{cases}$$

where $L_{k,\succ} = \{a_{k,\succ}, \dots, a_{K,\succ}\}$. For any $S \in \mathcal{X} - \mathcal{S}$, if there is $S' \in \mathcal{S}$ with $S' \supset S$, then

$$\mu(T|S) = \max_{S' \in \mathcal{S}: S' \supset S} \mu(T|S') \qquad \text{if } T \subsetneq S$$

and $\mu(S|S) = 1 - \sum_{T \subsetneq S} \mu(T|S)$. Finally, for $S \in \mathcal{X} - \mathcal{S}$, if there is no $S' \in \mathcal{S}$ with $S' \supset S$, then $\mu(S|S) = 1$ and $\mu(T|S) = 0$ for all $T \subsetneq S$.

It is easily seen that $(\succ, \mu)$ represents π_{obs}. We first need to show that $\mu(\cdot|S)$ is a probability distribution. The only case we need to check is when $S \in \mathcal{X} - \mathcal{S}$ and there exists $S' \supset S$ with $S' \in \mathcal{S}$. We need to show that $\mu(S|S) \geq 0$ or that $\sum_{T \subsetneq S} \mu(T|S) \leq 1$. Suppose $\sum_{T \subsetneq S} \mu(T|S) > 1$. By definition, for each $T \subsetneq S$ such that $\mu(T|S) > 0$, there exists a pair (a_{k_T}, S_T) such that $S_T \in \mathcal{S}$ with $S_T \supset S$ and $\mu(T|S) = \mu(T|S_T) = \pi_{\text{obs}}(a_{k_T}|S_T)$ where $T = L_{a_{k_T},\succ} \cap S_T$. Then, $\sum_{T \subsetneq S} \pi_{\text{obs}}(a_{k_T}|S_T) > 1$. Notice that since μ is triangular, a_{k_T} are distinct. By definition of P, there exist $T \subsetneq S$ and an alternative b_{k_T} in $S_T - S$ such that $a_{k_T} \mathsf{P} b_{k_T}$. But this is a contradiction as $b_{k_T} \notin T$ and $T = L_{a_{k_T},\succ} \cap S_T$.

We now need to show that μ defined as above is monotonic. We have a few cases to consider.

Case 1: $S, S - b \in \mathcal{S}$. Suppose $\mu(T|S) > \mu(T|S - b)$ where $b \notin T$. Since $S \in \mathcal{S}$ and $\mu(T|S) > 0$ it must be that $T = L_{k,\succ} \cap S$ for some k and $\mu(T|S) = \pi_{\text{obs}}(a_{k,\succ}|S)$. Since $b \notin T$ and $T = L_{k,\succ} \cap S$ we must have $b \succ a_{k,\succ}$. Therefore, it must be the case that $T = L_{k,\succ} \cap (S - b)$. By definition, $\mu(T|S - b) = \pi_{\text{obs}}(a_{k,\succ}|S - b)$. But then we have $\pi_{\text{obs}}(a_{k,\succ}|S) > \pi_{\text{obs}}(a_{k,\succ}|S - b)$ which implies that $a_{k,\succ} \mathsf{P} b$, a contradiction.

Case 2: $S \in \mathcal{S}$, $S - b \notin \mathcal{S}$. Let T with $b \notin T$ given. Since $S \in \mathcal{S}$ it must be that either $\mu(T|S) = 0$ in which case monotonicity is trivial or $T = L_{k,\succ} \cap S$ for some k and $\mu(T|S) = \pi_{\text{obs}}(a_{k,\succ}|S)$. First, suppose $T \subsetneq S - b$. Now $S \supset S - b$, and hence by definition, $\mu(T|S - b) = \max_{S' \in \mathcal{S}: S' \supset S - b} \mu(T|S') \geq \mu(T|S)$. This establishes that the claim holds for all $T \subsetneq S - b$. Notice that if $\mu(T|S-b) = \mu(T|S)$ for all $T \subsetneq S - b$, then $\mu(S-b|S-b) \geq \mu(S-b|S)$ also follows.

Suppose $\mu(S-b|S) > \mu(S-b|S-b)$. Then since $\mu(S-b|S) > 0$ and μ is triangular we must have that b is $\succ$ maximal in S. Thus $\mu(S|S) = \pi_{\text{obs}}(b|S)$. Furthermore, by the argument in the previous paragraph, there exists $T \subsetneq S - b$ such that $\mu(T|S-b) > \mu(T|S)$. Suppose there exists only one such T. (A similar argument will work if there is more than one such T.) By definition, there exists $S_T \in \mathcal{S}$ with $S_T \supset S - b$ such that $\mu(T|S_T) > \mu(T|S)$. Thus there exists a_{k_T} such that $\mu(T|S-b) = \mu(T|S_T) = \pi_{\text{obs}}(a_{k_T}|S_T)$ and $T = L_{a_{k_T},\succ} \cap S_T = L_{a_{k_T},\succ} \cap (S-b)$ where the second equality follows from the fact that $T \subset S - b$. Notice that $a_{k_T} \in S - b$ and since b is $\succ$ maximal in S we have $b \succ a_{k_T}$. Hence we have $T = L_{a_{k_T},\succ} \cap S$ which by definition implies $\mu(T|S) = \pi_{\text{obs}}(a_{k_T}|S) < \pi_{\text{obs}}(a_{k_T}|S_T)$. Now since by assumption $\mu(T'|S) = \mu(T'|S-b)$ for all $T' \subsetneq S - b$ with $T' \neq T$, by using the definition of μ, $\mu(S-b|S) > \mu(S-b|S-b)$ implies that $\pi_{\text{obs}}(b|S) + \pi_{\text{obs}}(a_{k_T}|S) < \pi_{\text{obs}}(a_{k_T}|S_T)$. Consider the collection $\{(a_{k_T}, S_T)\} \cup \{(a_i, S) | a_i \in S \text{ and } a_i \neq b, a_i \neq a_{k_T}\}$. Since $\pi_{\text{obs}}(a_{k_T}|S_T) + (1 - \pi_{\text{obs}}(a_{k_T}|S) - \pi_{\text{obs}}(b|S)) > 1$, the observed choice probabilities summed over this collection adds up to greater than one. By definition of P, either there exists an alternative in $a_i \in S - b$ such that $a_i \mathsf{P} b$ or there exists an alternative $c \in S_T - (S-b)$ such that $a_{k_T} \mathsf{P} c$. The first case leads to a contradiction since b is $\succ$ maximal in S. The second case leads to a contradiction since $T = L_{a_{k_T},\succ} \cap S_T = L_{a_{k_T},\succ} \cap (S-b) \not\ni c$.

Case 3: $S \notin \mathcal{S}$, $S - b \in \mathcal{S}$. If there exists no $S' \supset S$ such that $S' \in \mathcal{S}$, then monotonicity property is trivial as $\mu(S|S) = 1$. Hence, suppose there is $S' \supset S$ such that $S' \in \mathcal{S}$ and there exists $T \subset S - b$ such that $\mu(T|S) > \mu(T|S-b)$. Let a_{k_T} and S_T be such that $S_T \supset S$ and $\mu(T|S) = \pi_{\text{obs}}(a_{k_T}|S_T)$ where $T = L_{a_{k_T},\succ} \cap S_T$. Also notice that since $a_{k_T} \in T \subset S - b$, $a_{k_T} \neq b$. By definition of μ, it must be that $\pi_{\text{obs}}(a_{k_T}|S_T) > \pi_{\text{obs}}(a_{k_T}|S-b)$. Now consider the collection $\{(a_{k_T}, S_T)\} \cup \{(a_i, S-b) | a_i \in S - b, a_i \neq a_{k_T}\}$. If we add the choice probabilities over this collection, they will add up to greater than one. Hence, by definition of P, there exists $c \in S_T - (S-b)$ such that $a_{k_T} \mathsf{P} c$. But this is a contradiction as $T = L_{a_{k_T},\succ} \cap S_T$ and $T \subset S - b$.

Case 4: $S \notin \mathcal{S}$, $S - b \notin \mathcal{S}$. If there is no $S' \supset S$ such that $S' \in \mathcal{S}$, then the claim is trivial. Suppose there exists such S'. Consider $T \subsetneq S - b$. By definition,

$$\mu(T|S) = \underset{S' \supset S : S' \in \mathcal{S}}{\mu}(T|S') \leq \underset{S'' \supset S-b : S'' \in \mathcal{S}}{\mu}(T|S'') = \mu(T|S-b).$$

Hence, the claim holds for all $T \subsetneq S - b$. We need to show that the claim also holds when $T = S - b$. Notice that if $\mu(T|S) = \mu(T|S-b)$ for all $T \subsetneq S - b$, then $\mu(S-b|S) \leq \mu(S-b|S-b)$ follows immediately. Hence suppose there is at least one T such that $\mu(T|S) < \mu(T|S-b)$.

Now if b is not $\succ$ maximal in S, then $\mu(S-b|S) = 0$ and monotonicity is trivial. So suppose b is $\succ$ maximal and $\mu(S-b|S) > \mu(S-b|S-b)$. For each $T \subsetneq S-b$ such that $\mu(T|S-b) > 0$, let a_{k_T} and S_T be such that $\mu(T|S-b) = \pi_{\text{obs}}(a_{k_T}|S_T)$. Let $a_{k_{S-b}}$ and S_{S-b} be such that $\mu(S-b|S) = \pi_{\text{obs}}(a_{k_{S-b}}|S_{S-b})$. Consider the collection $\{(a_{k_T}, S_T)|\ T \subset S-b\}$. Since $\mu(S-b|S) > \mu(S-b|S-b)$, if we sum choice probabilities over this collection, they add up to greater than one. This implies that $a_{k_T}\mathsf{P}c$ for some $T \subset S-b$ and for some $c \notin S-b$. But this is a contradiction since only sets of the form $T = L_{a_{k_T},\succ} \cap S_T$ are considered with positive probability and $c \notin T$. ■

SA.4 Estimation and Inference

Here we first provide the proof of Theorem 4 in the main paper. We then discuss the computational cost of our inference procedure, following which we connect our inference approach to the "J-test" method. Then, we demonstrate how the matrix $\mathbf{R}_\succ$ can be constructed in a limited data scenario. Finally we discuss other possible ways to construct the critical value.

SA.4.1 Proof of Theorem 4

For completeness and notational clarity, we first present two lemmas, which follow from the central limit theorem.

Lemma SA.1. Under Assumption 2 in the main paper,

$$\sqrt{N_S}\Big(\hat{\boldsymbol{\pi}}_S - \boldsymbol{\pi}_S\Big) \rightsquigarrow \mathcal{N}\Big(\mathbf{0},\ \boldsymbol{\Omega}_{\pi,S}\Big), \qquad \boldsymbol{\Omega}_{\pi,S} = \text{diag}(\boldsymbol{\pi}_S) - \boldsymbol{\pi}_S\boldsymbol{\pi}_S',$$

where $N_S = \sum_{1\leq i\leq N} \mathbb{1}(Y_i = S)$ is the effective sample size of menu S, $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_{\pi,S})$ is the $|S|$-dimensional Gaussian distribution with mean zero and covariance $\boldsymbol{\Omega}_{\pi,S}$, and diag is the operator constructing diagonal matrices.

The asymptotic distribution is degenerate for two reasons. First, the choice rule, by construction, has to sum up to 1. Second, it is possible that some of the alternatives in S is never chosen (either in the sample or the population).

Lemma SA.2. Under Assumption 2 in the main paper,

$$\sqrt{N}\Big(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\Big) \rightsquigarrow \mathcal{N}\Big(\mathbf{0},\ \boldsymbol{\Omega}_\pi\Big),$$

where the asymptotic variance $\boldsymbol{\Omega}_\pi$ is block diagonal, with blocks given by $\frac{1}{\mathbb{P}[Y_i=S]}\boldsymbol{\Omega}_{\pi,S}$, $S \in \mathcal{X}$. (Block diagonality of the asymptotic variance follows directly from the fact that across choice problems, choice rules are estimated with independent samples, hence are independent.)

Recall that $\mathscr{T}(\succ) = \sqrt{N}\max((\mathbf{R}_\succ\hat{\boldsymbol{\pi}}) \oslash \hat{\boldsymbol{\sigma}}_\succ)_+$. The next proposition shows that its distribution is approximated by the infeasible statistic $\sqrt{N}\max(\mathbf{R}_\succ(\mathbf{z}^\star + \boldsymbol{\pi}) \oslash \hat{\boldsymbol{\sigma}}_\succ)_+$. For clarity, we do not seek uniformity here. Later we will show how the same argument can be used to demonstrate distributional approximation uniformly in a class of DGPs (see Proposition SA.2).

Proposition SA.1. Suppose Assumption 2 in the main paper holds and $\min(\boldsymbol{\sigma}_{\pi,\succ}) > 0$. Then

$$\Big|\mathbb{P}\Big[\mathscr{T}(\succ) \le t\Big] - \mathbb{P}^\star\Big[\sqrt{N}\max(\mathbf{R}_\succ(\mathbf{z}^\star + \boldsymbol{\pi}) \oslash \hat{\boldsymbol{\sigma}}_\succ)_+ \le t\Big]\Big| \to_{\mathbb{P}} 0, \qquad \forall t \neq 0.$$

Remark SA.2. We exclude $t = 0$ since the limiting distribution may have a point mass at the origin. $\lrcorner$

Proof. Let f be a bounded Lipschitz function. Without loss of generality assume its Lipschitz constant is 1, and $2 \cdot \|f\|_\infty = c$. For convenience, denote $\mathbb{E}[f(\cdot)]$ by $\mathbb{E}_f[\cdot]$.

By the central limit theorem, it is possible to construct a (sequence of) random vector $\tilde{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_\pi/N)$ such that $|\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) - \sqrt{N}\tilde{\mathbf{z}}| = O_{\mathbb{P}}(1/\sqrt{N})$ (we postpone the proof to the end). Further, let $\mathbf{w}^\star \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/\sqrt{N})$ with suitable dimension such that $\hat{\boldsymbol{\Omega}}^{1/2}\mathbf{w}^\star \sim \mathbf{z}^\star$. Then consider bounds on the following quantities.

$$\begin{aligned}
&\left|\mathbb{E}_f\left[\sqrt{N}\max\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}} \oslash \hat{\boldsymbol{\sigma}}_\succ\right)_+\right] - \mathbb{E}_f\left[\sqrt{N}\max\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}} \oslash \boldsymbol{\sigma}_{\pi,\succ}\right)_+\right]\right| \\
&\le \mathbb{E}\left[\left|\sqrt{N}\max\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}} \oslash \hat{\boldsymbol{\sigma}}_\succ\right)_+ - \sqrt{N}\max\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}} \oslash \boldsymbol{\sigma}_{\pi,\succ}\right)_+\right| \wedge c\right] \\
&\le \mathbb{E}\left[\left\|\sqrt{N}\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}} \oslash \hat{\boldsymbol{\sigma}}_\succ\right)_+ - \sqrt{N}\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}} \oslash \boldsymbol{\sigma}_{\pi,\succ}\right)_+\right\|_\infty \wedge c\right] \\
&= \mathbb{E}\left[\left\|\sqrt{N}\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}} \oslash \hat{\boldsymbol{\sigma}}_\succ - \mathbf{R}_\succ\hat{\boldsymbol{\pi}} \oslash \boldsymbol{\sigma}_{\pi,\succ}\right)_+ \odot \mathbb{1}\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}} \ge 0\right)\right\|_\infty \wedge c\right] \\
&\le \mathbb{E}\left[\left\|\sqrt{N}\left(\mathbf{R}_\succ(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \oslash \hat{\boldsymbol{\sigma}}_\succ - \mathbf{R}_\succ(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \oslash \boldsymbol{\sigma}_{\pi,\succ}\right)_+\right\|_\infty \wedge c\right] \\
&\qquad + \mathbb{E}\left[\left\|\sqrt{N}\left(\mathbf{R}_\succ\boldsymbol{\pi} \oslash \hat{\boldsymbol{\sigma}}_\succ - \mathbf{R}_\succ\boldsymbol{\pi} \oslash \boldsymbol{\sigma}_{\pi,\succ}\right)_+ \odot \mathbb{1}\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}} \ge 0\right)\right\|_\infty \wedge c\right].
\end{aligned}$$

The first inequality uses Lipschitz property of f and the fact that the whole term is bounded by $2 \cdot \|f\|_\infty = c$. The second inequality uses basic property of the max operator. The third inequality follows from triangle inequality of the norm $\|\cdot\|_\infty$.

We further split in order to control the denominator:

$$\begin{aligned}
&\text{previous display} \\
&\le \mathbb{E}\left[\left\|\sqrt{N}\left(\mathbf{R}_\succ(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \oslash \hat{\boldsymbol{\sigma}}_\succ - \mathbf{R}_\succ(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \oslash \boldsymbol{\sigma}_{\pi,\succ}\right)_+\right\|_\infty \mathbb{1}\left(\min(\hat{\boldsymbol{\sigma}}_\succ) \ge \min(\boldsymbol{\sigma}_{\pi,\succ})/2\right) \wedge c\right]
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{P}\Big[\min(\hat{\boldsymbol{\sigma}}_{\succ}) < \min(\boldsymbol{\sigma}_{\pi,\succ})/2\Big]\cdot c\\
&+\mathbb{E}\Big[\Big\|\sqrt{N}\Big(\mathbf{R}_{\succ}\boldsymbol{\pi}\oslash\hat{\boldsymbol{\sigma}}_{\succ}-\mathbf{R}_{\succ}\boldsymbol{\pi}\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}\odot\mathbb{1}\Big(\mathbf{R}_{\succ}\hat{\boldsymbol{\pi}}\geq 0\Big)\Big\|_{\infty}\wedge c\Big]\\
=\;& O\Big(\frac{1}{\sqrt{N}}\Big)+\mathbb{E}\Big[\Big\|\sqrt{N}\Big(\mathbf{R}_{\succ}\boldsymbol{\pi}\oslash\hat{\boldsymbol{\sigma}}_{\succ}-\mathbf{R}_{\succ}\boldsymbol{\pi}\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}\odot\mathbb{1}\Big(\mathbf{R}_{\succ}\hat{\boldsymbol{\pi}}\geq 0\Big)\Big\|_{\infty}\wedge c\Big]\\
\leq\;& O\Big(\frac{1}{\sqrt{N}}\Big)+\mathbb{E}\Big[\Big\|\sqrt{N}\Big(\mathbf{R}_{\succ}\boldsymbol{\pi}\oslash\hat{\boldsymbol{\sigma}}_{\succ}-\mathbf{R}_{\succ}\boldsymbol{\pi}\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}\odot\mathbb{1}\Big(\mathbf{R}_{\succ}\boldsymbol{\pi}\geq -\frac{a_n}{\sqrt{N}}\Big)\Big\|_{\infty}\wedge c\Big]\\
&+\mathbb{E}\Big[\Big\|\sqrt{N}\Big(\mathbf{R}_{\succ}\boldsymbol{\pi}\oslash\hat{\boldsymbol{\sigma}}_{\succ}-\mathbf{R}_{\succ}\boldsymbol{\pi}\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}\odot\mathbb{1}\Big(\mathbf{R}_{\succ}\hat{\boldsymbol{\pi}}\geq 0,\ \mathbf{R}_{\succ}\boldsymbol{\pi}< -\frac{a_n}{\sqrt{N}}\Big)\Big\|_{\infty}\wedge c\Big]\\
\leq\;& O\Big(\frac{1}{\sqrt{N}}\Big)+\mathbb{E}\Big[\Big\|\sqrt{N}\Big(\mathbf{R}_{\succ}\boldsymbol{\pi}\oslash\hat{\boldsymbol{\sigma}}_{\succ}-\mathbf{R}_{\succ}\boldsymbol{\pi}\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}\odot\mathbb{1}\Big(\mathbf{R}_{\succ}\boldsymbol{\pi}\geq -\frac{a_n}{\sqrt{N}}\Big)\Big\|_{\infty}\wedge c\Big]\\
&+\mathbb{E}\Big[\Big\|\mathbb{1}\Big(\mathbf{R}_{\succ}\hat{\boldsymbol{\pi}}\geq 0,\ \mathbf{R}_{\succ}\boldsymbol{\pi}< -\frac{a_n}{\sqrt{N}}\Big)\Big\|_{\infty}\wedge c\Big]\\
=\;& O\Big(\frac{1}{\sqrt{N}}+\frac{a_n}{\sqrt{N}}+\frac{1}{a_n}\Big),
\end{aligned}
$$

with $a_n \to \infty$ chosen so that the last line vanishes.

$$
\begin{aligned}
&\Big|\mathbb{E}_f\Big[\sqrt{N}\max\Big(\mathbf{R}_{\succ}\hat{\boldsymbol{\pi}}\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}\Big]-\mathbb{E}_f\Big[\sqrt{N}\max\Big(\mathbf{R}_{\succ}(\tilde{\mathbf{z}}+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}\Big]\Big|\\
\leq\;&\mathbb{E}\Big[\Big\|\sqrt{N}\Big(\mathbf{R}_{\succ}\hat{\boldsymbol{\pi}}\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}-\sqrt{N}\Big(\mathbf{R}_{\succ}(\tilde{\mathbf{z}}+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}\Big\|_{\infty}\wedge c\Big]\\
\leq\;&\mathbb{E}\Big[\sqrt{N}\Big\|\mathbf{R}_{\succ}\hat{\boldsymbol{\pi}}\oslash\boldsymbol{\sigma}_{\pi,\succ}-\mathbf{R}_{\succ}(\tilde{\mathbf{z}}+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big\|_{\infty}\wedge c\Big]\\
=\;&\mathbb{E}\Big[\sqrt{N}\Big\|\mathbf{R}_{\succ}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}-\mathbf{R}_{\succ}\tilde{\mathbf{z}}\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big\|_{\infty}\wedge c\Big]\\
=\;&O_{\mathbb{P}}\Big(\frac{1}{\sqrt{N}}\Big).
\end{aligned}
$$

The first inequality uses Lipschitz continuity of f and property of the max operator. The second inequality drops positivity. The rate in the last line comes from the coupling requirement of $\tilde{\mathbf{z}}$.

$$
\Big|\mathbb{E}_f\Big[\sqrt{N}\max\Big(\mathbf{R}_{\succ}(\tilde{\mathbf{z}}+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}\Big]-\mathbb{E}^{\star}_f\Big[\sqrt{N}\max\Big(\mathbf{R}_{\succ}(\boldsymbol{\Omega}_{\pi}^{1/2}\mathbf{w}^{\star}+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}\Big]\Big|=0,
$$

since the two terms have the same distribution.

$$
\begin{aligned}
&\Big|\mathbb{E}^{\star}_f\Big[\sqrt{N}\max\Big(\mathbf{R}_{\succ}(\boldsymbol{\Omega}_{\pi}^{1/2}\mathbf{w}^{\star}+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}\Big]-\mathbb{E}^{\star}_f\Big[\sqrt{N}\max\Big(\mathbf{R}_{\succ}(\hat{\boldsymbol{\Omega}}^{1/2}\mathbf{w}^{\star}+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}\Big]\Big|\\
\leq\;&\mathbb{E}^{\star}\Big[\Big\|\sqrt{N}\Big(\mathbf{R}_{\succ}(\boldsymbol{\Omega}_{\pi}^{1/2}\mathbf{w}^{\star}+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}-\sqrt{N}\Big(\mathbf{R}_{\succ}(\hat{\boldsymbol{\Omega}}^{1/2}\mathbf{w}^{\star}+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_{+}\Big\|_{\infty}\wedge c\Big]\\
\leq\;&\mathbb{E}^{\star}\Big[\sqrt{N}\Big\|\mathbf{R}_{\succ}(\boldsymbol{\Omega}_{\pi}^{1/2}\mathbf{w}^{\star}+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}-\mathbf{R}_{\succ}(\hat{\boldsymbol{\Omega}}^{1/2}\mathbf{w}^{\star}+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big\|_{\infty}\wedge c\Big]\\
=\;&\mathbb{E}^{\star}\Big[\sqrt{N}\Big\|\mathbf{R}_{\succ}\boldsymbol{\Omega}_{\pi}^{1/2}\mathbf{w}^{\star}\oslash\boldsymbol{\sigma}_{\pi,\succ}-\mathbf{R}_{\succ}\hat{\boldsymbol{\Omega}}^{1/2}\mathbf{w}^{\star}\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big\|_{\infty}\wedge c\Big]
\end{aligned}
$$

$$= O_{\mathbb{P}}\Big(\frac{1}{\sqrt{N}}\Big).$$

The first inequality uses Lipschitz continuity of f and property of the max operator. The second inequality drops positivity. The rate in the last line comes from the fact $\sqrt{N}\|\hat{\mathbf{\Omega}}-\mathbf{\Omega}_\pi\|_\infty = O_{\mathbb{P}}(1)$. Also, by construction $\mathbf{w}^\star = O_{\mathbb{P}^\star}(1/\sqrt{N})$.

$$\begin{aligned}
&\left|\mathbb{E}_f^\star\left[\sqrt{N}\max\left(\mathbf{R}_\succ(\hat{\mathbf{\Omega}}^{1/2}\mathbf{w}^\star+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\right)_+\right]-\mathbb{E}_f^\star\left[\sqrt{N}\max\left(\mathbf{R}_\succ(\hat{\mathbf{\Omega}}^{1/2}\mathbf{w}^\star+\boldsymbol{\pi})\oslash\hat{\boldsymbol{\sigma}}_\succ\right)_+\right]\right|\\
&\leq \mathbb{E}^\star\left[\left\|\sqrt{N}\Big(\mathbf{R}_\succ(\hat{\mathbf{\Omega}}^{1/2}\mathbf{w}^\star+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\Big)_+-\sqrt{N}\Big(\mathbf{R}_\succ(\hat{\mathbf{\Omega}}^{1/2}\mathbf{w}^\star+\boldsymbol{\pi})\oslash\hat{\boldsymbol{\sigma}}_\succ\Big)_+\right\|_\infty\wedge c\right]\\
&= \mathbb{E}^\star\left[\left\|\sqrt{N}\Big(\mathbf{R}_\succ(\hat{\mathbf{\Omega}}^{1/2}\mathbf{w}^\star+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}-\mathbf{R}_\succ(\hat{\mathbf{\Omega}}^{1/2}\mathbf{w}^\star+\boldsymbol{\pi})\oslash\hat{\boldsymbol{\sigma}}_\succ\Big)_+\odot\mathbb{1}\Big(\mathbf{R}_\succ(\hat{\mathbf{\Omega}}^{1/2}\mathbf{w}^\star+\boldsymbol{\pi})\geq 0\Big)\right\|_\infty\wedge c\right]\\
&\leq \mathbb{E}^\star\left[\left\|\sqrt{N}\Big(\mathbf{R}_\succ\hat{\mathbf{\Omega}}^{1/2}\mathbf{w}^\star\oslash\boldsymbol{\sigma}_{\pi,\succ}-\mathbf{R}_\succ\hat{\mathbf{\Omega}}^{1/2}\mathbf{w}^\star\oslash\hat{\boldsymbol{\sigma}}_\succ\Big)_+\right\|_\infty\wedge c\right]\\
&\qquad+\mathbb{E}^\star\left[\left\|\sqrt{N}\Big(\mathbf{R}_\succ\boldsymbol{\pi}\oslash\boldsymbol{\sigma}_{\pi,\succ}-\mathbf{R}_\succ\boldsymbol{\pi}\oslash\hat{\boldsymbol{\sigma}}_\succ\Big)_+\odot\mathbb{1}\Big(\mathbf{R}_\succ(\hat{\mathbf{\Omega}}^{1/2}\mathbf{w}^\star+\boldsymbol{\pi})\geq 0\Big)\right\|_\infty\wedge c\right]\\
&= O_{\mathbb{P}}\left(\frac{1}{\sqrt{N}}+\frac{a_n}{\sqrt{N}}+\frac{1}{a_n}\right).
\end{aligned}$$

The first inequality uses Lipschitz continuity of f and property of the max operator. The second inequality applies triangle inequality to the norm $\|\cdot\|_\infty$. The rest is essentially the same as that of the first part.

The only missing part is to show the existence of the coupling variable $\tilde{\mathbf{z}}$. Since the choice probabilities are averages of indicators, Corollary 4.1 of Chen, Goldstein, and Shao (2010) implies the following non-asymptotic bound on the Wasserstein metric:

$$\inf\Big\{\mathbb{E}\left[|X-Y|\right]:\ X\sim\sqrt{N}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}),\ Y\sim\mathcal{N}(\mathbf{0},\mathbf{\Omega}_\pi)\Big\}\leq \text{Const.}\sqrt{\frac{1}{\min_{S\in\mathcal{X}}N_S}},$$

where the infimum is taken over all joint distributions with the given marginals, and the constant in the above display is universal. By Assumption 2 in the main paper, the rate on the RHS is proportional to $\sqrt{1/N}$. Existence of the coupling variable follows from the bounds on the Wasserstein metric.

Hence we showed that

$$\left|\mathbb{E}_f\left[\sqrt{N}\max\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}}\oslash\hat{\boldsymbol{\sigma}}_\succ\right)_+\right]-\mathbb{E}_f^\star\left[\sqrt{N}\max\left(\mathbf{R}_\succ(\hat{\mathbf{\Omega}}^{1/2}\mathbf{w}^\star+\boldsymbol{\pi})\oslash\hat{\boldsymbol{\sigma}}_\succ\right)_+\right]\right|=o_{\mathbb{P}}(1).$$

■

Now we show how the previous result can be generalized to be uniform among a class of distributions. The main argument used in Proposition SA.1 remains almost unchanged.

Proposition SA.2. Under the assumptions of Theorem 4 in the main paper,

$$\sup_{\pi\in\Pi}\mathbb{P}\left[\left|\mathbb{P}\left[\mathscr{T}(\succ)\le t\right]-\mathbb{P}^\star\left[\sqrt{N}\max(\mathbf{R}_\succ(\mathbf{z}^\star+\boldsymbol{\pi})\oslash\hat{\boldsymbol{\sigma}}_\succ)_+\le t\right]\right|>\varepsilon\right]\to 0,\qquad \forall\varepsilon>0,\ \forall t\neq 0.$$

Proof. The proof remains almost the same, while extra care is needed for the coupling argument, which we demonstrate below. We would like to bound the following quantity *uniformly*:

$$\left|\mathbb{E}_f\left[\sqrt{N}\max\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}}\oslash\boldsymbol{\sigma}_{\pi,\succ}\right)_+\right]-\mathbb{E}^\star_f\left[\sqrt{N}\max\left(\mathbf{R}_\succ(\boldsymbol{\Omega}_\pi^{1/2}\mathbf{w}^\star+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\right)_+\right]\right|,$$

where f is a bounded function with Lipschitz constant 1. By the coupling argument in the proof of Proposition SA.1, it is possible to construct, for each $\pi\in\Pi$, a random variable $\tilde{\mathbf{z}}_\pi\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Omega}_\pi/N)$ (the covariance matrix $\boldsymbol{\Omega}_\pi$ depends on π, indicated by the subscript), such that

$$\mathbb{E}\left[\left|\sqrt{N}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi})-\sqrt{N}\tilde{\mathbf{z}}_\pi\right|\right]\le \text{Const.}\sqrt{\frac{1}{\min_{S\in\mathcal{X}}N_S}}.$$

Then we can bound the aforementioned quantity by

$$\begin{aligned}
&\left|\mathbb{E}_f\left[\sqrt{N}\max\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}}\oslash\boldsymbol{\sigma}_{\pi,\succ}\right)_+\right]-\mathbb{E}^\star_f\left[\sqrt{N}\max\left(\mathbf{R}_\succ(\boldsymbol{\Omega}_\pi^{1/2}\mathbf{w}^\star+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\right)_+\right]\right|\\
&\le\left|\mathbb{E}_f\left[\sqrt{N}\max\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}}\oslash\boldsymbol{\sigma}_{\pi,\succ}\right)_+\right]-\mathbb{E}_f\left[\sqrt{N}\max\left(\mathbf{R}_\succ(\tilde{\mathbf{z}}_\pi+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\right)_+\right]\right|\\
&\qquad+\left|\mathbb{E}_f\left[\sqrt{N}\max\left(\mathbf{R}_\succ(\tilde{\mathbf{z}}_\pi+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\right)_+\right]-\mathbb{E}^\star_f\left[\sqrt{N}\max\left(\mathbf{R}_\succ(\boldsymbol{\Omega}_\pi^{1/2}\mathbf{w}^\star+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\right)_+\right]\right|\\
&=\left|\mathbb{E}_f\left[\sqrt{N}\max\left(\mathbf{R}_\succ\hat{\boldsymbol{\pi}}\oslash\boldsymbol{\sigma}_{\pi,\succ}\right)_+\right]-\mathbb{E}_f\left[\sqrt{N}\max\left(\mathbf{R}_\succ(\tilde{\mathbf{z}}_\pi+\boldsymbol{\pi})\oslash\boldsymbol{\sigma}_{\pi,\succ}\right)_+\right]\right|,
\end{aligned}$$

where the second line comes from triangle inequality, and the equality uses the fact that $\tilde{\mathbf{z}}_\pi$ and $\boldsymbol{\Omega}_\pi^{1/2}\mathbf{w}^\star$ have the same distribution. Hence we have the bound:

$$\text{previous display}\precsim\sqrt{\frac{1}{\min_{S\in\mathcal{X}}N_S}},$$

with the estimate in the above display being uniformly valid over $\pi\in\Pi$. Hence the proof reduces to bound the probability (choose some $\varepsilon>0$ small enough):

$$\begin{aligned}
&\sup_{\pi\in\Pi}\mathbb{P}\left[\min_{S\in\mathcal{X}}N_S<\varepsilon N\right]=\sup_{\pi\in\Pi}\mathbb{P}\left[\frac{N_S}{N}<\varepsilon,\ \exists S\right]\le\sup_{\pi\in\Pi}\sum_{S\in\mathcal{X}}\mathbb{P}\left[\frac{N_S}{N}<\varepsilon\right]\\
&=\sup_{\pi\in\Pi}\sum_{S\in\mathcal{X}}\mathbb{P}\left[\frac{N_S-\mathbb{P}[Y_i=S]N}{N}<\varepsilon-\mathbb{P}[Y_i=S]\right]\\
&\le\sup_{\pi\in\Pi}\sum_{S\in\mathcal{X}}\exp\left(-(\varepsilon-\mathbb{P}[Y_i=S])^2\frac{N}{2}\right)\to 0.
\end{aligned}$$

For the inequality in the first line, we use union bound. The last line uses Hoeffding's inequality, which is valid for any ε smaller than $\inf_{\pi\in\Pi}\min_{S\in\mathcal{X}}\mathbb{P}[Y_i=S]\geq \underline{p}$ (see Assumption 2 in the main paper). Hence we demonstrated that

$$\sqrt{\frac{1}{\min_{S\in\mathcal{X}} N_S}} \asymp_{\mathbb{P}} \sqrt{\frac{1}{N}}, \qquad \text{uniformly in } \Pi.$$

■

Now we demonstrate how Theorem 4 in the main paper follows from the previous proposition.

Recall that in constructing $\mathscr{T}^\star(\succ)$, we use $\kappa_N^{-1}(\mathbf{R}_\succ\hat{\boldsymbol{\pi}})_-$ to replace the unknown $\mathbf{R}_\succ\boldsymbol{\pi}$. This unknown quantity is bounded above by 0 uniformly in Π, since this class only contains choice rules that are compatible with $\succ$. At the same time, $\kappa_N^{-1}(\mathbf{R}_\succ\hat{\boldsymbol{\pi}})_-$ converges to 0 in probability uniformly for the class Π. Therefore asymptotically $\mathscr{T}^\star(\succ)$ stochastically dominates $\mathscr{T}(\succ)$ uniformly in Π, which proves Theorem 4 in the main paper.

SA.4.2 Computation

The number of rows in the constraint matrix is

$$\#\mathrm{row}(\mathbf{R}_\succ) = \sum_{S\in\mathcal{X}}\sum_{a,b\in S}\mathbb{1}(b\prec a) = \sum_{S\in\mathcal{X},\ |S|\geq 2}\binom{|S|}{2} = \sum_{k=2}^{K}\binom{K}{k}\binom{k}{2},$$

where $K=|X|$ is the number of alternatives in the grand set X. Not surprisingly, the number of constraints increases very fast with the size of the grand set. However, we note that as long as the matrix $\mathbf{R}_\succ$ has been constructed for one preference $\succ$, constraint matrices for other preference orderings can be obtained by column permutations of $\mathbf{R}_\succ$. As a result, even though the construction of $\mathbf{R}_\succ$ might be computationally demanding with many alternatives available (i.e., when K is large), this step (Algorithm 1 in the main paper) only needs to be implemented once. This is particular useful and saves computation if there are multiple hypotheses to be tested.

We showcase how the proposed inference procedure is computationally attractive, especially when inference is conducted for many preferences. Table SA.1 records the computing time needed using our companion `R` package `ramchoice`. Note that the time reflects not only the time needed to construct the constraint matrices, but also 2,000 simulations from a multivariate normal distribution to obtain critical values.

Our general-purpose implementation executes very fast even when 720 different preference orderings are tested, and given that there are 240 constraints for testing one single preference. The full execution takes less than 7 minutes. Moreover, since the major computing time comes from constructing the constraint matrix $\mathbf{R}_\succ$ when testing many preferences, which involves looping over all choice problems and alternatives, Table SA.1. Computing Time ($K = 6$ and $K! = 720$)

Num. of Preferences	Time (seconds)	Num. of Preferences	Time (seconds)
1	1.117	50	21.185
5	2.668	100	40.422
10	4.584	400	195.907
20	8.948	720	407.177

There are $K = 6$ alternatives, leading to potentially $K! = 720$ preference orderings. All choice problems are observable in the data. For each preference, there are 240 inequality constraints in the matrix $\mathbf{R}_{\succ}$. The sample size is $N = 12,600$, and $M = 2,000$ simulations are used to construct critical value. System: MacOS 10.13.1, R 3.4.1.

it is possible to further speed up our general-purpose implementation by employing low-level programming languages such as `C++`, which are faster for simple for-loop structures.

SA.4.3 Connection to J-tests

The testing methods we proposed are connected to a class of inference strategies based on projection residuals, sometimes also known as J-tests. To describe this alternative inference approach, recall that any preference induces a surjective map from attention rules to choice rules (Definition 3 in the main paper), denoted by $\tilde{\mathbf{C}}_{\succ}$. Then, by our definition of partially identified preferences, $\succ \in \Theta_\pi$ if and only if $\boldsymbol{\pi}$ belongs to $\{\tilde{\mathbf{C}}_{\succ}\boldsymbol{\mu} :\ \mathbf{R}\boldsymbol{\mu} \leq 0\}$, where $\mathbf{R}$ represents the monotonicity assumption imposed on attention rules. Therefore, inference can be based on

$$\succ \in \Theta_\pi \quad \text{if and only if} \quad \inf_{\boldsymbol{\mu}:\ \mathbf{R}\boldsymbol{\mu}\leq 0} (\boldsymbol{\pi} - \tilde{\mathbf{C}}_{\succ}\boldsymbol{\mu})'\mathbf{W}(\boldsymbol{\pi} - \tilde{\mathbf{C}}_{\succ}\boldsymbol{\mu}) = 0,$$

where $\mathbf{W}$ is some positive definite weighting matrix. Hence, a preference compatible with π is equivalent to a zero residual from projecting $\boldsymbol{\pi}$ to the corresponding set $\{\tilde{\mathbf{C}}_{\succ}\boldsymbol{\mu} :\ \mathbf{R}\boldsymbol{\mu} \leq 0\}$. For example, this strategy is used by Kitamura and Stoye (2018) in random utility models.

To see the connection of J-tests to moment inequality testing, observe that if the mapping defined by $\tilde{\mathbf{C}}_{\succ}$ is invertible, then the above reduces to

$$\succ \in \Theta_\pi \quad \text{if and only if} \quad \mathbf{R}\tilde{\mathbf{C}}_{\succ}^{-1}\boldsymbol{\pi} \leq 0.$$

Such reduction may not be feasible analytically, or it may be numerically prohibitive. With a careful inspection of our problem, we showed that it is without loss of generality to focus on triangular attention rules, which can be uniquely constructed from preferences and choice rules. That is, we showed how the inversion $\tilde{\mathbf{C}}_{\succ}^{-1}$ is constructed, which is denoted by $\mathbf{C}_{\succ}$ in Theorem 3 in the main paper and its proof. Moreover, we provided an algorithm which constructs the constraint matrix directly, avoiding the detour of forming it as the product of two matrices.

Compared to directly testing inequality constraints as we propose, employing a J-test inference procedure has two potential drawbacks. First, the J-test statistic is essentially the (weighted) Euclidean norm of the projection residual, which may suffer from low power if only a few inequalities are violated. Second, constructing the projection residual requires numerical minimization, which can be computationally costly especially when the dimension of $\boldsymbol{\mu}$ is nontrivial. This is apparent from Table SA.1: testing a single preference takes about one second and testing for all 720 preference orderings takes about 7 minutes with our procedure, while employing the J-test can easily take a prohibitive amount of time because of the costly numerical optimization step over a possibly high-dimensional parameter space and the fact that this numerical optimization has to be done multiple times to construct a critical value. For example, in the same setting of Table SA.1, employing the J-test with $2,000$ bootstraps takes about 90 minutes for just one single preference when employing the quadratic programming procedure `quadprog` in `Matlab` (R2016b).

SA.4.4 Implementation with Limited Data

In some cases not all choice problems $S \in \mathcal{X}$ may be observed for two reasons. First, some choice problems are ruled out *a priori* in the population due to, for instance, capacity or institutional constraints. Second, certain choice problems may not be available in a given finite sample due to sampling variability. Even if all choice problems are observed in the data, some may have only a few appearances, and usually are dropped from empirical analysis to avoid dealing with small (effective) sample sizes. We descibe how our econometric methods (and assumptions) can also be adapted to situations of limited data.

Recall that $\mathcal{S} \subset \mathcal{X}$ denotes the collection of all observable choice problems. From an econometric perspective, this can also be seen as the collection of realized choice problems in the data. Assumption 2 in the main paper now takes the following form.

Assumption SA.1 (DGP with Limited Data)**.** The data is a random sample of choice problems Y_i and corresponding choices y_i, $\{(y_i, Y_i) : \ y_i \in Y_i, \ 1 \le i \le N\}$, generated by the underlying choice rule $\mathbb{P}[y_i = a|Y_i = S] = \pi(a|S)$, and that $\mathbb{P}[Y_i = S] \ge \underline{p} > 0$ for all $S \in \mathcal{S}$.

The most critical concern is on Assumption 1 in the main paper, which directly affects how the constraint matrix $\mathbf{R}_{\succ}$ is constructed. We state a seemingly "stronger" version of that assumption.

Assumption SA.2 (Monotonic Attention with Limited Data)**.** For any $A \subset S - T$, $\mu(T|S) \le \mu(T|S - A)$.

When $\mathcal{S} = \mathcal{X}$, this assumption is equivalent to Assumption 1 in the main paper. To see this fact, pick an arbitrary $A \subset S - T$ and let $a_1, a_2, \cdots$ be an enumeration of elements of A. Then, $\mu(T|S) - \mu(T|S - A)$ can be written as $\mu(T|S) - \mu(T|S - a_1) + \sum_{j \geq 1}[\mu(T|S - a_1 - \cdots - a_j) - \mu(T|S - a_1 - \cdots - a_j - a_{j+1})]$, where by Assumption 1 in the main paper each summand is nonpositive, hence Assumption 1 in the main paper implies Assumption SA.2. The other direction is obvious: Assumption SA.2 implies Assumption 1 in the main paper.

If $\mathcal{S}$ does not contain all possible choice problems, however, Assumption SA.2 provides a proper notion of monotonicity. To see the extra identification and statistical power implied by Assumption SA.2, two examples are provided after we give a proper definition of compatible preferences and the algorithm of constructing the constraint matrix in this context.

With limited data, unfortunately, there are two ways to generalize the previous definition, which are not equivalent. Recall that π_{obs} denotes the observed choice rule, defined only on $\mathcal{S}$, while we reserve the notation π to a choice rule that is defined on $\mathcal{X}$. Following is the first version.

Definition SA.1 (Compatible Preferences with Limited Data, I)**.** Let π_{obs} be the underlying choice rule/data generating process. A preference $\succ$ is compatible with π_{obs} on $\mathcal{S}$, denoted by $\succ \in \Theta_{\pi_{\text{obs}}}$, if $(\pi_{\text{obs}}, \succ)$ is a RAM on $\mathcal{S}$.

The other version is the following:

Definition SA.2 (Compatible Preferences with Limited Data, II)**.** Let π_{obs} be the underlying choice rule/data generating process. A preference $\succ$ is compatible with π_{obs} on $\mathcal{X}$, denoted by $\succ \in \Xi_{\pi_{\text{obs}}}$, if π_{obs} has an extension π to $\mathcal{X}$ such that $(\pi, \succ)$ is a RAM on $\mathcal{X}$.

When $\mathcal{S} = \mathcal{X}$, the two definitions agree, and also agree with the definition given in the main paper. If $\mathcal{S}$ is a proper subset of $\mathcal{X}$, however, $\Theta_{\pi_{\text{obs}}}$ can be larger than $\Xi_{\pi_{\text{obs}}}$. Depending on the goal of analysis, both can be of interest, while $\Xi_{\pi_{\text{obs}}}$ is much more difficult to characterize empirically.

The following algorithm generalizes the one given in the main text to the limited data scenario. The constraint matrix provided in this algorithm can be used to characterize $\Theta_{\pi_{\text{obs}}}$.

Consider now two examples with limited data.

Example SA.3 (Limited Data and Identification)**.** Assume $\mathcal{S} = \{\{a, b, c, d\}, \{a, b\}\}$, and we are interested in the null hypothesis that $c \succ b \succ d \succ a$. Assumption 1 in the main paper does not impose any constraint, since it only requires comparing choice problems that differ by one element. On the other hand, SA.2 implies the constraint that $\pi(a|\{a, b, c, d\}) \leq \pi(a|\{a, b\})$, since violating this constraint is equivalent to $a \succ c$ or $a \succ d$, which is incompatible with our hypothesis. ⌟

Algorithm 1' Construction of $\mathbf{R}_{\succ}$ with Limited Data.

```
Require: Set a preference $\succ$.
  $\mathbf{R}_{\succ} \leftarrow$ empty matrix
  for $S$ in $\mathcal{S}$ do
    for $T$ in $\mathcal{S}$ and $T \subset S$ do
      for $a \prec S - T$ do
        $\mathbf{R}_{\succ} \leftarrow$ add row corresponding to $\pi_{\text{obs}}(a|S) - \pi_{\text{obs}}(a|T) \leq 0$.
      end for
    end for
  end for
```

Example SA.4 (Limited Data and Statistical Power)**.** Consider $\mathcal{S} = \{\{a,b,c,d\},\{a,b,c\},\{a,b\}\}$. This is one example of limited data, but due to the special structure of $\mathcal{S}$, constraints implied by the two assumptions are equivalent *in population*. Consider the preference $a \succ c \succ d \succ b \succ$, then Assumption 1 in the main paper gives two constraints, (i) $\pi(b|\{a,b,c,d\}) \leq \pi(b|\{a,b,c\})$ and (ii) $\pi(b|\{a,b,c\}) \leq \pi(b|\{a,b\})$. Assumption SA.2, however, will give the *extra* condition (iii) $\pi(b|\{a,b,c,d\}) \leq \pi(b|\{a,b\})$. In population, this extra constraint is redundant, since it is not possible to violate (iii) without violating at least one of (i) and (ii).

When applied to finite sample, however, the extra condition (iii) is no longer redundant, and may help to improve statistical power. To see this, assume (i) is violated by $\delta > 0$, a small margin, so is (ii). Then it is hard to reject any of them with finite sample. On the other hand, (iii) combines (i) and (ii), hence is violated by 2δ, which is much easier to detect in a finite sample. $\lrcorner$

Test statistics and critical values are constructed in a similar way based on $\mathbf{R}_{\succ}\hat{\boldsymbol{\pi}}_{\text{obs}}$, hence not repeated here. We provide an analogous result of Theorem 4 in the main paper.

Theorem SA.2 (Validity of Critical Values with Limited Data I)**.** Assume Assumption SA.1 holds. Let Π_{obs} be a class of choice rules restricted to $\mathcal{S}$, and $\succ$ a preference, such that: (i) for each $\pi_{\text{obs}} \in \Pi_{\text{obs}}$, $\succ \in \Theta_{\pi_{\text{obs}}}$; and (ii) $\inf_{\pi_{\text{obs}}\in\Pi_{\text{obs}}} \min(\boldsymbol{\sigma}_{\pi_{\text{obs}},\succ}) > 0$. Then

$$\limsup_{N\to\infty} \sup_{\pi_{\text{obs}}\in\Pi_{\text{obs}}} \mathbb{P}\left[\mathscr{T}(\succ) > c_\alpha(\succ)\right] \leq \alpha.$$

One natural question is whether it is possible to make any claim on $\Xi_{\pi_{\text{obs}}}$ with limited data. Indeed, the three tests remain valid when applied to $\Xi_{\pi_{\text{obs}}}$ (i.e. controls size uniformly), which can be easily seen from the fact that $\Xi_{\pi_{\text{obs}}} \subset \Theta_{\pi_{\text{obs}}}$. We provide the following theorem. Also, for a choice rule π defined on $\mathcal{X}$, π_{obs} denotes its restriction to $\mathcal{S}$.

Theorem SA.3 (Validity of Critical Values with Limited Data II)**.** Assume Assumption SA.1 holds. Let Π be a class of choice rules on $\mathcal{X}$, and $\succ$ a preference, such that: (i) for each $\pi \in \Pi$, $\succ \in \Xi_{\pi_{\text{obs}}}$; and (ii)

$\inf_{\pi_{\mathrm{obs}}\in\Pi_{\mathrm{obs}}}\min(\boldsymbol{\sigma}_{\pi_{\mathrm{obs}},\succ}) > 0$. Then

$$\limsup_{N\to\infty}\sup_{\pi\in\Pi}\mathbb{P}\left[\mathscr{T}(\succ) > c_\alpha(\succ)\right] \leq \alpha.$$

In Theorem SA.3, we only need positive variance for moment conditions corresponding to the observed components of the choice rule. The reason is simple: In constructing the tests and critical values, we never use the unobserved part $\boldsymbol{\pi} - \boldsymbol{\pi}_{\mathrm{obs}}$.

What is lost from Theorem SA.2 to Theorem SA.3? With limited data, there may exist $\succ\in\Theta_{\pi_{\mathrm{obs}}} - \Xi_{\pi_{\mathrm{obs}}}$ which is not rejected by the test statistic $\mathscr{T}(\succ)$ asymptotically.

SA.4.5 Other Critical Values

There are many proposals for constructing critical values in the literature of testing moment inequalities (Canay and Shaikh, 2017; Ho and Rosen, 2017; Molinari, 2019, and references therein). Here we discuss some of them for completeness. Throughout, let $\mathbf{z}^\star$ denote a random vector independent of the original data and $\mathbf{z}^\star \sim \mathcal{N}(\mathbf{0},\ \hat{\boldsymbol{\Omega}}/N)$. To save notation, define $(x)_+$ as the positive parts of x, and $(x)_-$ as the negative parts multiplied by -1 (truncation above at zero).

Plug-in Method

The first method simply plugs-in an estimate of $\mathbf{R}_\succ\boldsymbol{\pi}$ subject to the non-positivity constraint. Define

$$\mathscr{T}^\star_{\mathtt{PI}}(\succ) = \sqrt{N}\cdot\max\Big(\big(\mathbf{R}_\succ\mathbf{z}^\star + (\mathbf{R}_\succ\hat{\boldsymbol{\pi}})_-\big)\oslash\hat{\boldsymbol{\sigma}}_\succ\Big)_+,$$

then the critical value with level α with the plug-in method is

$$c_{\alpha,\mathtt{PI}}(\succ) = \inf\Big\{t:\ \mathbb{P}^\star\left[\mathscr{T}^\star_{\mathtt{PI}}(\succ)\leq t\right]\geq 1-\alpha\Big\},$$

where $\mathbb{P}^\star$ denotes probability operator conditional on the data, and in practice, it is replaced by the simulated average $M^{-1}\sum_{m=1}^{M}\mathbb{1}(\cdot)$, where recall that $\mathscr{T}^\star_{\mathtt{PI}}(\succ)$ is simulated M times.

We note the critical values obtained here are *not* uniformly valid in the sense of Theorem 4 in the main paper.

Least Favorable Model

The critical values are nondecreasing in the centering, hence a conservative method is to consider the least favorable model, $\mathbf{R}_{\succ}\boldsymbol{\pi} = 0$, which assumes all the moment inequalities are binding. That is,

$$\mathscr{T}^{\star}_{\texttt{LF}}(\succ) = \sqrt{N} \cdot \max\Big((\mathbf{R}_{\succ}\mathbf{z}^{\star}) \oslash \hat{\boldsymbol{\sigma}}_{\succ} \Big)_{+},$$

and

$$c_{\alpha,\texttt{LF}}(\succ) = \inf\Big\{ t :\ \mathbb{P}^{\star}\left[\mathscr{T}^{\star}_{\texttt{LF}}(\succ) \le t \right] \ge 1 - \alpha \Big\}.$$

Undoubtedly, using such critical value may severely decrease the power of the test, especially when one or two moment restrictions are violated and the rest are far from binding. Next we introduce two methods seeking to improve power property of the test. The first one relies on the idea of moment selection, and the third one replaces the unknown moment conditions with upper bounds.

Two-step Moment Selection

To illustrate this method, we first represent $\mathbf{R}_{\succ}\hat{\boldsymbol{\pi}}$ by its individual components, as $\{\mathbf{r}'_{\succ,\ell}\hat{\boldsymbol{\pi}} :\ 1 \le \ell \le L\}$, where $\mathbf{r}'_{\succ,\ell}$ is the ℓ-th row of $\mathbf{R}_{\succ}$, so that there are in total L moment inequalities. The first step is to conduct moment selection. Let $0 < \beta < \alpha/3$, and the following set of indices of "active moment restrictions":

$$\mathcal{L} = \left\{ \ell :\ \sqrt{N} \cdot \frac{\mathbf{r}'_{\succ,\ell}\hat{\boldsymbol{\pi}}}{\hat{\sigma}_{\succ,\ell}} \ge -2 \cdot c_{\beta,\texttt{LF}}(\succ) \right\},$$

then

$$\mathscr{T}^{\star}_{\texttt{MS}}(\succ) = \sqrt{N} \cdot \max_{\ell \in \mathcal{L}} \left(\frac{\mathbf{r}'_{\succ,\ell}\mathbf{z}^{\star}}{\hat{\sigma}_{\succ,\ell}} \right)_{+},$$

and the critical value is computed as

$$c_{\alpha,\texttt{MS}}(\succ) = \inf\Big\{ t :\ \mathbb{P}^{\star}\left[\mathscr{T}^{\star}_{\texttt{MS}}(\succ) \le t \right] \ge 1 - \alpha + 2\beta \Big\}.$$

Constructing the coupling statistic $\mathscr{T}^{\star}(\mathcal{P}_{\forall})$ requires more delicate work. Note that simply taking maximum of individual $\mathscr{T}^{\star}_{\texttt{MS}}(\succ)$ does not work, mainly due to the two-step nature of the construction. More precisely, the first step moment selection controls error probability to be β for each individual preference, but not jointly, and the reason is that we used $c_{\beta,\texttt{LF}}(\succ)$ for moment selection, which is too small jointly for a collection of preferences.

For a collection of preferences, the correct moment selection is the following:

$$\mathcal{L}_{\succ,\mathcal{P}} = \left\{\ell:\ \sqrt{N}\cdot\frac{\mathbf{r}'_{\succ,\ell}\hat{\boldsymbol{\pi}}}{\hat{\sigma}_{\succ,\ell}} \geq -2\cdot c_{\beta,\texttt{LF}}(\mathcal{P}_\forall)\right\}, \qquad \text{for each } \succ\in\mathcal{P}.$$

Now we use a much larger critical value for moment selection: $c_{\beta,\texttt{LF}}(\mathcal{P}_\forall)$ instead of $c_{\beta,\texttt{LF}}(\succ)$. Then the coupling statistic is

$$\mathscr{T}^\star_{\texttt{MS}}(\mathcal{P}_\forall) = \sqrt{N}\cdot\max_{\succ\in\mathcal{P}}\max_{\ell\in\mathcal{L}_{\succ,\mathcal{P}}}\left(\frac{\mathbf{r}'_{\succ,\ell}\mathbf{z}^\star}{\hat{\sigma}_{\succ,\ell}}\right)_+,$$

then the critical value is computed accordingly as the $1-\alpha$ quantiles:

$$c_{\alpha,\texttt{MS}}(\mathcal{P}_\forall) = \inf\left\{t:\ \mathbb{P}^\star\left[\mathscr{T}^\star_{\texttt{MS}}(\mathcal{P}_\forall)\leq t\right]\geq 1-\alpha+2\beta\right\}.$$

See, for example, Chernozhukov, Chetverikov, and Kato (2019) and references therein.

Two-step Moment Upper Bounding

This method uses a first step to construct a confidence region for $\mathbf{R}_\succ\boldsymbol{\pi}$, and the upper bound of such region is used as a conservative estimate for $\mathbf{R}_\succ\boldsymbol{\pi}$. Let $0<\beta<\alpha$, and

$$\mathscr{T}^\star_{\texttt{UB}}(\succ) = \sqrt{N}\cdot\max\left((\mathbf{R}_\succ\mathbf{z}^\star + (\mathbf{R}_\succ\hat{\boldsymbol{\pi}} + c_{\beta,\texttt{LF}}(\succ)\hat{\boldsymbol{\sigma}}_\succ/\sqrt{N})_-)\oslash\hat{\boldsymbol{\sigma}}_\succ\right)_+,$$

then

$$c_{\alpha,\texttt{UB}}(\succ) = \inf\left\{t:\ \mathbb{P}^\star\left[\mathscr{T}^\star_{\texttt{UB}}(\succ)\leq t\right]\geq 1-\alpha+\beta\right\}.$$

Note that in the first step, we use the critical values from the least favorable method to construct upper bounds on the moment inequalities $\mathbf{R}_\succ\hat{\boldsymbol{\pi}}+c_{\beta,\texttt{LF}}(\succ)\hat{\boldsymbol{\sigma}}/\sqrt{N}$, which is guaranteed to have coverage $1-\beta$. Then the significance level in the second step is adjusted to account for errors incurred in the first step. See, e.g., Romano, Shaikh, and Wolf (2014) for further details. In particular, they recommend to use $\beta/\alpha=0.1$.

Same as the moment selection method, the upper bound (or confidence region for $\mathbf{R}_\succ\boldsymbol{\pi}$) constructed above only controls error probability for individual preference, but not jointly for a collection of preferences. Hence we need to make further adjustments. The solution is almost the same: replace the critical value used for constructing upper bounds by one that controls error probability jointly for the collection $\mathcal{P}$:

$$\mathscr{T}^\star_{\texttt{UB}}(\mathcal{P}_\forall) = \sqrt{N}\cdot\max_{\succ\in\mathcal{P}}\max\left((\mathbf{R}_\succ\mathbf{z}^\star + (\mathbf{R}_\succ\hat{\boldsymbol{\pi}} + c_{\beta,\texttt{LF}}(\mathcal{P}_\forall)\hat{\boldsymbol{\sigma}}_\succ/\sqrt{N})_-)\oslash\hat{\boldsymbol{\sigma}}_\succ\right)_+,$$

then the critical value is computed accordingly as the $1-\alpha$ quantiles:

$$c_{\alpha,\mathtt{UB}}(\mathcal{P}_\forall) = \inf\Big\{t:\ \mathbb{P}^\star\left[\mathscr{T}^\star_{\mathtt{UB}}(\mathcal{P}_\forall) \le t\right] \ge 1-\alpha+\beta\Big\}.$$

SA.5 Additional Simulation Results

We first recall the simulation setup. We consider a class of logit attention rules indexed by ς:

$$\mu_\varsigma(T|S) = \frac{w_{T,\varsigma}}{\sum_{T'\subset S} w_{T',\varsigma}}, \qquad w_{T,\varsigma} = |T|^\varsigma,$$

where $|T|$ is the cardinality of T. In this Supplemental Appendix, we give simulation evidence for $\varsigma \in \{0,1\}$.

As in the main paper, we list five hypotheses (preference orderings), and whether they are compatible with our RAM model and specific values of ϕ.

$\varsigma = 0$	ϕ										
	1	.95	.90	.85	.80	.75	.70	.65	.60	.55	.50
$\mathsf{H}_{0,1}: a_1 \succ a_2 \succ a_3 \succ a_4 \succ a_5$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$\mathsf{H}_{0,2}: a_2 \succ a_3 \succ a_4 \succ a_5 \succ a_1$	✓	✓	✓	✓	✓	✓	✓	×	×	×	×
$\mathsf{H}_{0,3}: a_3 \succ a_4 \succ a_5 \succ a_2 \succ a_1$	✓	✓	✓	✓	✓	✓	✓	×	×	×	×
$\mathsf{H}_{0,4}: a_4 \succ a_5 \succ a_3 \succ a_2 \succ a_1$	✓	✓	✓	✓	✓	✓	✓	×	×	×	×
$\mathsf{H}_{0,5}: a_5 \succ a_4 \succ a_3 \succ a_2 \succ a_1$	✓	✓	✓	✓	✓	✓	✓	×	×	×	×

$\varsigma = 1$	ϕ										
	1	.95	.90	.85	.80	.75	.70	.65	.60	.55	.50
$\mathsf{H}_{0,1}: a_1 \succ a_2 \succ a_3 \succ a_4 \succ a_5$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$\mathsf{H}_{0,2}: a_2 \succ a_3 \succ a_4 \succ a_5 \succ a_1$	✓	✓	✓	✓	✓	✓	×	×	×	×	×
$\mathsf{H}_{0,3}: a_3 \succ a_4 \succ a_5 \succ a_2 \succ a_1$	✓	✓	✓	✓	✓	✓	×	×	×	×	×
$\mathsf{H}_{0,4}: a_4 \succ a_5 \succ a_3 \succ a_2 \succ a_1$	×	×	×	×	×	×	×	×	×	×	×
$\mathsf{H}_{0,5}: a_5 \succ a_4 \succ a_3 \succ a_2 \succ a_1$	×	×	×	×	×	×	×	×	×	×	×

With $\varsigma = 0$ being the data generating process, we have the identified set to be the set of all preferences, and for $\varsigma = 1$, it is $\{\succ: a_3 \succ a_4 \succ a_5\}$. Simulation results are collected in Figure SA.1 and SA.2.

References

Abaluck, Jason and Abi Adams. 2017. "What Do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses." NBER Working Paper No. 23566.

Aguiar, Victor H. 2015. "Stochastic Choice and Attention Capacities: Inferring Preferences from Psychological Biases." SSRN Working Paper No. 2607602.

———. 2017. "Random Categorization and Bounded Rationality." *Economics Letters* 159:46–52.

Aguiar, Victor H, María José Boccardi, and Mark Dean. 2016. "Satisficing and Stochastic Choice." *Journal of Economic Theory* 166:445–482.

Ahumada, Alonso and Levent Ulku. 2018. "Luce Rule with Limited Consideration." *Mathematical Social Sciences* 93:52–56.

Barberà, Salvador and Birgit Grodal. 2011. "Preference for Flexibility and the Opportunities of Choice." *Journal of Mathematical Economics* 47 (3):272–278.

Barseghyan, Levon, Maura Coughlin, Francesca Molinari, and Joshua C. Teitelbaum. 2018. "Heterogeneous Consideration Sets and Preferences." Work in Progress, Cornell University.

Brady, Richard L and John Rehbeck. 2016. "Menu-Dependent Stochastic Feasibility." *Econometrica* 84 (3):1203–1223.

Canay, Ivan A. and Azeem M. Shaikh. 2017. "Practical and Theoretical Advances for Inference in Partially Identified Models." In *Advances in Economics and Econometrics: Volume 2: Eleventh World Congress*, edited by B. Honore, A. Pakes, M. Piazzesi, and L. Samuelson. Cambridge: Cambridge University Press, 271–306.

Chen, Louis H.Y., Larry Goldstein, and Qi-Man Shao. 2010. *Normal Approximation by Stein's Method*. New York: Springer.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. 2019. "Inference on Causal and Structural Parameters using Many Moment Inequalities." *Review of Economic Studies*, forthcoming.

Dardanoni, Valentino, Paola Manzini, Marco Mariotti, and Christopher J Tyson. 2018. "Inferring Cognitive Heterogeneity from Aggregate Choices." Working Paper Series No. 1018, Department of Economics, University of Sussex.

de Clippel, Geoffroy and Kareen Rozen. 2014. "Bounded Rationality and Limited Datasets." SSRN Working Paper No. 2018463.

Demirkan, Yusufcan and Mert Kimya. 2018. "Hazard Rate, Stochastic Choice and Consideration Sets." Working Paper.

Echenique, Federico and Kota Saito. 2019. “General Luce Model.” *Economic Theory*, forthcoming.

Echenique, Federico, Kota Saito, and Gerelt Tserenjigmid. 2018. “The Perception-Adjusted Luce Model.” *Mathematical Social Sciences* 93:67–76.

Fudenberg, Drew, Ryota Iijima, and Tomasz Strzalecki. 2015. “Stochastic Choice and Revealed Perturbed Utility.” *Econometrica* 83 (6):2371–2409.

Gul, Faruk, Paulo Natenzon, and Wolfgang Pesendorfer. 2014. “Random Choice as Behavioral Optimization.” *Econometrica* 82 (5):1873–1912.

Ho, Kate and Adam M. Rosen. 2017. “Partial Identification in Applied Research: Benefits and Challenges.” In *Advances in Economics and Econometrics: Volume 2: Eleventh World Congress*, edited by B. Honore, A. Pakes, M. Piazzesi, and L. Samuelson. Cambridge: Cambridge University Press, 307–359.

Horan, Sean. 2018a. “Random Consideration and Choice: A Case Study of “Default” Options.” Working Paper, Université de Montréal and CIREQ.

———. 2018b. “Threshold Luce Rules.” Working Paper, Université de Montréal and CIREQ.

Kitamura, Y. and J. Stoye. 2018. “Nonparametric Analysis of Random Utility Models.” *Econometrica* 86 (6):1883–1909.

Manzini, Paola and Marco Mariotti. 2014. “Stochastic Choice and Consideration Sets.” *Econometrica* 82 (3):1153–1176.

Masatlioglu, Yusufcan and Elchin Suleymanov. 2017. “Decision Making within a Product Network.” Working Paper.

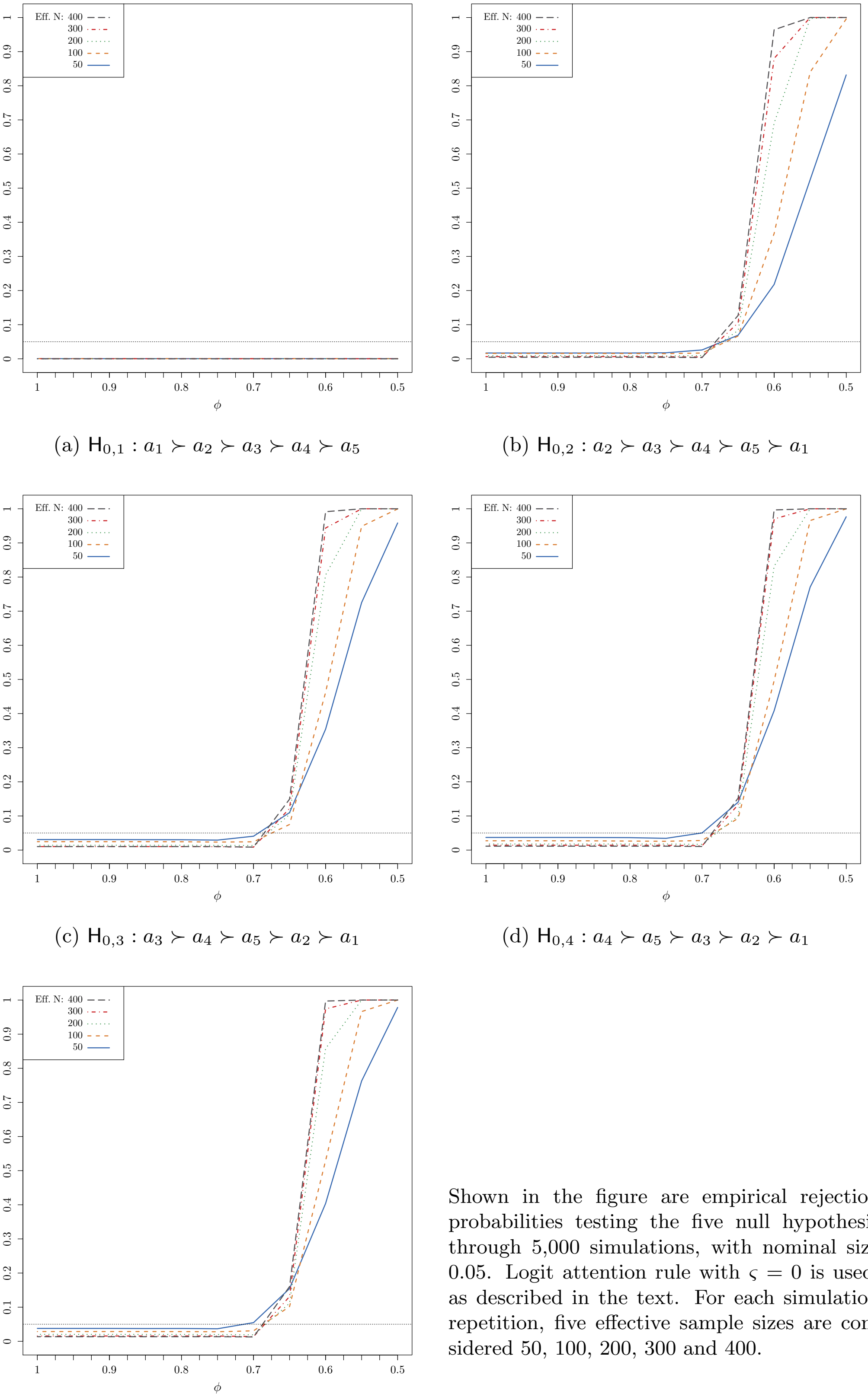
Molinari, Francesca. 2019. “Econometrics with Partial Identification.” In *Handbook of Econometrics VII*, forthcoming.

Romano, Joseph P, Azeem M Shaikh, and Michael Wolf. 2014. “A Practical Two-Step Method for Testing Moment Inequalities.” *Econometrica* 82 (5):1979–2002.

Salant, Yuval and Ariel Rubinstein. 2008. “(A, f): Choice with Frames.” *Review of Economic Studies* 75 (4):1287–1296.

Simon, Herbert A. 1955. “A Behavioral Model of Rational Choice.” *Quarterly Journal of Economics* 69 (1):99–118.

Suleymanov, Elchin. 2018. “Stochastic Attention and Search.” Working Paper.



(a) $\mathsf{H}_{0,1} : a_1 \succ a_2 \succ a_3 \succ a_4 \succ a_5$

(b) $\mathsf{H}_{0,2} : a_2 \succ a_3 \succ a_4 \succ a_5 \succ a_1$

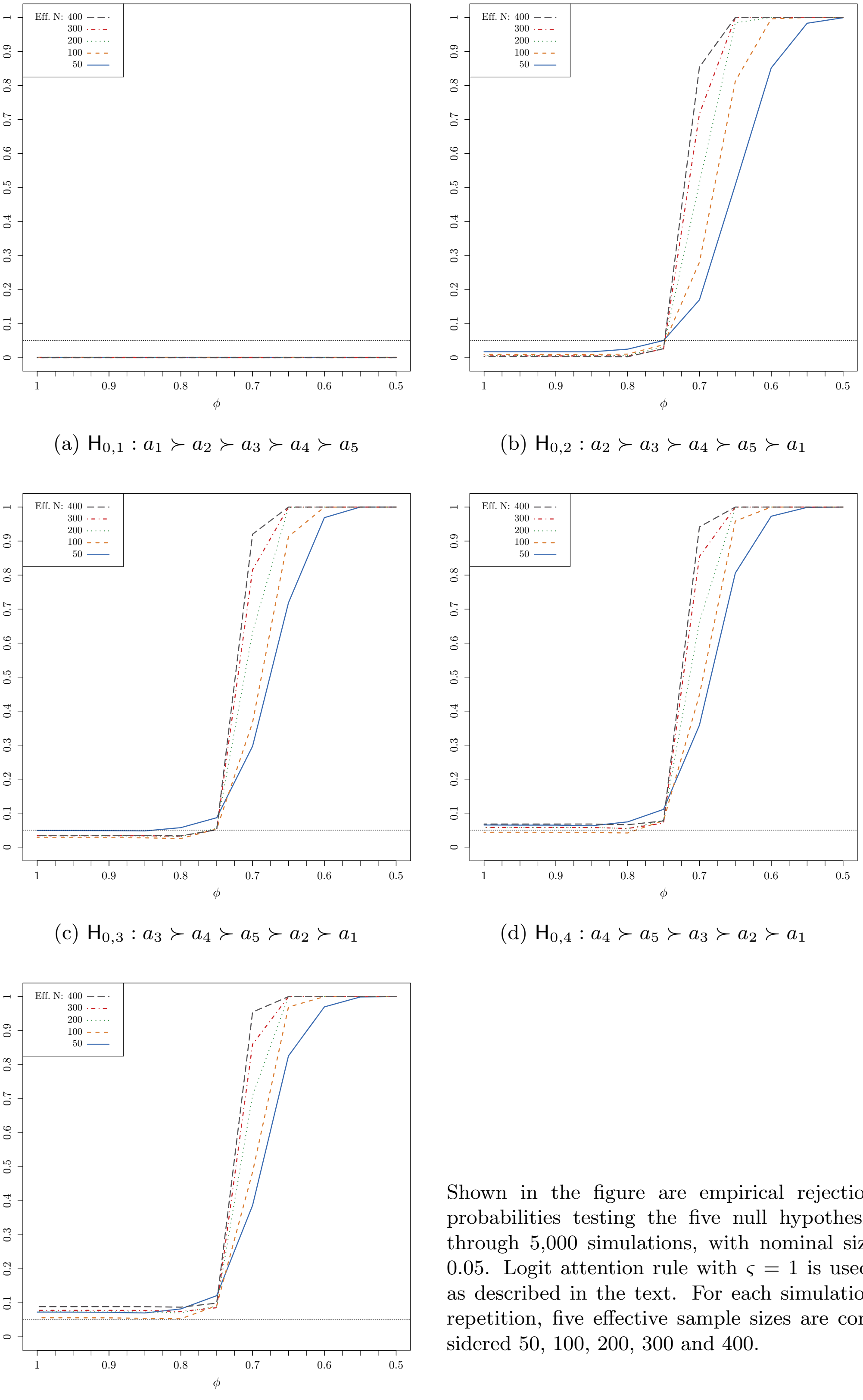
(c) $\mathsf{H}_{0,3} : a_3 \succ a_4 \succ a_5 \succ a_2 \succ a_1$

(d) $\mathsf{H}_{0,4} : a_4 \succ a_5 \succ a_3 \succ a_2 \succ a_1$

(e) $\mathsf{H}_{0,5} : a_5 \succ a_4 \succ a_3 \succ a_2 \succ a_1$

Shown in the figure are empirical rejection probabilities testing the five null hypothesis through 5,000 simulations, with nominal size 0.05. Logit attention rule with $\varsigma = 0$ is used, as described in the text. For each simulation repetition, five effective sample sizes are considered 50, 100, 200, 300 and 400.

Figure SA.1. Empirical Rejection Probabilities ($\varsigma = 0$)



(a) $\mathsf{H}_{0,1} : a_1 \succ a_2 \succ a_3 \succ a_4 \succ a_5$

(b) $\mathsf{H}_{0,2} : a_2 \succ a_3 \succ a_4 \succ a_5 \succ a_1$

(c) $\mathsf{H}_{0,3} : a_3 \succ a_4 \succ a_5 \succ a_2 \succ a_1$

(d) $\mathsf{H}_{0,4} : a_4 \succ a_5 \succ a_3 \succ a_2 \succ a_1$

Shown in the figure are empirical rejection probabilities testing the five null hypothesis through 5,000 simulations, with nominal size 0.05. Logit attention rule with $\varsigma = 1$ is used, as described in the text. For each simulation repetition, five effective sample sizes are considered 50, 100, 200, 300 and 400.

(e) $\mathsf{H}_{0,5} : a_5 \succ a_4 \succ a_3 \succ a_2 \succ a_1$

Figure SA.2. Empirical Rejection Probabilities ($\varsigma = 1$)