Lu, Mia; Mononen, Lasse; Netzer, Nick

Working Paper
The swaps index for consumer choice

Working Paper, No. 418

Provided in Cooperation with:
Department of Economics, University of Zurich

Suggested Citation: Lu, Mia; Mononen, Lasse; Netzer, Nick (2026) : The swaps index for consumer choice, Working Paper, No. 418, University of Zurich, Department of Economics, Zurich, https://doi.org/10.5167/uzh-220550

This Version is available at:
https://hdl.handle.net/10419/343174

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 418

The Swaps Index for Consumer Choice

Mia Lu, Lasse Mononen and Nick Netzer

First version: August 2022
This version: August 2026

The Swaps Index for Consumer Choice

Mia Lu, Lasse Mononen and Nick Netzer*

This version: August 2026
First version: August 2022

Abstract

We extend the swaps index of rationality, introduced by Apesteguia and Ballester (2015) for a finite set of alternatives, to the standard infinite consumer choice setting in which other rationality indices are typically defined. We show that the swaps approach recovers the decision-maker's true preference from choice data for a class of behavioral models that satisfy a monotonicity condition. We also derive methods to compute the swaps index and swaps preferences for real-world data sets.

Keywords: measures of rationality, revealed preference, behavioral welfare economics
JEL Classification: D01, D11, D60, D90

*Mia Lu: Department of Economics, University of Zurich. Lasse Mononen: Center for Mathematical Economics, Bielefeld University, lasse.mononen@uni-bielefeld.de. Nick Netzer: Department of Economics, University of Zurich, nick.netzer@econ.uzh.ch. We are grateful for very helpful comments by Tim Netzer, Jakub Steiner, and seminar participants at the 2022 Asian School in Economic Theory, Swiss Theory Day 2022, and the University of Zurich. An earlier version of this paper was part of Mia Lu's PhD thesis. This research project and related results were made possible with the support of the Alfred P. Sloan Foundation and the NOMIS Foundation.

1 Introduction

Measuring the degree of rationality of choices has a long tradition in economics. The early literature on revealed preference theory (Samuelson, 1938; Houthakker, 1950; Arrow, 1959) provided conditions under which choices can be classified as perfectly rational. Subsequent authors have acknowledged that human decision-makers often deviate from rationality, if only by mistake or lapse of attention, and have proposed measures that quantify these deviations. The most prominent examples of rationality measures are Afriat's efficiency index (Afriat, 1973) and Varian's goodness-of-fit index (Varian, 1990).[1] These indices have been applied in various empirical studies and have facilitated important insights, for example on the development of rationality during childhood (Harbaugh et al., 2001), the correlation between rationality and wealth (Choi et al., 2014), and the causal impact of education on rationality (Kim et al., 2018).

More recently, Apesteguia and Ballester (2015, henceforth AB) have proposed the *swaps index* of rationality, which is based on counting the minimal number of choices that would have to be swapped to reconcile behavior with the maximization of a preference. This not only quantifies deviations from rationality but additionally delivers a revealed preference, as in the original revealed preference literature mentioned above. The *swaps preference* is the one closest to the observed behavior in terms of the number of required swaps. AB have shown that their approach recovers the decision-maker's true preference from choice data for a large class of boundedly rational behavioral models.[2] The swaps preference can thus be used as a welfare measure, connecting the literature on measuring rationality with the literature on behavioral welfare economics (e.g. Kőszegi and Rabin, 2007; Bernheim and Rangel, 2009; Rubinstein and Salant, 2012; Benkert and Netzer, 2018).

AB obtained their results in a setting with a finite number of alternatives. We extend the swaps approach to the standard consumer choice setting with infinite commodity spaces. This is the canonical setting in which most other common rationality measures are defined. Applications include, among others, consumer demand from competitive budget sets and choice under uncertainty where investors purchase Arrow-Debreu securities. We show that the swaps approach still recovers the true preference in this setting under a suitably defined monotonicity condition. We discuss the plausibility of the monotonicity condition by showing that it is satisfied for various behavioral models of interest but violated for others. We then

[1] Other rationality measures have been proposed by, among others, Houtman and Maks (1985), Echenique et al. (2011), Dean and Martin (2016), and Echenique et al. (2023).

[2] Halevy et al. (2018) and Lanier and Quah (2026) show that other rationality indices can also be given an interpretation as a minimal distance between observed behavior and preferences, for suitably defined distance measures, but AB (p. 1292f.) argue that these indices do not generally recover true preferences from behavioral data.

provide methods for computing the swaps index and swaps preferences for real-world data sets.

The swaps approach rests on the *swaps distance* between observed behavior and any candidate preference. In the finite case, if alternative x was chosen from budget set B and the candidate preference is $\succeq$, then the number of swaps required to reconcile this behavior with the maximization of $\succeq$ is simply the number of alternatives in B which are strictly preferred by $\succeq$ over x. This is the number of alternatives a decision-maker with preference $\succeq$ must have "overlooked" when choosing x. The swaps distance between a stochastic choice function and a preference is the probability-weighted average of these numbers across observations. A swaps preference is a strict preference minimizing the swaps distance to the stochastic choice function. The swaps index is the minimum swaps distance.

For the infinite case, AB (Online Appendix D.3) proposed replacing the number of preferred alternatives in a budget set by their Lebesgue measure, but did not derive general results in that setting. We implement their proposal. Our approach is more general than the proposal by AB in several dimensions. First, we allow for arbitrary stochastic choice from a budget set, not only finite randomization. Special cases include deterministic choice or finite randomization, as well as continuous distributions like uniform trembles over the entire budget set or more complex singular distributions like continuous trembles on the budget frontier. Second, we allow for arbitrary collections of choice situations, not only finite collections of competitive budget sets. This allows us to embed conventional, deterministic Walrasian demand functions for all prices as a special case. We are also flexible enough to model nonlinear prices, endowment points, and even frames which affect behavior without affecting the budget set. Third, we impose only mild regularity conditions on preferences and obtain various results without requiring continuous utility functions or convexity of preferences.

Our first set of results centers around the recoverability of preferences from behavioral data. The central property that guarantees recoverability is monotonicity. For the finite case, AB define a stochastic choice function to be monotone with respect to preference $\succeq$ if $x \succeq y$ implies that the probability of choosing x is larger than the probability of choosing y, on average across all budget sets in which both x and y are available. Several behavioral models generate choice data which satisfy monotonicity with respect to the decision-maker's true preference, for example a simple trembles model in which uniform choice mistakes happen with small probability. AB show that $\succeq$-monotonicity implies that $\succeq$ is a swaps preference. For our definition of monotonicity in the consumer choice setting, we first apply a Lebesgue decomposition of the choice probability measure into its absolutely continuous and singular components. Then we require that $x \succeq y$ implies that the density of the absolutely continuous component is weakly larger for x than for y, on average across all budget sets in which both x and y are available. We furthermore require that all bundles in the support of the singular component are optimal with respect to $\succeq$. Based on this definition, we show that $\succeq$-monotonicity implies that $\succeq$ is a swaps preference. The swaps approach can therefore be seen as a preference estimator for the entire range of behavioral models that generate monotone data.

To illustrate this range, we study examples of boundedly rational behavioral models. As in the finite case, many of them generate data satisfying monotonicity. The first example is a uniform trembles model in which the decision-maker chooses optimally from a budget set with probability $1 - \epsilon$ and randomizes uniformly over the budget set with probability ϵ. The decision-maker's true preference is a swaps preference when choice mistakes take this form. Furthermore, and again in analogy to AB, the swaps preference is unique when the data are rich. To establish this result, we exploit a connection to the problem of unique recoverability of preferences from conventional, deterministic Walrasian demand functions (Mas-Colell, 1977). A second example for which monotonicity obtains is a continuous version of the multinomial logit model. A third example is a model of framing, in which choice is influenced by defaults or presentation effects more generally. As long as this affects only the absolutely continuous component of choice and averages out across frames, monotonicity is satisfied. There are, however, also behavioral models which violate monotonicity. For example, even a noisy decision-maker may understand that consumption bundles in the interior of the budget set are dominated and may randomize over the budget frontier instead of the entire budget set. As we will show, monotonicity with respect to the true preference breaks down in that case. Swaps preferences will be systematically distorted by the need to explain the choices on the budget frontier. Other models for which monotonicity fails include a limited attention version of the logit model (Matejka and McKay, 2015) and a random parameter model in the context of choice under uncertainty (Apesteguia and Ballester, 2018).

We conclude the first part of the paper with a discussion of approximate recoverability. If a data set is close to $\succeq$-monotone, are the swaps preferences close to $\succeq$? We derive a bound on the Kendall tau distance between swaps preferences and any arbitrary preference, which reflects how well the arbitrary preference explains the data. This bound allows us to show that approximate recoverability holds under a strict notion of monotonicity but not necessarily otherwise.

Our second set of results concerns the computation of the swaps index and swaps preferences for data sets typically encountered in empirical applications. These data sets consist of a finite number of deterministic choices from different budget sets, a special case of our general setting. We restrict attention to monotone preferences here, as such data can otherwise always be explained by preferences that rank the finitely many chosen bundles above all others, generating a Lebesgue swaps distance of zero. We show that the swaps index can be calculated by solving a simple minimization problem. Each ranking of the chosen bundles is associated with a unique minimal swaps value, obtained by adding across all observations the Lebesgue measure of how much the higher ranked bundles and their principal upper sets cut into each observation's budget set. The swaps index is the minimum of these values across all rankings. Our result is more general than an analogous recent result by Lanier and Quah (2026), because it holds for arbitrary budget sets and for preferences that are not necessarily continuous. Furthermore, we provide an alternative characterization as a binary minimization problem where revealed preferences are deleted at a cost until the revealed preference graph becomes acyclic. This follows the approach of Mononen (2023), who studied similar techniques for the conventional rationality measures discussed above and showed that the approach is computationally feasible.

Deriving swaps preferences is complicated by the fact that they are neither guaranteed to be unique nor to exist in the infinite consumer choice setting. Lanier and Quah (2026) have shown that a swaps preference does not exist when GARP is violated in the standard case of linear budgets and when continuity of preferences is required. Perhaps surprisingly, this finding turns out to be driven by the requirement of continuity. We show that a swaps preference always exists within the class of monotone but not necessarily continuous preferences. We then derive *robust swaps preferences* among the chosen bundles, which are the pairwise rankings that all preferences with a swaps distance sufficiently close to the swaps index must have in common. The idea of robust preferences was proposed by Lanier and Quah (2026), who study them for the distance functions underlying the indices of Afriat, Varian, and Houtman-Maks, but not the swaps distance. If a swaps preference does not exist (for example when continuity is required and GARP is violated), these robust swaps preferences are a reasonable alternative welfare measure. If swaps preferences do exist (for example when only monotonicity is required), then robust swaps preferences are the rankings that all swaps preferences agree on, capturing what is unambiguously revealed by the data, similar in spirit to Bernheim and Rangel (2009). We show that our minimization problems that characterize the swaps index also characterize the robust swaps preferences. For example, when all rankings that solve the above-described minimization problem rank two bundles in the same way, this ranking is a robust swaps preference.

Finally, we study a short empirical application using the experimental data of Castillo and Freer (2023). Their experiment involves nonlinear budget sets. Several older measures of rationality are not defined in this setting, which makes it an ideal application for the swaps index. Castillo and Freer (2023) use the Houtman-Maks index as a distance measure to test the hypothesis of quasilinear preferences. We compare the Houtman-Maks index with the swaps index, without putting additional requirements on preferences. We show that the swaps index differentiates levels of rationality more finely than the Houtman-Maks index and thus allows for more fine-grained hypothesis tests. Furthermore, robust swaps preferences are more complete than robust preferences based on the Houtman-Maks distance. In addition to showcasing the applicability and computational feasibility of our results, these findings support the swaps index as an informative approach that generates rich preferences useful for welfare analysis.

The paper is organized as follows. Section 2 studies the recoverability of preferences from behavioral data. Section 3 provides methods to compute the swaps index and swaps preferences. Section 4 contains our empirical application. Section 5 concludes. Proofs are relegated to the Appendix.

2 Recoverability

In this section, we study whether the swaps approach can correctly recover a decision-maker's true preferences. We assume that the decision-maker sometimes chooses differently from her most preferred option, for example due to random stochastic errors or lack of attention, but we observe only the decision-maker's choices and not her true preferences.

2.1 Formal Framework

The choice set is $X = \mathbb{R}^L_+$, where L denotes the number of commodities. The commodities could be different consumption goods, or they could refer to states of the world. Budget sets $B_\delta \subseteq X$ are parameterized by $\delta \in \Delta \subseteq \mathbb{R}^K$, where K denotes the number of parameters. The budget sets are assumed to be bounded and Lebesgue measurable for each $\delta \in \Delta$. In the standard setting, $\delta \in \Delta \subseteq \mathbb{R}^L_{++}$ is a price vector and $B_\delta = \{x \in X \mid \delta \cdot x \leq 1\}$ is the competitive budget set.[3] Our approach is more general and δ could, for example, also encode kinks in the budget line generated by endowment points in an insurance setting, other nonlinear budget lines as generated by nonlinear screening mechanisms, and even frames which affect choice without affecting the budget set.

The observed behavior of the decision-maker is summarized by her choice data, here defined as a collection D of probability measures, $D = (\nu, (\mu_\delta)_{\delta \in \Delta})$.[4] First, ν is a full support

[3] The implicit assumption then is that the decision-maker's demand satisfies homogeneity of degree zero. Hence we do not need to distinguish between demand on $\{x \mid \delta \cdot x \leq w\}$ and $\{x \mid (\delta/w) \cdot x \leq 1\}$ and can normalize wealth to 1.

[4] Throughout, all probability measures are defined on the Borel algebra, and expressions like "absolutely

probability measure on the sample space Δ, representing how often the decision-maker faces a choice situation indexed by δ. We think of this as exogenous to the decision-maker. For example, the measure ν could describe the frequency of budget problems as determined by an experimenter. Since Δ can be finite or infinite, this encompasses the case of choice from finitely many different budget sets but also allows observation of a demand function for all prices, as in the conventional Walrasian approach to consumer demand. Second, the choice behavior of the decision-maker for any choice situation $\delta \in \Delta$ is captured by a probability measure μ_δ on the sample space B_δ. That is, for any measurable subset of alternatives $A \subseteq B_\delta$, $\mu_\delta(A)$ is the probability that the decision-maker chooses a consumption bundle in the subset A conditional on facing the choice situation δ. Again, we consider general probability measures here. This includes the cases of deterministic choice, discrete randomization, and stochastic choice described by a density function.

By Lebesgue's decomposition theorem, we can decompose every μ_δ into an absolutely continuous measure μ_δ^c and a singular measure μ_δ^s such that

$$\mu_\delta = \mu_\delta^c + \mu_\delta^s.$$

Moreover, this decomposition is always unique. Due to the absolute continuity, we can then write the measure μ_δ^c in terms of a density f_δ such that

$$\mu_\delta^c(A) = \int_A f_\delta(x)dx$$

for every measurable subset $A \subseteq B_\delta$.[5] The singular measure μ_δ^s captures the possibility of deterministic choice, discrete randomization, or stochastic choice on other subsets of B_δ that have Lebesgue measure zero.

Unlike in the finite case of AB, indifferences cannot be assumed away in the consumer choice setting with infinite commodity spaces. Instead, we only impose two weak regularity conditions on the decision-maker's preferences. First, we assume that her complete and transitive preference relation $\succeq$ on X can be represented by some (Borel) measurable utility function $u : X \to \mathbb{R}$. A sufficient condition for this would be continuity of $\succeq$. Second, we assume that $\succeq$ has thin level sets, i.e., the sets $L(z) = \{x \in X \mid u(x) = z\}$ have Lebesgue measure zero for all $z \in \mathbb{R}$. One consequence is that we can use $\lambda(\{y \in A \mid y \succeq x\})$ and $\lambda(\{y \in A \mid y \succ x\})$ interchangeably, where λ denotes the Lebesgue measure and $\succ$ the

continuous" or "singular" refer to the Lebesgue measure.

[5]Note that, while μ_δ is a probability measure, μ_δ^c and μ_δ^s are generally not probability measures, unless μ_δ itself is already absolutely continuous or singular. Hence f_δ is not a probability density function, because $\int_{B_\delta} f_\delta(x)dx = \mu_\delta^c(B_\delta)$ can be smaller than one. This does not have to concern us in the following.

asymmetric part of $\succeq$. We denote the set of preference relations that have thin level sets and for which measurable utility representations exist by $\mathscr{R}$.

We now define the swaps index in this setting. It is a natural extension of the swaps index for finite sets and was already proposed by AB (Online Appendix D.3). We generalize it here by allowing for infinite collections of observed choice situations, unconstrained stochastic choice in each of those situations, and our larger class of preferences. The *swaps distance* between a preference $\succeq$ and choice data D is defined as

$$W(\succeq, D) = \int_{\Delta} \int_{B_\delta} \lambda(\{y \in B_\delta \mid y \succ x\}) d\mu_\delta(x) d\nu(\delta).$$

It measures the inconsistency of the given choice data D with the given preference $\succeq$. For any chosen bundle x in a budget set B_δ, inconsistency with $\succeq$ is measured by the Lebesgue volume of the upper contour set of x in B_δ. That is, the inconsistency is measured by how many available alternative bundles were strictly preferred to x but not chosen and would thus have to be "swapped" with x to make choice consistent with maximization of $\succeq$. The overall inconsistency of D with $\succeq$ is then computed by first taking the weighted average of this volume over all bundles in the budget set, and then taking the weighted average over all choice situations, where the respective weights are the observed frequencies in the data.[6,7]

The *swaps index* for the given choice data D is defined as

$$I(D) = \inf_{\succeq \in \mathscr{R}} W(\succeq, D).$$

If a preference attains the infimum, we refer to it as a *swaps preference*. A swaps preference is a preference with the smallest possible inconsistency with the choice data. Note that swaps preferences are neither guaranteed to be unique nor to exist in the infinite consumer choice setting. We will return to this issue repeatedly in the following sections.

2.2 Main Result

Our goal is to establish a condition which ensures that the decision-maker's true preference is a swaps preference. To this end, we introduce a monotonicity property, which we adapt

[6] The fact that u is a measurable function implies that all upper contour sets are Lebesgue measurable and their weighted average volume within each budget set is well-defined. In our definition of the swaps distance, we implicitly restrict attention to choice data sets for which the weighted average over all choice situations is also well-defined. In particular, we assume that $f_\delta(x)$ is jointly measurable in x and δ.

[7] A similar measure was proposed by Heufer (2008). The main difference is that Heufer (2008) does not use the Lebesgue measure of the upper contour set of a chosen bundle with respect to a candidate preference, but instead uses the Lebesgue measure of the bundles which are revealed to be preferred over the chosen bundle by the decision-maker's other choices and monotonicity.

from the finite setting of AB. Informally speaking, monotonicity with respect to a given preference, typically interpreted as the decision-maker's true preference, requires that more preferred bundles are chosen more often.

Definition 1. *The choice data $D = (\nu, (\mu_\delta)_{\delta\in\Delta})$ are $\succeq_*$-monotone for a preference $\succeq_* \in \mathscr{R}$ if they satisfy the following two conditions:*

(*i*) *The absolutely continuous component has expected densities that are monotone in $\succeq_*$,*

$$\forall x, y \in X : x \succeq_* y \Rightarrow \int_{\{\delta\in\Delta | \{x,y\}\subseteq B_\delta\}} f_\delta(x) d\nu(\delta) \geq \int_{\{\delta\in\Delta | \{x,y\}\subseteq B_\delta\}} f_\delta(y) d\nu(\delta).$$

(*ii*) *The support of the singular component is optimal with respect to $\succeq_*$,*

$$\forall \delta \in \Delta, \forall x \in supp(\mu_\delta^s) : x \succeq_* y \;\; \forall y \in B_\delta.$$

Condition (*i*) requires that a more preferred bundle x has a higher expected density in the absolutely continuous component of the data than a less preferred bundle y, and in this sense is observed more frequently on average across all decision problems in which the two bundles are available. A strong sufficient condition is $f_\delta(x) \geq f_\delta(y)$ for all $\delta \in \Delta$, so that the inequality holds pointwise in the integrals. Bundles in the support of the singular component (e.g., mass points) are observed even "more frequently," so condition (*ii*) requires them to be preference maximizing.[8]

We will discuss the plausibility of the monotonicity property in the next subsection. First, we state the main result of Section 2.

Proposition 1. *If D satisfies $\succeq_*$-monotonicity, then $\succeq_*$ is a swaps preference.*

The intuition for Proposition 1 is that $\succeq_*$-monotonicity ensures that the choice data are informative about $\succeq_*$. This preference then minimizes the weighted inconsistency with the data and is a swaps preference.

As a first interpretation, the swaps approach can thus be seen as a preference estimator that is robust across the range of behavioral models that generate data satisfying monotonicity with respect to the decision-maker's true preference. We will follow this interpretation in Subsection 2.3, where we start from behavioral models and establish monotonicity, so that

[8]We could relax Definition 1 in two ways but refrain from doing so for ease of notation. First, we could replace condition (*ii*) by the weaker requirement that $\lambda(\{y \in B_\delta \mid y \succ_* x\}) = 0$ for all $x \in supp(\mu_\delta^s)$, thus allowing for non-optimal bundles in the support of μ_δ^s provided that the Lebesgue measure of better bundles is zero. This would give some additional degrees of freedom when $\succeq_*$ is not a continuous preference. Second, we could relax the requirements in Definition 1 on null sets of the product measures $\nu \times \mu_\delta$.

the swaps approach is justified because it recovers the correct preference from the noisy data. A second interpretation is to start from given data and assume that a preference has been found for which those data are monotone. Proposition 1 then implies that this preference minimizes the swaps distance to the data, which could be seen as a desirable welfare property in itself, without recourse to a specific behavioral model.

Monotonicity implies the existence of a swaps preference. However, the swaps preference is not necessarily unique, so the true preference cannot necessarily be recovered uniquely from the data. We will return to this issue in the next subsection, where we illustrate that uniqueness can be achieved with rich data sets. Furthermore, in Section 3 we will study the notion of a robust swaps preference, which is useful when multiple swaps preferences exist.

2.3 Applications

To illustrate the plausibility of the monotonicity condition and to elaborate on the unique recoverability of the preference, we now discuss several examples of behavioral models for which monotonicity is satisfied or violated.

2.3.1 Uniform Trembles

A first example satisfying $\succeq_*$-monotonicity is a uniform trembles model, which AB have studied for the finite case. In choice situation $\delta \in \Delta$, the decision-maker chooses optimally in budget set B_δ according to her true preference $\succeq_*$ with probability $1-\epsilon > 0$. For example, Figure 1 illustrates a case where the unique optimal bundle is x^*. With probability $\epsilon \geq 0$, the decision-maker trembles uniformly over all bundles in B_δ, as indicated by the shaded red area in Figure 1.

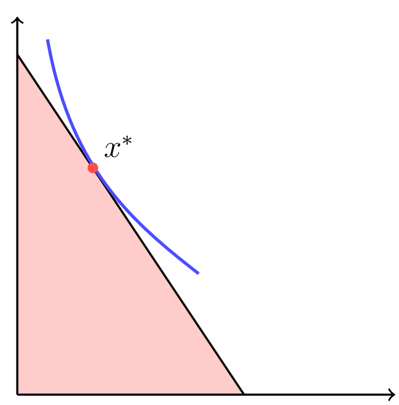

Figure 1: Uniform Trembles.

Formally, the choice data $D = (\nu, (\mu_\delta)_{\delta\in\Delta})$ are given by an exogenous distribution ν over choice situations and, for each $\delta \in \Delta$, a probability measure μ_δ that can be decomposed into an absolutely continuous component μ_δ^c with constant density $f_\delta(x) = \epsilon/\lambda(B_\delta)$ and a singular component μ_δ^s that satisfies $\mu_\delta^s(A) = 1-\epsilon$ if $x_\delta^* \in A$ and $\mu_\delta^s(A) = 0$ if $x_\delta^* \notin A$, where x_δ^* denotes the optimal bundle in B_δ according to $\succeq_*$ (assumed to be unique for simplicity). It is easy to see that these data satisfy $\succeq_*$-monotonicity. The density condition (i) in the definition holds with equality for each $\delta \in \Delta$ separately, and condition (ii) is satisfied since the non-tremble choice is optimal according to $\succeq_*$. It follows from Proposition 1 that the true preference $\succeq_*$ is a swaps preference.

The swaps preference will generally not be unique. For example, if only a single budget set is observed, as in Figure 1, then any preference for which bundle x^* is optimal is a swaps preference (the same would be true in a finite setting). Observing choice from additional budget sets puts more constraints on the set of swaps preferences, because all mass points have to be explained as optimal to achieve the same minimum swaps index as the true preference $\succeq_*$. Our next result gives conditions under which the swaps preference becomes unique when the data set becomes complete.

We make use of an existing recoverability result for conventional, deterministic Walrasian demand functions by Mas-Colell (1977). We restrict attention to preferences $\succeq$ that are continuous, monotone, and strictly convex. As defined in Mas-Colell (1977, p. 1411), such a preference is *lipschitzian* if, for every $0 < r < \infty$, the function $V_{\succeq,r} : K_r \to \mathcal{C}$ given by $V_{\succeq,r}(x) = \{y \in K_r \mid y \succeq x\}$ is Lipschitz continuous, where $K_r = \{y \in \mathbb{R}_+^L \mid 1/(1+r) \leq y_i \leq 1+r, \forall i = 1,\ldots,L\}$ and $\mathcal{C}$ is the set of nonempty, compact, and convex subsets of $\mathbb{R}_+^L$.

Proposition 2. *Let $\Delta = \mathbb{R}_{++}^L$ and $B_\delta = \{x \in X \mid \delta \cdot x \leq 1\}$. Let $D = (\nu, (\mu_\delta)_{\delta\in\Delta})$ be choice data generated by a uniform trembles model with a true preference $\succeq_*$ that is continuous, monotone, and strictly convex. If $\succeq_*$ is additionally lipschitzian, then it is the unique swaps preference in the class of continuous, monotone, and strictly convex preferences.*

To understand the logic of the result, it is instructive to consider the Walrasian demand function x^* generated by preference $\succeq_*$. Under the lipschitzian assumption, Mas-Colell (1977) has shown that any other preference $\succeq \neq \succeq_*$ generates a Walrasian demand function x that differs from x^* for at least one price vector δ_0. Continuity of Walrasian demand then implies that x^* and x must differ on an open set of prices. Furthermore, the set of bundles that $\succeq$ prefers strictly over x^* has strictly positive Lebesgue measure on that open set of prices. Since the singular part of the choice data generated by the uniform trembles model is described by x^*, its swaps distance to $\succeq_*$ is zero but strictly positive to $\succeq$. We also show in the proof that all preferences explain uniform randomization and hence the absolutely continuous part of the choice data equally well (Lemma 1 in Appendix A.2). It follows that $\succeq$ is not a swaps preference.

We can calculate the value of the swaps index for data generated by a uniform trembles model as follows (where the third equality follows from Lemma 1):

$$\begin{aligned}
I(D) &= \int_\Delta \int_{B_\delta} \lambda(\{y \in B_\delta \mid y \succ_* x\}) d\mu_\delta(x) d\nu(\delta) \\
&= \int_\Delta \frac{\epsilon}{\lambda(B_\delta)} \int_{B_\delta} \lambda(\{y \in B_\delta \mid y \succ_* x\}) dx d\nu(\delta) \\
&= \int_\Delta \frac{\epsilon}{\lambda(B_\delta)} \frac{\lambda(B_\delta)^2}{2} d\nu(\delta) \\
&= \epsilon \int_\Delta \frac{\lambda(B_\delta)}{2} d\nu(\delta).
\end{aligned}$$

The degree of irrationality is increasing linearly in the noise parameter ϵ. It also depends on the decision-maker's environment described by ν. If she is confronted with larger budget sets, then her uniform trembles over those sets imply a larger degree of irrationality.

2.3.2 Uniform Trembles on the Budget Frontier

Now consider a variant of the uniform trembles model. With probability $1 - \epsilon_1 - \epsilon_2 > 0$, the decision-maker chooses optimally in budget set B_δ according to $\succeq_*$. With probability $\epsilon_1 \geq 0$, she trembles uniformly over all bundles in B_δ. With probability $\epsilon_2 > 0$, she trembles uniformly over all bundles on the frontier of B_δ. For example, the decision-maker may find it easier to understand that bundles in the interior of the budget set are dominated than to determine the exact location of the optimal bundle on the budget frontier. Trembles on the frontier would then arise when the decision-maker is mildly distracted, while trembles on the entire budget set would arise only if distraction is strong.

Formally, the measure μ_δ is decomposed into the absolutely continuous component μ_δ^c that captures trembles on the entire budget set and the singular component μ_δ^s that captures trembles on the budget frontier and the optimal choice. This is illustrated in Figure 2, where the optimal bundle is x^* and trembles take place sometimes on the budget line and sometimes on the entire budget set.

Suppose for simplicity that we only observe one budget set B as in Figure 2. It follows immediately that the choice data are best explained by an indifference between all bundles on the budget line. With an indifference curve like the bold red line, all singular choices are perfectly explained. Since all preferences explain the uniform trembles on the entire budget set equally well, it follows that some perfect-substitutes preference is a swaps preference.

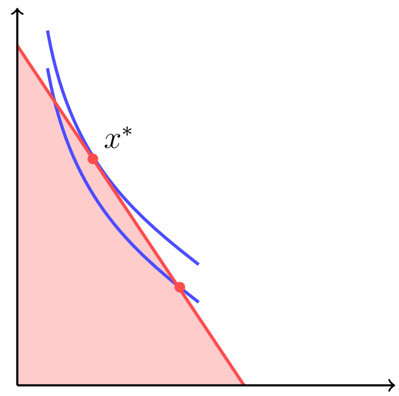

Figure 2: Uniform Trembles on the Budget Frontier.

In fact, the choice data are monotone with respect to that perfect-substitutes preference. Whenever the true preference is different, as depicted by the convex blue indifference curves in Figure 2, its swaps distance to the data is strictly larger, because it generates upper contour sets with strictly positive volume for all non-optimal choices on the budget line. Hence, the true preference is not a swaps preference and recoverability fails. The value of the swaps index is $I(D) = \epsilon_1 \lambda(B)/2$, missing the trembles that arise with probability ϵ_2.[9]

Indifferences cannot be ruled out in the consumer choice setting, in contrast to the finite setting. The bold red line in Figure 2 is also compatible with strictly monotone preferences, a restriction proposed by AB (Online Appendix D.2) to deal with indifferences. Note furthermore that the recoverability problem would not disappear when imposing the additional restriction of strictly convex preferences, so that an indifference among all bundles on the budget line is no longer possible. Preferences with indifference curves closer to the bold red line still perform better than the true preference. The problem is also not due to the assumption of observing only one budget set. Figuring out the swaps preference is more difficult when the environment is richer (see Section 3), but the previous arguments imply that preference elicitation will be systematically distorted by the need to explain the singular choices on the budget frontiers.

An interesting special case arises when $\epsilon_1 = 0$, so that trembles take place entirely on the budget frontier. We can now attempt to solve the problem by reducing the dimensionality of the commodity space. Consider again an example with two commodities, $x = (x_1, x_2)$, and a single budget set, as in Figure 2. Write the decision-maker's problem as a choice of $x_1 \in B = [0, x_1^{max}]$ and determine x_2 as a residual on the budget line. The data are

[9]The value $\epsilon_1 \lambda(B)/2$ follows from the same calculation as in Subsection 2.3.1, after observing that only the uniform randomization over the entire budget set contributes to the swaps distance for the optimal perfect-substitutes preferences.

then given by the choice of x_1^* with probability $1 - \epsilon_2$ (singular component) and a uniform distribution on $[0, x_1^{max}]$ with probability ϵ_2 (now the absolutely continuous component). We can look for a swaps preference on $B = [0, x_1^{max}]$. Note that $\mathscr{R}$ now excludes the complete indifference, which would not have thin level sets in the lower dimension. As before, all preferences in $\mathscr{R}$ explain the uniform randomization equally well. The same arguments as in Subsection 2.3.1 then imply that any preference for which x_1^* is optimal in B is a swaps preference, generating a swaps index of $I(D) = \epsilon_2 x_1^{max}/2$. In particular, the decision-maker's true preference on B induced by her true preference $\succeq_*$ on $\mathbb{R}^2_+$ satisfies that condition. Hence, reducing the dimensionality of the commodity space solves the problem when the observed choice is entirely singular.

2.3.3 Random Parameter Models

There are behavioral models for which recoverability fails even after a reduction of the dimensionality of the commodity space. Consider the following example of a random parameter model in the context of choice under uncertainty (e.g. Apesteguia and Ballester, 2018).

Let $x = (x_1, x_2) \in X = \mathbb{R}^2_+$ denote consumption in two states of the world. The decision-maker is endowed with a bundle $(w, w-d) \in X$, reflecting that her wealth is w if no accident occurs (state 1) but is reduced by $d > 0$ if an accident occurs (state 2). The probability of an accident is $0 < p < 1$. The decision-maker can buy insurance. Each unit of insurance costs q, to be paid in both states of the world, and pays 1 in case of an accident. We assume $q > p$ so that the insurance company earns profits. The resulting budget set B is illustrated in Figure 3. The dashed 45-degree line indicates perfect insurance bundles.

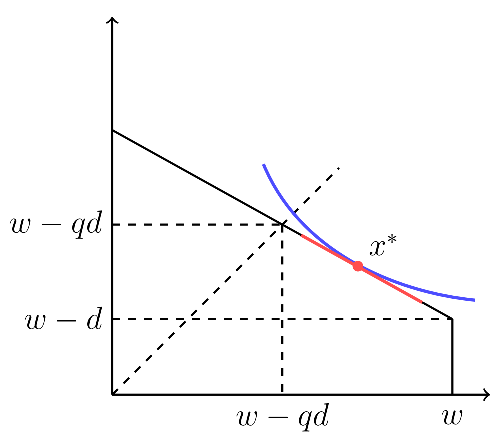

Figure 3: Random Parameter Model.

The decision-maker is an expected utility maximizer with CARA utility $u(x_i) = -e^{-\beta x_i}$ where $\beta > 0$. It is an easy exercise to show that the optimal bundle is

$$(x_1, x_2) = \left(w - qd + \frac{q\delta}{\beta}, w - qd - \frac{(1-q)\delta}{\beta}\right),$$

with

$$\delta = \ln\left(\frac{1-p}{p}\right) - \ln\left(\frac{1-q}{q}\right) > 0,$$

where we assume $\beta > \delta/d$ to rule out corner solutions at the endowment point.

Suppose choice is governed by a uniform distribution of β on $[\beta_* - \epsilon/2, \beta_* + \epsilon/2]$ for some $\epsilon > 0$, i.e., the preference parameter is uniformly distributed around β_*. This model generates singular stochastic choice on the budget line, as indicated by the bold red line segment around x^* in Figure 3. Hence, by our previous arguments, the swaps approach applied in $\mathbb{R}^2_+$ predicts an indifference among all chosen bundles, which has no plausible relation to the underlying random parameter model.

Reducing the dimensionality of the commodity space does not solve that problem. The decision-maker can be conceptualized as choosing x_1 from $B = [x_1^{min}, x_1^{max}]$, where

$$x_1^{min} = w - qd + \frac{q\delta}{\beta_* + \epsilon/2} \quad \text{and} \quad x_1^{max} = w - qd + \frac{q\delta}{\beta_* - \epsilon/2}.$$

Due to nonlinearity in the risk aversion coefficient, choice is not uniform on B. It is described by the density

$$f(x_1) = \frac{q\delta}{\epsilon\,(x_1 - w + qd)^2},$$

which is strictly decreasing in x_1. These data are monotone with respect to preferences that always favor smaller values of x_1, i.e., higher levels of insurance. The decision-maker's average risk aversion coefficient β_* induces a preference on B that is unimodal with an interior peak at $x_1^* = w - qd + q\delta/\beta_*$ and therefore does not satisfy that condition. More generally, within the parametric class of CARA utility functions, swaps preferences are the preferences with $\beta \geq \beta_* + \epsilon/2$. Thus, the lowest swaps coefficient of risk aversion is the upper bound of the interval of possible true coefficients. The swaps approach overestimates risk aversion systematically.

2.3.4 Logit Models

In a conventional multinomial logit model (or Luce model) for a finite set of alternatives, the probability of choosing an alternative x from a set B is given by

$$p(x, B) = \frac{e^{u(x)/\gamma}}{\sum_{y \in B} e^{u(y)/\gamma}}, \tag{1}$$

where $\gamma > 0$ is a noise parameter. As $\gamma \to 0$, choice concentrates on the utility maximizing alternatives in B. As $\gamma \to \infty$, choice converges to uniform randomization. The model can be justified as a random utility model with Gumbel error terms (see e.g. Anderson et al., 1992).[10]

We can translate this idea to our consumer choice setting by postulating that each μ_δ is absolutely continuous and the decision-maker chooses from B_δ in choice situation $\delta \in \Delta$ according to the density

$$f_\delta(x) = \frac{e^{u(x)/\gamma}}{\int_{B_\delta} e^{u(y)/\gamma} dy}.$$

It is immediate that the resulting data D are monotone with respect to the preference represented by the utility function u, which is hence a swaps preference.

There are modifications of the logit model for which monotonicity fails. Matejka and McKay (2015) study a rational inattention model in which the decision-maker has to learn about her true utility function and pays an entropy-based cost of information. Consider their setting with a finite set of alternatives and assume that the utility function u_i arises with prior probability q_i, where $i = 1, \ldots, m$. Matejka and McKay (2015) show that the rationally inattentive decision-maker chooses stochastically according to a modified multinomial logit formula. When the true utility function is u_i, her conditional choice probabilities are

$$p(x, B|u_i) = \frac{p(x, B) e^{u_i(x)/\gamma}}{\sum_{y \in B} p(y, B) e^{u_i(y)/\gamma}}, \tag{2}$$

where

$$p(y, B) = \sum_{i=1}^{m} q_i p(y, B|u_i)$$

[10] See Alós-Ferrer et al. (2021) for a critique of the distributional assumptions made in conventional random utility models like the multinomial logit and for a constructive way to dispense with these assumptions by using response time data.

are the unconditional choice probabilities. When all alternatives are ex ante homogeneous, this simultaneous system of equations yields the conventional logit formula (1) from above. Otherwise, conditional choice probabilities are distorted towards unconditional choice probabilities. We show in Appendix A.3 that the conditional choice probabilities are not always monotone with respect to the preference $\succeq_i$ represented by u_i. We may have $p(x, B|u_i) > p(y, B|u_i)$ even though $u_i(x) < u_i(y)$. We also show that the unconditional choice probabilities are not always monotone with respect to the prior expected preference. We can have $p(x, B) > p(y, B)$ even though $\sum_{i=1}^m q_i u_i(x) < \sum_{i=1}^m q_i u_i(y)$. It follows that there is not necessarily a plausible relation between swaps preferences and true preferences for data generated by the rational inattention model.

2.3.5 Framing

Consider a situation in which choice is affected by frames (Salant and Rubinstein, 2008). For example, one of the bundles in a budget set could be marked as the default, and different defaults could lead to different choice behaviors. A choice situation could then be described by $\delta = (p, d) \in \Delta$, where $p \in \mathbb{R}^L_{++}$ is the commodity prices and $d \in B_\delta = \{x \in X \mid p \cdot x \leq 1\}$ is the default bundle. More generally, we can always combine budget sets with different frames that describe the way in which the budget set is presented or that encode other relevant features of the environment such as the level of distraction for the decision-maker.

Formally, given an arbitrary set of choice situations Δ, we can define an equivalence relation $\simeq$ on Δ by letting $\delta \simeq \delta'$ if and only if $B_\delta = B_{\delta'}$. In words, $\delta \simeq \delta'$ means that the two choice situations δ and δ' involve the same budget set and hence can differ only in their framing. In the above example, we would have $\delta = (p, d) \simeq (p', d') = \delta'$ if and only if $p = p'$. Denote the equivalence class of δ by $[\delta]$.

If frames have no systematic directional effect on choices but only encode different levels of noise, then our previous arguments apply. For example, if we have a uniform trembles model as in Subsection 2.3.1 and different frames only affect the level of noise ϵ, because they generate different levels of distraction, then the data still satisfy $\succeq_*$-monotonicity for the true preference $\succeq_*$ and the true preference can be recovered from the data as before. In other applications, such as the one with defaults, frames have systematic directional effects on choices. Here, the fact that condition (i) in the definition of $\succeq_*$-monotonicity requires only a ranking of the expected densities becomes important. A natural sufficient condition for this ranking to remain valid despite directional distortions is that $x \succeq_* y$ implies

$$\int_{[\delta]} f_{\delta'}(x) d\nu(\delta') \geq \int_{[\delta]} f_{\delta'}(y) d\nu(\delta')$$

for each equivalence class $[\delta]$ with $\{x, y\} \subseteq B_\delta$. In words, as long as the directional effects arise only in the absolutely continuous part of the data and do not reverse the preference ranking on average, monotonicity and preference recoverability are preserved.

This result may be useful for the analysis of nudging (Thaler and Sunstein, 2008). In the spirit of the behavioral welfare approach by Benkert and Netzer (2018), one could check monotonicity of frame-dependent choice data and then evaluate the quality of choices under different frames using as a welfare criterion the respective preference for which monotonicity obtains, with the goal of ranking frames according to their induced decision quality.

2.4 Approximate Recoverability

After having explored behavioral models which do or do not generate monotone data, we now discuss an approximate version of the recoverability result. For example, we can stick to our interpretation that $\succeq_*$ is the decision-maker's true preference and that the behavioral model generates $\succeq_*$-monotone data. When those data are slightly perturbed, if only due to measurement error that does not necessarily obey preference monotonicity, are swaps preferences guaranteed to remain close to $\succeq_*$? Alternatively, and in line with our second interpretation of Proposition 1 above, if we find a candidate preference $\succeq_*$ for which the data set is almost monotone, will $\succeq_*$ be close to a swaps preference?

As a measure of distance between preferences, we use a weighted Kendall tau distance

$$d(\succeq_1, \succeq_2) = \int_{X \times X} \mathbb{1}(y \succ_1 x)\mathbb{1}(x \succ_2 y) w(x, y) d(x, y),$$

where the weighting function $w : X \times X \to \mathbb{R}_+$ is an exogenous, integrable function chosen to make the integral well-defined for all pairs of preferences. For instance, when w is a symmetric density, then $d(\succeq_1, \succeq_2)$ captures the probability of disagreement (divided by 2) between the preferences for randomly chosen bundles (and takes values between 0 and 1/2).

We now restrict attention to the case where the data D are absolutely continuous, i.e., $\mu_\delta = \mu_\delta^c$ for all $\delta \in \Delta$. For all $(x, y) \in X \times X$, define

$$m(x, y) = \int_{\{\delta \in \Delta | \{x,y\} \subseteq B_\delta\}} (f_\delta(x) - f_\delta(y))\, d\nu(\delta).$$

The previous $\succeq_*$-monotonicity becomes the requirement that $x \succeq_* y$ implies $m(x, y) \geq 0$. As discussed earlier, this means that the data provide systematic evidence for $\succeq_*$. The size of $m(x, y)$ reflects the strength of this evidence. Strength of evidence has not played a role so far but becomes important now.

Proposition 3. *Suppose D is absolutely continuous and $\succeq_S$ is a swaps preference. For any preference $\succeq_* \in \mathscr{R}$,*

$$d(\succeq_S, \succeq_*) \leq \inf_{\alpha>0} \int_{X\times X} \mathbb{1}(x \succ_* y) \max\left\{w(x,y) - \alpha m(x,y), 0\right\} d(x,y). \tag{3}$$

Our bound on the distance between the swaps preference $\succeq_S$ and $\succeq_*$ reflects the data's strength of evidence for or against preference $\succeq_*$. For any fixed $\alpha > 0$, the expression on the right-hand side of (3) is relatively small when $x \succ_* y$ implies that $m(x,y)$ is positive and large. It is relatively large when $x \succ_* y$ implies that $m(x,y)$ is small or even negative.

Consider the case where D is actually $\succeq_*$-monotone. Suppose first that monotonicity is strict, in the sense that $x \succ_* y$ implies $m(x,y) > 0$. This holds, for example, for an absolutely continuous logit version of a Walrasian demand function observed on all linear budget sets.[11] The bound in (3) is zero in that case.[12] A first implication is the novel insight that, for strictly $\succeq_*$-monotone data, $\succeq_*$ is the unique swaps preference up to almost-everywhere pairwise agreement. As an immediate corollary, we obtain a unique recoverability result for the logit model, analogous to Proposition 2. A second insight is that, if we perturb strictly monotone data slightly and this changes the function m continuously, then the bound in (3) grows from zero continuously. Even after the perturbation, swaps preferences recover $\succeq_*$ approximately.

Suppose second that monotonicity is not strict, i.e., there are instances with $x \succ_* y$ but $m(x,y) = 0$. In this case, the bound simplifies to

$$\int_{X\times X} \mathbb{1}(x \succ_* y)\mathbb{1}(m(x,y) = 0) w(x,y)\, d(x,y).$$

This bound reflects that a possible multiplicity of swaps preferences arises from switching pairs where there is no evidence of preferences in either direction. An extreme case is pure uniform trembles over all budget sets without any singular choice, which implies $m(x,y) = 0$ for all pairs. In this extreme case, the bound takes its maximal value of $1/2$ (when w is a symmetric density). It is not possible to improve this bound, because any preference is a swaps preference and hence the maximal distance of $1/2$ does indeed arise between any given $\succeq_*$ and some swaps preference. If we perturb such data slightly, the swaps preference can change substantially. Non-strict monotonicity is not sufficient for approximate recoverability.

[11] Strict monotonicity requires each non-indifferent pair of alternatives to be jointly observed in a collection of budget sets with strictly positive measure, as otherwise $m(x,y) = m(y,x) = 0$ for the respective pair.

[12] When $x \succ_* y$ implies $m(x,y) > 0$, the integrand in (3) converges pointwise to zero as $\alpha \to \infty$. Since the integrand is bounded in absolute value by the integrable function $w(x,y)$, the dominated convergence theorem implies that the entire integral converges to zero as $\alpha \to \infty$, which implies that the infimum is zero.

Finally, if we start from a data set D that is not $\succeq_*$-monotone in the first place, then the bound in (3) can be interpreted as quantifying how close D is to $\succeq_*$-monotonicity. For example, the bound is maximal for "strictly anti-monotone" data sets. By contrast, when $\succeq_*$ and $m(x, y)$ sometimes agree and sometimes disagree on pairwise comparisons, taking the infimum over $\alpha > 0$ optimally trades off the countervailing effects, taking into account both the strength of evidence $m(x, y)$ and the weight $w(x, y)$ of the different pairs.

3 Computation

In this section, we provide methods to compute the swaps index and swaps preferences. In contrast to the previous section, we restrict attention to the case where only finitely many choices are observed and where preferences are monotone.

3.1 Formal Framework

The decision-maker is assumed to choose one bundle from each of finitely many budget sets. With a slight abuse of notation, we summarize the observed choice data by

$$D = ((B_1, x_1), \dots, (B_T, x_T)),$$

where x_t is the bundle chosen from budget $B_t \subseteq \mathbb{R}^L_+$ and T is the number of decisions. Denote the set of bundles that are weakly larger in all coordinates than a given bundle x by $x^{\geq} = \{y \in \mathbb{R}^L_+ \mid y^i \geq x^i \ \forall i = 1, \dots, L\}$ and the set of bundles that are strictly larger in all coordinates by $x^{\gg} = \{y \in \mathbb{R}^L_+ \mid y^i > x^i \ \forall i = 1, \dots, L\}$. The sets $x^{\leq}$ and $x^{\ll}$ of pointwise smaller bundles are defined analogously. Following Forges and Minelli (2009), we assume that budget sets can be described by a function $g_t : \mathbb{R}^L_+ \to \mathbb{R}$ such that $B_t = \{x \in \mathbb{R}^L_+ \mid g_t(x) \leq 0\}$. For example, in the linear case with prices $\delta_t \in \mathbb{R}^L_{++}$ and income $w_t > 0$, we have $g_t(x) = \delta_t \cdot x - w_t$. We assume that g_t is continuous and strictly increasing, i.e., $g_t(y) > g_t(x)$ whenever $y \in x^{\geq}$ and $x \neq y$. We also assume that all budget sets are nontrivial (there exists $x \in \mathbb{R}^L_{++}$ with $g_t(x) < 0$) and that all chosen bundles are on the boundary of the budget set ($g_t(x_t) = 0$) and mutually distinct ($x_t \neq x_{t'}$ for all $t \neq t'$).

Denote by $\mathscr{R}_M \subset \mathscr{R}$ the subset of preferences that are monotone, by which we mean that $y \succeq x$ when $y \in x^{\geq}$ and $y \succ x$ when $y \in x^{\gg}$. Denote by $\mathscr{R}_{CM} \subset \mathscr{R}_M$ the subset of monotone preferences that are continuous. For the following results, we fix an arbitrary set $\mathscr{R}^*$ with $\mathscr{R}_{CM} \subseteq \mathscr{R}^* \subseteq \mathscr{R}_M$ and define the swaps index as the infimum of the swaps distance between the data and those preferences, $I(D) = \inf_{\succeq \in \mathscr{R}^*} W(\succeq, D)$.

3.2 Swaps Index

Denote the set of strict rankings of the chosen bundles that respect monotonicity by

$$\Pi^D = \{\pi : \{x_t\}_{t=1}^T \to \{1, \ldots, T\} \mid \pi \text{ is bijective, and } \pi(x_{t'}) \geq \pi(x_t) \text{ when } x_{t'} \in x_t^{\gtrsim}\}.$$

The following result provides a simple minimization problem that yields the swaps index.

Proposition 4. *Given choice data $D = ((B_1, x_1), \ldots, (B_T, x_T))$, the swaps index is*

$$I(D) = \min_{\pi \in \Pi^D} \sum_{t=1}^{T} \lambda\left(B_t \cap \left(\bigcup_{t' \mid \pi(x_{t'}) > \pi(x_t)} x_{t'}^{\gtrsim}\right)\right). \tag{4}$$

According to the proposition, each ranking $\pi \in \Pi^D$ of the chosen bundles is associated with a unique value, obtained by adding across all observations t the Lebesgue measure of how much the principal upper sets $x_{t'}^{\gtrsim}$ of all bundles $x_{t'}$ that are ranked above x_t cut into the budget set B_t. This is a crucial simplification, as we can restrict attention to one specific and easy-to-calculate configuration of upper contour sets. Then we need to optimize only over the rankings of the observed bundles.[13]

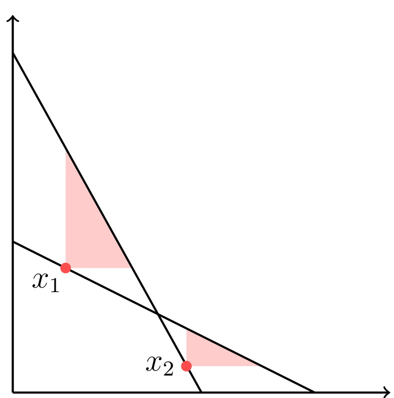

Figure 4: GARP Violation.

In Figure 4, the procedure is illustrated for a simple case with $T = 2$ linear budgets and a GARP violation, where the problem reduces to comparing the sizes of the two shaded red areas. The area next to x_1 corresponds to the case where we rank $\pi(x_1) > \pi(x_2)$, and the

[13] Our construction differs from the suggestion in AB (Online Appendix D.3) because they impose the requirement of convexity of preferences. Our result generalizes Lanier and Quah (2026, Lemma 2) by allowing for nonlinear budgets and preferences that are not necessarily continuous. Our result holds irrespective of the chosen set $\mathscr{R}^*$ of preferences. This has important consequences for the existence of swaps preferences, as we will show in the next subsection. Lastly, problem (4) is not one of the linear ordering problems that AB discuss for the finite case, because the right-hand side cannot be written as a sum of pairwise costs.

area next to x_2 corresponds to the case where $\pi(x_2) > \pi(x_1)$. Since the latter area is smaller in the figure, $\pi(x_2) > \pi(x_1)$ is the solution to problem (4) and determines the swaps index.

Next, we offer an alternative representation of the swaps index as a binary minimization problem, where revealed preferences are removed to make the revealed preference graph acyclical. This type of problem has previously been studied by Mononen (2023) and is known to be computationally feasible for typical data sets from the literature. To state the result, define the strict revealed preference P by $x_t P x_{t'} \Leftrightarrow g_t(x_{t'}) < 0$. Furthermore, define the dominance relation M by $x_t M x_{t'} \Leftrightarrow x_t \in x_{t'}^{\geqq}$.

Proposition 5. *Given choice data* $D = \big((B_1, x_1), \ldots, (B_T, x_T)\big)$*, the swaps index is*

$$I(D) = \min_{U \subseteq P \setminus M} \sum_{t=1}^{T} \lambda\Big(B_t \cap \Big(\bigcup_{t' | (x_t, x_{t'}) \in U} x_{t'}^{\geqq}\Big)\Big) \text{ such that } P \setminus U \text{ is acyclical.} \tag{5}$$

This reformulation of the previous problem has a graphical interpretation. The strict revealed preferences correspond to a directed graph where each chosen bundle x_t is a vertex and there is a directed edge between x_t and $x_{t'}$ when $x_t P x_{t'}$. We want to make this revealed preference graph acyclical by removing edges $(x_t, x_{t'})$, keeping those that arise from dominance of the bundles. There is a cost of removing edges, based on the Lebesgue measure of how much the union of the sets $x_{t'}^{\geqq}$ cuts into B_t, like before. Hence, the cost of removing multiple edges from a given x_t is not additive if the sets $x_{t'}^{\geqq}$ overlap. Deleting one edge can reduce the marginal cost of deleting another. This is in contrast to Mononen (2023), where the cost of removing each edge is defined independently.

3.3 Swaps Preferences

Propositions 4 and 5 give expressions for the swaps index but remain silent about the swaps preferences. A swaps preference does not even have to exist. Indeed, Lanier and Quah (2026) have shown that a swaps preference often does not exist in the conventional setting with linear budget sets: it never exists when the data violate GARP and one requires continuity of preferences.

This is illustrated in Figure 5, based on the previous example in Figure 4. For the solution $\pi(x_2) > \pi(x_1)$ to problem (4), the bold orange and blue lines sketch limiting indifference curves that explain the two choices with minimal swaps distance. These limiting indifference curves are overlapping (but depicted with a small offset for better readability). Continuous preferences can only approximate such overlapping indifference curves, which explains the non-existence of a continuous swaps preference as established by Lanier and Quah (2026).

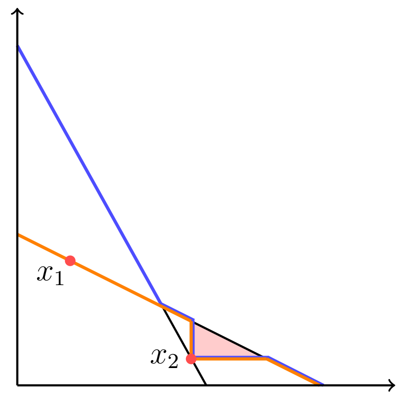

Figure 5: Limiting Indifference Curves.

As the next result shows, the conclusion is very different if we allow for all monotone preferences.

Proposition 6. *Given choice data $D = \big((B_1, x_1), \ldots, (B_T, x_T)\big)$, if $\mathscr{R}^* = \mathscr{R}_M$ then a swaps preference exists.*

In the proof, we construct a monotone utility function u that agrees with an arbitrary given solution π to problem (4) on the chosen bundles and achieves the swaps index exactly. In the example in Figure 5, we start by assigning a utility of zero to all bundles. We then change the utility to $u(x) = 1$ for all bundles $x \in x_1^{\geq}$ (which includes the chosen bundle x_1). The next step is to further change the utility to $u(x) = 2$ for all bundles $x \in x_2^{\geq}$, possibly overriding earlier changes. In the general case, this procedure is iterated following the order given by the solution π to problem (4). The resulting utility function is weakly monotone and takes integer values. It can then be perturbed slightly to achieve monotonicity without overturning the ranking of the chosen bundles. The resulting function is not continuous but, by construction, achieves the swaps index as characterized in Proposition 4. Hence a swaps preference is guaranteed to exist in the set $\mathscr{R}_M$.

Lanier and Quah (2026) suggest the notion of robust preferences for a given measure of distance between preferences and data. Robust preferences are those binary comparisons that all preferences close enough to the infimum distance must have in common. They study robust preferences for the distance measures underlying the Afriat, Varian, and Houtman-Maks indices. In the following, we will study robust preferences for the swaps distance.

Definition 2. *Given choice data $D = \big((B_1, x_1), \ldots, (B_T, x_T)\big)$, x is robustly swaps-preferred to y if there exists $\succeq^0 \in \mathscr{R}^*$ such that all $\succeq \in \mathscr{R}^*$ with $W(\succeq, D) \leq W(\succeq^0, D)$ satisfy $x \succeq y$.*

Denote by $\succeq^*$ the binary relation that summarizes all robust swaps preferences and by $\succ^*$ its asymmetric part. The notion of a robust swaps preference is of interest not only when a swaps preference does not exist but also when multiple swaps preferences exist. In the latter case, it follows from the definition that $x \succeq^* y$ if and only if all swaps preferences agree on that ranking (see the discussion in Lanier and Quah, 2026). Hence, robust swaps preferences capture the preference components that are truly revealed from the data and remain silent on those preference components that are ambiguous due to multiplicity. This is similar in spirit to the unambiguous choice relation by Bernheim and Rangel (2009).

Denote by Π^{D*} the set of minimizers in our optimization problem (4). The following result shows that these solutions describe robust swaps preferences among the chosen bundles.

Proposition 7. *Given choice data $D = \big((B_1, x_1), \dots, (B_T, x_T)\big)$, for each x_t and $x_{t'}$, the following (i) implies (ii):*

(i) $\pi(x_t) > \pi(x_{t'})$ for all $\pi \in \Pi^{D*}$.

(ii) $x_t \succ^* x_{t'}$.

If $\mathscr{R}^* = \mathscr{R}_M$ or if a swaps preference does not exist, then (ii) also implies (i).

If all solutions of problem (4) agree on the ranking of two distinct bundles, then that ranking corresponds to a robust strict swaps preference. An immediate corollary is that, if there is a unique minimizer in (4), then it represents the robust swaps preferences among the chosen bundles and these are complete. This is the case in the example in Figures 4 and 5, where we obtain $x_2 \succ^* x_1$. The converse implication is that any robust strict swaps preference generates agreement among all solutions to problem (4) and can therefore be detected this way. Equivalently, disagreement among multiple solutions implies that there is no corresponding robust swaps preference. This converse is true either when a swaps preference does not exist (as in Lanier and Quah, 2026, when continuity of preferences is required and the data violate GARP) or when all monotone preferences are admissible (which is sufficient for the existence of a swaps preference by Proposition 6). It cannot be true for arbitrary preference sets. In a case where multiple solutions to (4) exist, and hence there are multiple swaps preferences in $\mathscr{R}_M$ that disagree on some bundles, we can always artificially create a robust swaps preference by removing exactly all disagreeing preferences from $\mathscr{R}_M$. Such artificial robust swaps preferences cannot be detected by the optimization problem (4). We conclude this discussion by pointing out that the proposition characterizes robust swaps preferences among the chosen bundles, but of course these can be extended to non-chosen bundles using monotonicity and transitivity.

It is instructive to also consider the solutions to our alternative minimization problem of removing revealed preferences. Denote by $\mathscr{U}^*$ the set of minimizers in problem (5) and by $\text{trans}(\cdot)$ the transitive closure of a binary relation.

Proposition 8. *Given choice data $D = \big((B_1, x_1), \ldots, (B_T, x_T)\big)$, it holds that $\pi(x_t) > \pi(x_{t'})$ for all $\pi \in \Pi^{D*}$ if and only if $(x_t, x_{t'}) \in \text{trans}(P \setminus U)$ for all $U \in \mathscr{U}^*$.*

This result implies that the solutions to problem (5) characterize robust swaps preferences among the chosen bundles in exactly the same way as the solutions to problem (4). A benefit of the alternative characterization is its computational convenience, which we will exploit in Section 4. Furthermore, the result shows that, under the assumptions for which Proposition 7 asserts $(ii) \Rightarrow (i)$, a necessary condition for a robust swaps preference $x_t \succ^* x_{t'}$ is that we observe a conventional revealed preference $(x_t, x_{t'}) \in \text{trans}(P)$ in the data.

4 Empirical Application

In this section, we provide a short empirical application. The goal is to show that our theoretical results can be used in practice and that the swaps index generates insights that go beyond older measures of rationality.

We use the experiment of Castillo and Freer (2023), whom we thank for sharing the data. In that experiment, 65 subjects chose bundles from 20 different budget sets. A bundle (m, x) specifies a monetary payment m and the number x of real-effort tasks (data entry) that the subject does *not* have to solve, out of ten tasks. Defining x in this way implies that x is a good rather than a bad. Importantly, the budget frontiers in the experiment are concave and only piecewise linear. This design allows Castillo and Freer (2023) to test whether subjects' preferences are quasilinear. For us, the design is interesting because several of the conventional measures of rationality are not defined for nonlinear budget sets, so that the swaps index is a natural candidate in this application.

Castillo and Freer (2023) use the Houtman-Maks index to calculate a distance between observed choices and different rational choice benchmarks. For example, they check what fraction of observations one would have to remove until a subject's remaining choices become rationalizable by quasilinear preferences. This approach is well-defined irrespective of the shape of budget sets. To obtain statistical tests, they bootstrap the distribution of the index for choices that are randomly sampled from the experimental data. The null hypothesis of random choice is rejected in favor of preference maximization when a subject's rationality index is better than some percentile of the distribution for random choices.

We mimic this approach and calculate the Houtman-Maks distance to the benchmark of maximizing monotone preferences. Table 1 reports cutoffs for rejecting random choice and the resulting individual pass rates. For example, among 400′000 randomly generated choice data sets, the most rational ten percent generate a Houtman-Maks index of 0.10 or less, because removing two of the 20 choices is sufficient for rationalizability. Among all 65

Table 1: Pass rates and cutoffs.

	Pass rates			Cutoffs		
	$p = 0.10$	$p = 0.05$	$p = 0.01$	$p = 0.10$	$p = 0.05$	$p = 0.01$
H-M	85% (55/65)	75% (49/65)	75% (49/65)	0.1000	0.0500	0.0500
Swaps	85% (55/65)	74% (48/65)	52% (34/65)	0.3463	0.0911	0.0001

subjects, 55 achieve this value or less. Hence the rate of subjects who pass as more rational than random is about 85% at a significance level of $p = 0.10$.

We conduct exactly the same exercise for the swaps index, using Proposition 5 for computation. With the MATLAB solver, the average computation time for a swaps index is 0.02 seconds and the maximum is 0.11 seconds. The results are reported in the second row of Table 1. As can be seen, the swaps index agrees with the Houtman-Maks index for the least stringent significance level $p = 0.10$. The two measures diverge for lower values of p. Since subjects make 20 choices, the Houtman-Maks index takes only discrete values in steps of size $1/20 = 0.05$. As a consequence, the cutoffs do not even differ between $p = 0.05$ and $p = 0.01$, generating the uniform pass rate of 75%. The swaps index is continuous in the geometry of the budget sets and separates these two cases. The pass rate is a slightly lower 74% at $p = 0.05$ and a substantially lower 52% at $p = 0.01$. We consider this finer differentiation a desirable property for a measure of rationality.[14]

We also derive robust preferences for both the Houtman-Maks index and the swaps distance. For swaps, we do this using Proposition 8. We consider all solutions that remove revealed preferences from the revealed preference graph (keeping those due to dominance) and achieve acyclicity at minimal cost. Robust preferences among the chosen bundles are those that survive in the transitive closure of each solution. We proceed analogously for Houtman-Maks. We consider all solutions that remove the minimal number of observations from the data to achieve rationalizability. Robust preferences are those that survive in the transitive closure of the revealed preferences (adding those due to dominance) in each solution.

Table 2 summarizes the results. On average across subjects, we obtain about 211 strict revealed preferences in the conventional sense, i.e., there are 211 ordered pairs of chosen bundles in the transitive closure of P. This is a large number. Since there exist 380 ordered pairs of the 20 chosen bundles, an upper bound on the number of strict revealed preferences

[14] Castillo and Freer (2023, Appendix 2) additionally study two versions of the Afriat index but point out that its definition and interpretation are ambiguous for nonlinear budget sets. In their test of locally nonsatiated preferences, one version of the index differentiates the two cases $p = 0.05$ and $p = 0.01$ strongly (their Table 5) but the other version much less (their Table 6).

Table 2: Robust preferences.

	# trans(P)	# $\succ^*$	% complete
H-M	211.03	157.13	0.83
Swaps	211.03	184.42	0.97

that are compatible with acyclicity is $380/2 = 190$. Hence a substantial number of revealed preference cycles exist. With the Houtman-Maks approach, about 157 of these revealed preferences are robust on average. With the swaps approach, about 184 preferences are robust. This number is close to the theoretical upper bound of 190, implying that robust swaps preferences among the chosen bundles are close to complete (average 97% completeness, compared to only 83% for Houtman-Maks). The swaps approach is highly informative and generates preferences suitable for meaningful welfare analysis.

5 Conclusions

We study the swaps index by Apesteguia and Ballester (2015) in a consumer choice setting with infinite commodity spaces. We show that the swaps approach has a solid foundation, because many (but not all) plausible boundedly rational models of behavior generate choice data for which the true preference of the decision-maker is recovered as a swaps preference. We also show how the swaps index and robust swaps preferences can be computed for standard consumer choice data sets. Our empirical application illustrates the applicability and usefulness of the approach.

The basic swaps index is nonparametric. Parametric assumptions on utility functions are often desirable and commonly used in practice (see e.g. Halevy et al., 2018). Information in the form of parametric assumptions can easily be incorporated in our setting. This has been done, for example, by Chung et al. (2025) in their study of the causal effect of attention-enhancing drugs on rationality.

References

Afriat, S. (1973). On a system of inequalities in demand analysis: An extension of the classical method. *International Economic Review*, 14:460–472.

Alós-Ferrer, C., Fehr, E., and Netzer, N. (2021). Time will tell: Recovering preferences when choices are noisy. *Journal of Political Economy*, 129:1828–1877.

Anderson, S., de Palma, A., and Thisse, J. (1992). *Discrete Choice Theory of Product Differentiation*. Cambridge: MIT Press.

Apesteguia, J. and Ballester, M. (2015). A measure of rationality and welfare. *Journal of Political Economy*, 123:1278–1310.

Apesteguia, J. and Ballester, M. (2018). Monotone stochastic choice models: The case of risk and time preferences. *Journal of Political Economy*, 126:74–106.

Arrow, K. (1959). Rational choice functions and orderings. *Economica*, 26:121–127.

Benkert, J.-M. and Netzer, N. (2018). Informational requirements of nudging. *Journal of Political Economy*, 126:2323–2355.

Bernheim, B. and Rangel, A. (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, 124:51–104.

Castillo, M. and Freer, M. (2023). A general revealed preference test for quasilinear preferences: Theory and experiments. *Experimental Economics*, 26:673–696.

Choi, S., Kariv, S., Müller, W., and Silverman, D. (2014). Who is (more) rational? *American Economic Review*, 104:1518–1550.

Chung, H.-K., Doren, N., Mononen, L., Lu, M., Grueschow, M., Hayward-Könnecke, H., Jetter, A., Quednow, B., Netzer, N., and Tobler, P. (2025). Improving rationality by increasing attention. Mimeo.

Dean, M. and Martin, D. (2016). Measuring rationality with the minimum cost of revealed preference violations. *Review of Economics and Statistics*, 98:524–534.

Echenique, F., Imai, T., and Saito, K. (2023). Approximate expected utility rationalization. *Journal of the European Economic Association*, 21:1821–1864.

Echenique, F., Lee, S., and Shum, M. (2011). The money pump as a measure of revealed preference violations. *Journal of Political Economy*, 119:1201–1223.

Forges, F. and Minelli, E. (2009). Afriat's theorem for general budget sets. *Journal of Economic Theory*, 144:135–145.

Halevy, Y., Persitz, D., and Zrill, L. (2018). Parametric recoverability of preferences. *Journal of Political Economy*, 126:1558–1593.

Harbaugh, W., Krause, K., and Berry, T. (2001). Garp for kids: On the development of rational choice behavior. *American Economic Review*, 91:1539–1545.

Heufer, J. (2008). A geometric measure for the violation of utility maximization. Ruhr Economic Papers 69.

Houthakker, H. (1950). Revealed preference and the utility function. *Economica*, 17:159–174.

Houtman, M. and Maks, J. (1985). Determining all maximal data subsets consistent with revealed preference. *Kwantitatieve Methoden*, 6:89–104.

Kőszegi, B. and Rabin, M. (2007). Mistakes in choice-based welfare analysis. *American Economic Review, Papers and Proceedings*, 97:477–481.

Kim, H., Choi, S., Kim, B., and Pop-Eleches, C. (2018). The role of education interventions in improving economic rationality. *Science*, 362:83–86.

Lanier, J. and Quah, J.-H. (2026). Goodness-of-fit and utility estimation: what's possible and what's not. Mimeo.

Mas-Colell, A. (1977). The recoverability of consumers' preferences from market demand behavior. *Econometrica*, 45:1409–1430.

Matejka, F. and McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105:272–298.

Mononen, L. (2023). Computing and comparing measures of rationality. Mimeo.

Rubinstein, A. and Salant, Y. (2012). Eliciting welfare preferences from behavioural data sets. *Review of Economic Studies*, 79:375–387.

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

Samuelson, P. (1938). A note on the pure theory of consumer's behavior. *Economica*, 56:61–71.

Thaler, R. and Sunstein, C. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press.

Varian, H. (1990). Goodness-of-fit in optimizing models. *Journal of Econometrics*, 43:125–140.

A Appendix

A.1 Proof of Proposition 1

Fix $\succeq_* \in \mathscr{R}$ and an arbitrary $\succeq \in \mathscr{R}$. The distance between $\succeq$ and D can be decomposed as

$$W(\succeq, D) = \int_\Delta \int_{B_\delta} \lambda(\{y \in B_\delta \mid y \succ x\}) d\mu_\delta^s(x) d\nu(\delta) \tag{6}$$

$$+ \int_\Delta \int_{B_\delta} \lambda(\{y \in B_\delta \mid y \succ x\}) f_\delta(x) dx d\nu(\delta). \tag{7}$$

For the inner integral of the first summand (6), we have, for all $\delta \in \Delta$,

$$\int_{B_\delta} \lambda(\{y \in B_\delta \mid y \succ x\}) d\mu_\delta^s(x) \geq 0 = \int_{B_\delta} \lambda(\{y \in B_\delta \mid y \succ_* x\}) d\mu_\delta^s(x),$$

because the singular component μ_δ^s is rationalized by $\succeq_*$ by definition of $\succeq_*$-monotonicity. It follows that the first summand (6) is minimized by $\succeq_*$ among all preferences in $\mathscr{R}$.

Considering the second summand (7), first define

$$A(x, y) = \int_{\{\delta \in \Delta \mid \{x,y\} \subseteq B_\delta\}} f_\delta(x) d\nu(\delta) = \int_\Delta \mathbb{1}(\{x, y\} \subseteq B_\delta) f_\delta(x) d\nu(\delta),$$

so that, for all $x, y \in X$, $x \succeq_* y$ implies $A(x, y) \geq A(y, x)$ by $\succeq_*$-monotonicity. We can then rewrite (7) as

$$\begin{aligned}
&\int_\Delta \int_{B_\delta} \lambda(\{y \in B_\delta \mid y \succ x\}) f_\delta(x) dx d\nu(\delta) \\
&= \int_\Delta \int_{B_\delta} \int_{B_\delta} \mathbb{1}(y \succ x) dy f_\delta(x) dx \, d\nu(\delta) \\
&= \int_\Delta \int_X \int_X \mathbb{1}(\{x, y\} \subseteq B_\delta) \mathbb{1}(y \succ x) dy f_\delta(x) dx d\nu(\delta) \\
&= \int_X \int_X \int_\Delta \mathbb{1}(\{x, y\} \subseteq B_\delta) \mathbb{1}(y \succ x) f_\delta(x) d\nu(\delta) dy dx \\
&= \int_X \int_X \mathbb{1}(y \succ x) \left(\int_\Delta \mathbb{1}(\{x, y\} \subseteq B_\delta) f_\delta(x) d\nu(\delta) \right) dy dx \\
&= \int_X \int_X \mathbb{1}(y \succ x) A(x, y) dy dx.
\end{aligned}$$

Fix an arbitrary linear order $\rhd$ on X with measurable upper contour sets (e.g. the lexicographical order) and define

$$Y = \{(x, y) \in X \times X \mid y \rhd x,\ x \not\sim y,\ x \not\sim_* y\},$$

so that every pair of bundles is contained in one order only, and not at all if either $\succeq$ or $\succeq_*$ involves an indifference. We can then further rewrite

$$\begin{aligned}
&\int_X \int_X \mathbb{1}(y \succ x) A(x, y) dy dx \\
&= \int_X \int_X \mathbb{1}(y \succ x) \mathbb{1}(y \not\sim_* x) A(x, y) dy dx \\
&= \int_{X \times X} \mathbb{1}(y \succ x) \mathbb{1}(y \not\sim_* x) A(x, y) d(x, y) \\
&= \int_Y (\mathbb{1}(y \succ x) A(x, y) + \mathbb{1}(x \succ y) A(y, x))\, d(x, y),
\end{aligned}$$

where the first equality uses that the Lebesgue measure of indifference sets is zero for all preferences in $\mathscr{R}$ and hence $\mathbb{1}(y \not\sim_* x) = 1$ almost everywhere in the inner integral. For any $(x, y) \in Y$ it holds that

$$\mathbb{1}(y \succ x) A(x, y) + \mathbb{1}(x \succ y) A(y, x) \geq \mathbb{1}(y \succ_* x) A(x, y) + \mathbb{1}(x \succ_* y) A(y, x),$$

because both sides of the inequality are identical when $\succeq$ and $\succeq_*$ agree, and the inequality follows from $\succeq_*$-monotonicity when $\succeq$ and $\succeq_*$ disagree. Hence

$$\begin{aligned}
&\int_\Delta \int_{B_\delta} \lambda(\{y \in B_\delta \mid y \succ x\}) f_\delta(x) dx d\nu(\delta) \\
&= \int_Y \big(\mathbb{1}(y \succ x) A(x, y) + \mathbb{1}(x \succ y) A(y, x)\big) d(x, y) \\
&\geq \int_Y \big(\mathbb{1}(y \succ_* x) A(x, y) + \mathbb{1}(x \succ_* y) A(y, x)\big) d(x, y) \\
&= \int_\Delta \int_{B_\delta} \lambda(\{y \in B_\delta \mid y \succ_* x\}) f_\delta(x) dx d\nu(\delta).
\end{aligned}$$

It follows that the second summand (7) is also minimized by $\succeq_*$ among all preferences in $\mathscr{R}$. Hence $\succeq_*$ is a swaps preference.

A.2 Proof of Proposition 2

Let $\succeq \neq \succeq_*$ be any continuous, monotone, and strictly convex preference. For each price vector $\delta \in \Delta = \mathbb{R}^L_{++}$, the optimal bundles in B_δ are unique so we can represent them by a Walrasian demand function, denoted by $x^*(\delta)$ for $\succeq_*$ and by $x(\delta)$ for $\succeq$. It follows from Berge's maximum principle that $x^*(\delta)$ and $x(\delta)$ are continuous functions.

For uniqueness, we need to show that $W(\succeq, D) > W(\succeq_*, D)$. As in the proof of Proposition 1, we can decompose the swaps distance for any preference $\succeq'$ into a singular part and an absolutely continuous part, $W(\succeq', D) = W^s(\succeq', D) + W^c(\succeq', D)$, where

$$W^s(\succeq', D) = \int_\Delta \lambda(\{y \in B_\delta \mid y \succ' x^*(\delta)\})(1-\epsilon)d\nu(\delta)$$

and

$$W^c(\succeq', D) = \int_\Delta \int_{B_\delta} \lambda(\{y \in B_\delta \mid y \succ' x\})\frac{\epsilon}{\lambda(B_\delta)}dxd\nu(\delta),$$

using the definition of the uniform trembles model with true preference $\succeq_*$.

We first state a useful lemma.

Lemma 1. *For any bounded and Lebesgue measurable set $B \subseteq \mathbb{R}^L_+$, the unweighted average volume of the upper contour sets is the same for all preferences,*

$$\int_B \lambda(\{y \in B \mid y \succ x\})dx = \frac{\lambda(B)^2}{2} \quad \forall \succeq \in \mathscr{R}. \tag{8}$$

Proof. The result is immediate when $\lambda(B) = 0$. When $\lambda(B) > 0$, the Lebesgue measure of the upper contour set of x within B is the probability that a uniformly drawn bundle from B is preferred over x, multiplied by the constant $\lambda(B)$. Averaging this probability over all x in B gives the probability that among two bundles that are sequentially drawn uniformly and independently from B, the second is preferred over the first. This probability is independent of the preference $\succeq \in \mathscr{R}$ and equals 1/2. Hence, integrating $\lambda(\{y \in B \mid y \succ x\})$ over all x in B yields $\lambda(B)^2/2$.[15] □

Lemma 1 immediately implies $W^c(\succeq, D) = W^c(\succeq_*, D)$. It also holds that $W^s(\succeq_*, D) = 0$ by definition of the uniform trembles model. Hence it suffices to show that $W^s(\succeq, D) > 0$.

Since $\succeq \neq \succeq_*$, it follows from Theorem 2 in Mas-Colell (1977) – using the property that $\succeq_*$ is lipschitzian – that there exists a price vector δ_0 such that $x(\delta_0) \neq x^*(\delta_0)$. Continuity of x and x^* then implies that there exists an open ball $U(\delta_0)$ around δ_0 such that $x(\delta) \neq x^*(\delta)$ and in particular $x(\delta) \succ x^*(\delta)$ for all $\delta \in U(\delta_0)$. Let u be a continuous utility representation of $\succeq$. Then $u(x(\delta)) > u(x^*(\delta))$ for all $\delta \in U(\delta_0)$.

It follows that $u^{-1}((u(x^*(\delta)), u(x(\delta)))) \subset \mathbb{R}^L_+$ is open and nonempty for all $\delta \in U(\delta_0)$. It

[15] We are grateful to Jakub Steiner for suggesting this short proof.

furthermore follows that $u^{-1}((u(x^*(\delta)), u(x(\delta)))) \cap \text{int}(B_\delta)$ is open and nonempty. Nonemptiness holds because $u(\alpha x(\delta) + (1-\alpha)x^*(\delta)) \in (u(x^*(\delta)), u(x(\delta)))$ for all $\alpha \in (0,1)$, by strict convexity and the fact that $x(\delta)$ uniquely maximizes u on B_δ. Thus by continuity there exists $\gamma \in (0,1)$ such that $u(\gamma(\alpha x(\delta) + (1-\alpha)x^*(\delta))) \in (u(x^*(\delta)), u(x(\delta)))$ and $\gamma(\alpha x(\delta) + (1-\alpha)x^*(\delta)) \in \text{int}(B_\delta)$. Therefore

$$\begin{aligned} W^s(\succeq, D) &= \int_\Delta \lambda(\{y \in B_\delta \mid y \succ x^*(\delta)\})(1-\epsilon)d\nu(\delta) \\ &\geq \int_{U(\delta_0)} \lambda(u^{-1}((u(x^*(\delta)), u(x(\delta)))) \cap \text{int}(B_\delta))(1-\epsilon)d\nu(\delta) > 0. \end{aligned}$$

A.3 Non-Monotonicity of Rational Inattention

Consider an example in the finite setting of Matejka and McKay (2015) with $B = \{x, y\}$. Suppose the decision-maker has one of two possible utility functions. Utility function u_1 with $u_1(x) = 2$ and $u_1(y) = 0$ has prior probability q. Utility function u_2 with $u_2(x) = 0$ and $u_2(y) = 1$ has prior probability $1-q$. Setting $\gamma = 1$ without loss of generality, the normalization condition (17) in Matejka and McKay (2015) becomes

$$\frac{q}{p(x,B) + (1-p(x,B))e^{-2}} + \frac{1-q}{p(x,B) + (1-p(x,B))e^{1}} = 1,$$

from which the unconditional choice probability $p(x, B)$ can be obtained. The two conditional choice probabilities $p(x, B|u_1)$ and $p(x, B|u_2)$ then follow from (2).

For $q = 0.35$, we obtain the numerical solution $p(x, B) \approx 0.45$. Hence $p(x, B) < p(y, B)$ even though $q2 + (1-q)0 = 0.7 > 0.65 = q0 + (1-q)1$, showing that unconditional choice probabilities are not monotone with respect to the prior preference.

For $q = 0.6$, we obtain $p(x, B) \approx 0.89$, from which it follows that $p(x, B|u_1) \approx 0.98$ and $p(x, B|u_2) \approx 0.74$. Hence $p(x, B|u_2) > p(y, B|u_2)$ even though $u_2(x) = 0 < 1 = u_2(y)$, showing that conditional choice probabilities are not monotone with respect to the true preference.

A.4 Proof of Proposition 3

As shown in the proof of Proposition 1, for any $\succeq \in \mathscr{R}$ we can write

$$W(\succeq, D) = \int_X \int_X \mathbb{1}(y \succ x) A(x, y) dy dx,$$

where we use absolute continuity of the data. Given any two preferences $\succeq_1, \succeq_2 \in \mathscr{R}$, define

$$Y = \{(x, y) \in X \times X \mid x \succ_2 y, x \not\succ_1 y\}.$$

Thus,

$$\begin{aligned} W(\succeq_1, D) &= \int_X \int_X \mathbb{1}(y \succ_1 x) A(x, y) dy dx \\ &= \int_X \int_X \mathbb{1}(y \succ_1 x) \mathbb{1}(x \not\succ_2 y) A(x, y) dy dx \\ &= \int_Y \left(\mathbb{1}(y \succ_1 x) A(x, y) + \mathbb{1}(x \succ_1 y) A(y, x)\right) d(x, y) \end{aligned}$$

and

$$\begin{aligned} W(\succeq_2, D) &= \int_X \int_X \mathbb{1}(y \succ_2 x) A(x, y) dy dx \\ &= \int_X \int_X \mathbb{1}(y \succ_2 x) \mathbb{1}(x \not\succ_1 y) A(x, y) dy dx \\ &= \int_Y \left(\mathbb{1}(y \succ_2 x) A(x, y) + \mathbb{1}(x \succ_2 y) A(y, x)\right) d(x, y) \\ &= \int_Y A(y, x) d(x, y). \end{aligned}$$

It follows that

$$\begin{aligned} W(\succeq_1, D) - W(\succeq_2, D) &= \int_Y \mathbb{1}(y \succ_1 x) \left(A(x, y) - A(y, x)\right) d(x, y) \\ &= \int_Y \mathbb{1}(y \succ_1 x) m(x, y) d(x, y). \end{aligned}$$

For every $(x, y) \in Y$ and every $\kappa > 0$,

$$m(x, y) = \kappa w(x, y) - (\kappa w(x, y) - m(x, y)) \geq \kappa w(x, y) - \max\{\kappa w(x, y) - m(x, y), 0\}.$$

Hence,

$$\begin{aligned} W(\succeq_1, D) - W(\succeq_2, D) &\geq \int_Y \mathbb{1}(y \succ_1 x) \left(\kappa w(x, y) - \max\{\kappa w(x, y) - m(x, y), 0\}\right) d(x, y) \\ &= \kappa \int_Y \mathbb{1}(y \succ_1 x) w(x, y) d(x, y) \\ &\quad - \int_Y \mathbb{1}(y \succ_1 x) \max\{\kappa w(x, y) - m(x, y), 0\} d(x, y) \end{aligned}$$

$$\begin{aligned}&\geq \kappa\, d(\succeq_1,\succeq_2)\\&\qquad - \int_Y \max\{\kappa w(x,y)-m(x,y),0\} d(x,y).\end{aligned}$$

For $\succeq_1=\succeq_S$ and $\succeq_2=\succeq_*$ we have $0 \geq W(\succeq_S,D)-W(\succeq_*,D)$ and therefore

$$\begin{aligned}d(\succeq_S,\succeq_*) &\leq \frac{1}{\kappa}\int_Y \max\{\kappa w(x,y)-m(x,y),0\} d(x,y)\\&= \frac{1}{\kappa}\int_{X\times X} \mathbb{1}(x\succ_* y)\mathbb{1}(x \not\succ_S y)\max\{\kappa w(x,y)-m(x,y),0\} d(x,y)\\&= \frac{1}{\kappa}\int_{X\times X} \mathbb{1}(x\succ_* y)\max\{\kappa w(x,y)-m(x,y),0\} d(x,y)\\&= \int_{X\times X} \mathbb{1}(x\succ_* y)\max\{w(x,y)-m(x,y)/\kappa,0\} d(x,y).\end{aligned}$$

Since the inequality holds for all $\kappa > 0$, it also holds for the infimum. Substituting $\alpha = 1/\kappa$ yields the result.

A.5 Proof of Proposition 4

We prove the result by two lemmas. The first lemma shows that (4) is an upper bound for the swaps index even when we minimize over the small set $\mathscr{R}_{CM}$.[16] The second lemma implies that the same expression (4) is a lower bound for the swaps index even when we minimize over the large set $\mathscr{R}_M$. Taken together, the swaps index must be equal to (4) for any $\mathscr{R}_{CM} \subseteq \mathscr{R}^* \subseteq \mathscr{R}_M$.

Lemma 2. *Given choice data $D = \big((B_1,x_1),\ldots,(B_T,x_T)\big)$, for any $\pi \in \Pi^D$,*

$$\inf_{\succeq\in\mathscr{R}_{CM}} W(\succeq,D) \;\leq\; \sum_{t=1}^{T} \lambda\Bigg(B_t \cap \Bigg(\bigcup_{t'|\pi(x_{t'})>\pi(x_t)} x_{t'}^{\geq}\Bigg)\Bigg). \tag{9}$$

Proof. Given any $\pi \in \Pi^D$, we construct monotone continuous preferences that coincide with π on the chosen bundles and achieve a swaps distance arbitrarily close to the right-hand side of (9). Hence the infimum of $W(\succeq,D)$ over $\mathscr{R}_{CM}$ cannot exceed this expression.

Step 1. We first construct limiting upper contour sets and indifference sets, which may still overlap for the different chosen bundles. By permuting indices, we can w.l.o.g. assume that $\pi(x_1) < \pi(x_2) < \ldots < \pi(x_T)$. For each $t \in \{1,\ldots,T\}$, let $\tilde{A}_t$ be the downward closure

[16] Our proof is constructive. Lanier and Quah (2026) prove the result non-constructively for the case of linear budgets. Their technique would apply to nonlinear budgets as well.

of int B_t,

$$\tilde{A}_t = \{x \in \mathbb{R}^L \mid \exists y \in \mathbb{R}^L_+, g_t(y) < 0, x \in y^{\leq}\}.$$

Define $A_t = \mathbb{R}^L \setminus \tilde{A}_t$, noting that A_t contains the bundles $x \in \mathbb{R}^L_+$ with $g_t(x) \geq 0$, but is a subset of $\mathbb{R}^L$ not of $\mathbb{R}^L_+$. Then define $\tilde{U}_0 = \mathbb{R}^L$ and, recursively for each $n \in \{1, \ldots, T\}$,

$$\tilde{U}_n = \tilde{U}_{n-1} \cap \Bigg(A_n \cup \Bigg(\bigcup_{t=n+1}^{T} x_t^{\geq}\Bigg)\Bigg).$$

Thus, the limiting upper contour sets $\tilde{U}_n$ are given by the bundles that are at or above the boundary of the budget set, plus all bundles that are pointwise larger than those bundles x_t which π ranks above x_n, then extended naturally to $\mathbb{R}^L$ and restricted to those bundles which were already in the earlier limiting upper contour set $\tilde{U}_{n-1}$. The construction implies $\tilde{U}_n \subseteq \tilde{U}_{n-1}$ and $x_t^{\geq} \subseteq \tilde{U}_n$ for all $t \in \{n, \ldots, T\}$. The limiting indifference sets are defined by the boundary of $\tilde{U}_n$ in $\mathbb{R}^L_+$. Formally, $\tilde{I}_n = \partial\tilde{U}_n \cap \mathbb{R}^L_+$ for all $n \in \{1, \ldots, T\}$. It holds that $x_n \in \tilde{I}_n$, because x_n is on the boundary of A_n and not contained in any $x_t^{\geq}$ for $t > n$, since π respects monotonicity. The Lebesgue measure of the intersection of B_n with $\tilde{U}_n$ is given by

$$\lambda\Bigg(B_n \cap \Bigg(\bigcup_{t=n+1}^{T} x_t^{\geq}\Bigg)\Bigg).$$

Summing across all $n \in \{1, \ldots, T\}$ yields the right-hand side of (9).

Step 2. We now shift the upper contour sets and indifference sets to remove overlap. Define for all $S \subseteq \mathbb{R}^L$ and $\eta \in \mathbb{R}$

$$S + \eta = \{y \in \mathbb{R}^L \mid \exists x \in S \text{ with } y^i = x^i + \eta \text{ for all } i = 1, \ldots, L\}.$$

Note that for some sufficiently small $\eta^0 > 0$, $x_n \notin (x_t^{\geq} - \eta^0)$ for all n and all $t > n$, since π respects monotonicity. Let $\eta < \eta^0$ and $\eta < \min\{x_t^i \mid 1 \leq i \leq L, 1 \leq t \leq T, x_t^i > 0\}$, where the second inequality ensures that no coordinate of $x_t^{\geq}$ is shifted exactly to zero. Then define $U_0 = \mathbb{R}^L$ and, recursively for each $n \in \{1, \ldots, T\}$,

$$U_n = \left(U_{n-1} + \frac{\eta}{T}\right) \cap \Bigg(A_n \cup \Bigg(\bigcup_{t=n+1}^{T} (x_t^{\geq} - \eta)\Bigg)\Bigg).$$

In contrast to the construction in step 1, the shifted upper contour set U_n contains some additional bundles around each better-ranked $x_t^{\geq}$, but is restricted to bundles strictly away from U_{n-1}, so that $U_n \subset U_{n-1}$. It still holds that $x_t^{\gtrsim} \subseteq U_n$ for all $t \in \{n, \ldots, T\}$, because even T subsequent upward shifts of the upper contour sets of size η/T do not exceed the downward shift of $x_t^{\gtrsim}$ of size η. The respective indifference sets are again defined by the boundary $I_n = \partial U_n \cap \mathbb{R}_+^L$ for all $n \in \{1, \ldots, T\}$. It still holds that $x_n \in I_n$ because $x_n \notin (x_t^{\gtrsim} - \eta)$ for $t > n$ when η is small enough, as pointed out above. Furthermore, any ray through the origin intersects these indifference sets in at most one point, because g_t is strictly increasing and η is small enough to not shift any coordinate of $x_t^{\gtrsim}$ to zero, as also pointed out above.

Step 3. We next construct a preliminary utility function $\tilde{u} : \mathbb{R}_+^L \to \mathbb{R}$ that respects the shifted upper contour sets and indifference sets. First define $\tilde{u}(0) = 0$. Consider then any non-zero $x \in \mathbb{R}_+^L$ with $x \notin U_T$. There exist unique $n \in \{0, 1, \ldots, T-1\}$, $\alpha \leq 1$, and $\beta > 1$ such that $\alpha x \in I_n$ and $\beta x \in I_{n+1}$, where we let $I_0 = \{0\}$. Define

$$\tilde{u}(x) = n + \frac{1-\alpha}{\beta - \alpha}.$$

Finally, consider any $x \in U_T$. There exists a unique $\alpha \leq 1$ such that $\alpha x \in I_T$. Define

$$\tilde{u}(x) = \frac{1}{\alpha} T.$$

The construction implies that $\tilde{u}$ is continuous and strictly increasing along rays through the origin, and that $\tilde{u}(x) = n$ whenever $x \in I_n$.

Step 4. We finally construct a utility function $u : \mathbb{R}_+^L \to \mathbb{R}$ that additionally respects monotonicity. For each $x \in \mathbb{R}_+^L$, define

$$u(x) = \max_{y \in x^{\leq}} \tilde{u}(y).$$

Hence, u is continuous and monotone, and $u(x) = n$ still holds when $x \in I_n$.

Step 5. When $\eta \to 0$, the upper contour sets of u converge to the limiting sets constructed in step 1. The result now follows from continuity of the Lebesgue measure. □

The second lemma implies that (4) is also a lower bound for the swaps index even when we minimize over the large set $\mathscr{R}_M$. The result is slightly stronger to be also useful later on. Denote the set of strict rankings of the chosen bundles that respect strict monotonicity by

$$\Pi_{\gg}^D = \{\pi : \{x_t\}_{t=1}^T \to \{1, \ldots, T\} \mid \pi \text{ is bijective, and } \pi(x_{t'}) \geq \pi(x_t) \text{ when } x_{t'} \in x_t^{\gg}\}.$$

Lemma 3. *Given choice data $D = ((B_1, x_1), \ldots, (B_T, x_T))$, for any $\pi \in \Pi^D_{\gg}$ and $\succeq \in \mathscr{R}_M$ such that $x_{t'} \succ x_t$ implies $\pi(x_{t'}) > \pi(x_t)$,*

$$\sum_{t=1}^{T} \lambda(\{y \in B_t \mid y \succ x_t\}) \geq \sum_{t=1}^{T} \lambda\Big(B_t \cap \Big(\bigcup_{t' \mid \pi(x_{t'}) > \pi(x_t)} x_{t'}^{\geq}\Big)\Big).$$

Proof. For all $t = 1, \ldots, T$,

$$\begin{aligned}\lambda(\{y \in B_t \mid y \succ x_t\}) &= \lambda(B_t \cap \{y \in \mathbb{R}^L_+ \mid y \succ x_t\}) \\ &\geq \lambda\Big(B_t \cap \Big(\bigcup_{t' \mid \pi(x_{t'}) > \pi(x_t)} x_{t'}^{\geq}\Big)\Big),\end{aligned}$$

where the inequality holds because for any t' with $\pi(x_{t'}) > \pi(x_t)$ we either have $x_{t'} \succ x_t$ and thus $x_{t'}^{\geq} \subseteq \{y \in \mathbb{R}^L_+ \mid y \succ x_t\}$, or $x_{t'} \sim x_t$ in which case $x_{t'}^{\geq} \setminus \{y \in \mathbb{R}^L_+ \mid y \succ x_t\}$ has Lebesgue measure zero. □

Lemma 3 implies that (4) is a lower bound for the swaps index defined over $\mathscr{R}_M$, after observing that, for any $\succeq \in \mathscr{R}_M$, there exists $\pi \in \Pi^D \subseteq \Pi^D_{\gg}$ such that $x_{t'} \succ x_t$ implies $\pi(x_{t'}) > \pi(x_t)$, by monotonicity of the preference $\succeq$.

A.6 Proof of Proposition 5

For a ranking $\pi \in \Pi^D$, define

$$C(\pi) = \sum_{t=1}^{T} \lambda\Big(B_t \cap \Big(\bigcup_{t' \mid \pi(x_{t'}) > \pi(x_t)} x_{t'}^{\geq}\Big)\Big).$$

By Proposition 4, the swaps index is $I(D) = \min_{\pi \in \Pi^D} C(\pi)$. Define the set of deleted revealed-preference edges induced by π as

$$A_\pi = \{(x_t, x_{t'}) \in P \mid \pi(x_{t'}) > \pi(x_t)\}.$$

Because π respects monotonicity, $A_\pi \subseteq P \backslash M$. Moreover, $P \backslash A_\pi$ is acyclical since π represents it. Observe that if $(x_t, x_{t'}) \notin P$, then $\lambda(B_t \cap x_{t'}^{\geq}) = 0$. Hence,

$$C(\pi) = \sum_{t=1}^{T} \lambda\Big(B_t \cap \Big(\bigcup_{t' \mid (x_t, x_{t'}) \in A_\pi} x_{t'}^{\geq}\Big)\Big).$$

Thus every feasible ranking π generates a feasible edge-removal set A_π with the same value. It follows that

$$\min_{U \subseteq P \setminus M} \sum_{t=1}^{T} \lambda\Big(B_t \cap \Big(\bigcup_{t' | (x_t, x_{t'}) \in U} x_{t'}^{\gtrsim} \Big)\Big) \text{ s.t. } P \setminus U \text{ is acyclical } \leq \min_{\pi \in \Pi^D} C(\pi) = I(D).$$

Conversely, take any $A \subseteq P \setminus M$ such that $P \setminus A$ is acyclic. Thus there exists a ranking π_A such that if $(x_t, x_{t'}) \in P \setminus A$ then $\pi_A(x_t) > \pi_A(x_{t'})$. Since $M \subseteq P \setminus A$, if $x_t \in x_{t'}^{\gtrsim}$ then $\pi_A(x_t) > \pi_A(x_{t'})$. Hence, $\pi_A \in \Pi^D$. Now suppose $(x_t, x_t') \in P$ and $\pi_A(x_{t'}) > \pi_A(x_t)$, hence $(x_t, x_t') \notin P \setminus A$. Thus $(x_t, x_t') \in A$ and so $A_{\pi_A} = \{(x_t, x_{t'}) \in P \mid \pi_A(x_{t'}) > \pi_A(x_t)\} \subseteq A$. Hence, we have

$$C(\pi_A) = \sum_{t=1}^{T} \lambda\Big(B_t \cap \Big(\bigcup_{t' | (x_t, x_{t'}) \in A_{\pi_A}} x_{t'}^{\gtrsim} \Big)\Big) \leq \sum_{t=1}^{T} \lambda\Big(B_t \cap \Big(\bigcup_{t' | (x_t, x_{t'}) \in A} x_{t'}^{\gtrsim} \Big)\Big).$$

Thus every feasible edge-removal set A generates a feasible ranking π_A with a weakly lower value. It follows that

$$I(D) = \min_{\pi \in \Pi^D} C(\pi) \leq \min_{U \subseteq P \setminus M} \sum_{t=1}^{T} \lambda\Big(B_t \cap \Big(\bigcup_{t' | (x_t, x_{t'}) \in U} x_{t'}^{\gtrsim} \Big)\Big) \text{ s.t. } P \setminus U \text{ is acyclical.}$$

These inequalities together imply that the optimization problem (5) has the value $I(D)$.

A.7 Proof of Proposition 6

Fix any ranking $\pi \in \Pi^D$. We construct a utility function $u : \mathbb{R}_+^L \to \mathbb{R}$ that respects π among the chosen bundles. We then show that the preference $\succeq$ represented by u satisfies $\succeq \in \mathscr{R}_M$ and generates a swaps distance

$$W(\succeq, D) \leq \sum_{t=1}^{T} \lambda\Bigg(B_t \cap \Bigg(\bigcup_{t' | \pi(x_{t'}) > \pi(x_t)} x_{t'}^{\gtrsim} \Bigg)\Bigg).$$

When π is a solution to problem (4), the resulting $\succeq$ is a swaps preference by Proposition 4.

Define $\tilde{v} : \mathbb{R}_+^L \to \mathbb{R}$ by

$$\tilde{v}(x) = \begin{cases} \pi(x_t) & \text{if } x = x_t \text{ for some } t = 1, \dots, T \\ 0 & \text{otherwise.} \end{cases}$$

Then define $v : \mathbb{R}^L_+ \to \mathbb{R}$ by $v(x) = \max_{y \in x^{\leq}} \tilde{v}(y)$. Since $\pi \in \Pi^D$ respects monotonicity, we obtain $v(x_t) = \pi(x_t)$ for all $t = 1, \ldots, T$. Furthermore, $y \in x^{\geq}$ implies $v(y) \geq v(x)$, so v is weakly monotone. Then define $u(x) = v(x) + \sigma(x)$ for $\sigma(x) = 1 - \exp(-\sum_{i=1}^L x^i)$. Since $\sigma(x) \in [0, 1)$, adding it never overturns a strict ranking of v between bundles. In particular, u still respects π among the chosen bundles. Furthermore, u is strictly increasing (hence is measurable and has thin level sets), so that the preference it represents satisfies $\succeq \in \mathscr{R}_M$.

Consider any $t = 1, \ldots, T$ and $y \succ x_t$, which requires $v(y) \geq v(x_t) = \pi(x_t)$. This implies that there must exist $x_{t'}$ with $\pi(x_{t'}) \geq \pi(x_t)$ and $x_{t'} \in y^{\leq}$, or, equivalently, $y \in x_{t'}^{\geq}$. Therefore,

$$y \in \bigcup_{t' | \pi(x_{t'}) \geq \pi(x_t)} x_{t'}^{\geq}.$$

We obtain

$$\begin{aligned} W(\succeq, D) = \sum_{t=1}^{T} \lambda(\{y \in B_t \mid y \succ x_t\}) &\leq \sum_{t=1}^{T} \lambda\Bigg(B_t \cap \Bigg(\bigcup_{t' | \pi(x_{t'}) \geq \pi(x_t)} x_{t'}^{\geq}\Bigg)\Bigg) \\ &= \sum_{t=1}^{T} \lambda\Bigg(B_t \cap \Bigg(\bigcup_{t' | \pi(x_{t'}) > \pi(x_t)} x_{t'}^{\geq}\Bigg)\Bigg), \end{aligned}$$

where the last equality follows because $\pi(x_{t'}) = \pi(x_t)$ requires $t' = t$, and $\lambda(B_t \cap x_t^{\geq}) = 0$.

A.8 Proof of Proposition 7

Step 1: (i) implies (ii).

As shown in the proof of Lemma 2, for any $\pi \in \Pi^D$ we can construct preferences $\succeq \in \mathscr{R}_{CM}$ that coincide with π on the chosen bundles and achieve a swaps distance arbitrarily close to the right-hand side of (9). If $\pi(x_{t^1}) > \pi(x_{t^2})$ for all $\pi \in \Pi^{D*}$, it follows that there exist preferences $\succeq \in \mathscr{R}_{CM}$ with $x_{t^1} \succ x_{t^2}$ that achieve a swaps distance arbitrarily close to $I(D)$, by Proposition 4.

Consider instead any preference $\succeq \in \mathscr{R}_M$ such that $x_{t^2} \succeq x_{t^1}$. If $x_{t^1} \notin x_{t^2}^{\geq}$, then choose any $\pi \in \Pi^D$ which satisfies $\pi(x_{t^2}) > \pi(x_{t^1})$ and additionally $\pi(x_{t'}) > \pi(x_t)$ whenever $x_{t'} \succ x_t$, which exists due to monotonicity of $\succeq$. Lemma 3 shows that the swaps distance is at least the right-hand side of (9) for the chosen π. Since $\pi \notin \Pi^{D*}$ and this set is finite, it follows that the swaps distance of all these preferences is strictly bounded away from $I(D)$.

If $x_{t^1} \in x_{t^2}^{\geq}$ and therefore $\pi(x_{t^1}) > \pi(x_{t^2})$ for all $\pi \in \Pi^D$, we must have $x_{t^1} \notin x_{t^2}^{\gg}$ and $x_{t^2} \sim x_{t^1}$ by monotonicity of $\succeq$. Choose any $\pi \in \Pi^D$ which satisfies $\pi(x_{t'}) > \pi(x_t)$ whenever

$x_{t'} \succ x_t$ and additionally that there does not exist t with $\pi(x_{t^1}) > \pi(x_t) > \pi(x_{t^2})$ (if there exists at least one t' with $x_{t'} \sim x_{t^2} \sim x_{t^1}$ and $x_{t'} \in x_{t^2}^{\geq}$ and $x_{t^1} \in x_{t'}^{\geq}$, we can replace t^1 by the appropriate t' in the following argument, which guarantees the existence of such a ranking π). Construct $\pi' \in \Pi^D_{\gg}$ from π by only switching the values of x_{t^1} and x_{t^2}. Since there does not exist t with $\pi(x_{t^1}) > \pi(x_t) > \pi(x_{t^2})$, it still holds that $x_{t'} \succ x_t$ implies $\pi'(x_{t'}) > \pi'(x_t)$. By Lemma 3,

$$\sum_{t=1}^{T} \lambda(\{y \in B_t \mid y \succ x_t\}) \geq \sum_{t=1}^{T} \lambda\bigg(B_t \cap \bigg(\bigcup_{t' \mid \pi'(x_{t'}) > \pi'(x_t)} x_{t'}^{\geq}\bigg)\bigg).$$

By the definition of π', for all $t \notin \{t^1, t^2\}$ we have

$$\lambda\bigg(B_t \cap \bigg(\bigcup_{t' \mid \pi'(x_{t'}) > \pi'(x_t)} x_{t'}^{\geq}\bigg)\bigg) = \lambda\bigg(B_t \cap \bigg(\bigcup_{t' \mid \pi(x_{t'}) > \pi(x_t)} x_{t'}^{\geq}\bigg)\bigg).$$

Additionally, it holds that $\lambda(B_{t^2} \cap x_{t^1}^{\geq}) = 0$ since $x_{t^1} \in x_{t^2}^{\geq}$, and therefore

$$\begin{aligned}
\lambda\bigg(B_{t^2} \cap \bigg(\bigcup_{t' \mid \pi(x_{t'}) > \pi(x_{t^2})} x_{t'}^{\geq}\bigg)\bigg) &= \lambda\bigg(B_{t^2} \cap \bigg(\bigcup_{t' \mid \pi'(x_{t'}) > \pi'(x_{t^2})} x_{t'}^{\geq} \cup x_{t^1}^{\geq}\bigg)\bigg) \\
&= \lambda\bigg(B_{t^2} \cap \bigg(\bigcup_{t' \mid \pi'(x_{t'}) > \pi'(x_{t^2})} x_{t'}^{\geq}\bigg)\bigg).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\lambda\bigg(B_{t^1} \cap \bigg(\bigcup_{t' \mid \pi'(x_{t'}) > \pi'(x_{t^1})} x_{t'}^{\geq}\bigg)\bigg) &= \lambda\bigg(B_{t^1} \cap \bigg(\bigcup_{t' \mid \pi(x_{t'}) > \pi(x_{t^1})} x_{t'}^{\geq} \cup x_{t^2}^{\geq}\bigg)\bigg) \\
&> \lambda\bigg(B_{t^1} \cap \bigg(\bigcup_{t' \mid \pi(x_{t'}) > \pi(x_{t^1})} x_{t'}^{\geq}\bigg)\bigg),
\end{aligned}$$

where the inequality follows from the observation that

$$\begin{aligned}
&\lambda\bigg(\Big(B_{t^1} \cap x_{t^2}^{\geq}\Big) \setminus \bigg(B_{t^1} \cap \bigg(\bigcup_{t' \mid \pi(x_{t'}) > \pi(x_{t^1})} x_{t'}^{\geq}\bigg)\bigg)\bigg) \\
&\quad = \lambda\bigg(B_{t^1} \cap \bigg(x_{t^2}^{\geq} \setminus \bigcup_{t' \mid \pi(x_{t'}) > \pi(x_{t^1})} x_{t'}^{\geq}\bigg)\bigg) > 0
\end{aligned}$$

since $x_{t^1} \in x_{t^2}^{\geq}$, $\pi \in \Pi^D$, and the function g_{t^1} defining B_{t^1} is strictly increasing. Thus

$$\sum_{t=1}^{T} \lambda(\{y \in B_t \mid y \succ x_t\}) \geq \sum_{t=1}^{T} \lambda\Bigg(B_t \cap \Bigg(\bigcup_{t' \mid \pi'(x_{t'}) > \pi'(x_t)} x_{t'}^{\geq}\Bigg)\Bigg)$$
$$> \min_{\pi \in \Pi^D} \sum_{t=1}^{T} \lambda\Bigg(B_t \cap \Bigg(\bigcup_{t' \mid \pi(x_{t'}) > \pi(x_t)} x_{t'}^{\geq}\Bigg)\Bigg),$$

which again shows that the swaps distance is strictly bounded away from $I(D)$.

Taken together, sufficiently close to $I(D)$, all preferences must satisfy $x_{t^1} \succ x_{t^2}$, which implies $(x_{t^1}, x_{t^2}) \in \succeq^*$ and $(x_{t^2}, x_{t^1}) \notin \succeq^*$, hence $x_{t^1} \succ^* x_{t^2}$.

Step 2: (ii) implies (i) if either $\mathscr{R}^* = \mathscr{R}_M$ or a swaps preference does not exist.

We will show that a violation of condition (i) implies that (ii) is also not true. This is immediate when the violation is due to $\pi(x_t) = \pi(x_{t'})$ because $t = t'$. It is also immediate when $\pi(x_t) < \pi(x_{t'})$ for all $\pi \in \Pi^{D*}$, because in that case $x_{t'} \succ^* x_t$ by the first part of the proof, and therefore $x_t \not\succ^* x_{t'}$ by asymmetry of $\succ^*$. Hence consider the remaining case where there exist $\pi, \tilde{\pi} \in \Pi^{D*}$ with $\pi(x_t) > \pi(x_{t'})$ and $\tilde{\pi}(x_t) < \tilde{\pi}(x_{t'})$.

Suppose first that a swaps preference does not exist. Let $\succeq^0 \in \mathscr{R}^*$ be arbitrary. Since a swaps preference does not exist, $I(D) < W(\succeq^0, D)$. As shown in the proof of Proposition 4, we can therefore construct a preference $\succeq \in \mathscr{R}_{CM} \subseteq \mathscr{R}^*$ that agrees with π and a preference $\tilde{\succeq} \in \mathscr{R}_{CM} \subseteq \mathscr{R}^*$ that agrees with $\tilde{\pi}$ and which satisfy $W(\succeq, D), W(\tilde{\succeq}, D) < W(\succeq^0, D)$. Hence $(x_t, x_{t'}) \notin \succeq^*$ and $(x_{t'}, x_t) \notin \succeq^*$, so there is no robust swaps preference.

Suppose next that $\mathscr{R}^* = \mathscr{R}_M$. As shown in the proof of Proposition 6, there exists a swaps preference $\succeq \in \mathscr{R}^*$ that agrees with π and a swaps preference $\tilde{\succeq} \in \mathscr{R}^*$ that agrees with $\tilde{\pi}$. Again, $(x_t, x_{t'}) \notin \succeq^*$ and $(x_{t'}, x_t) \notin \succeq^*$, so there is no robust swaps preference.

A.9 Proof of Proposition 8

Suppose that $\pi(x_t) > \pi(x_{t'})$ for all $\pi \in \Pi^{D*}$ but, by contradiction, $(x_t, x_{t'}) \notin \text{trans}(P \setminus U)$ for some $U \in \mathscr{U}^*$. Denote $\hat{P} = \text{trans}(P \setminus U)$ and define $\tilde{P} = \hat{P} \cup \{(x_{t'}, x_t)\}$. We first show that $\tilde{P}$ is acyclical. Since $P \setminus U$ is acyclical by definition of $\mathscr{U}^*$, it follows that $\hat{P}$ is acyclical. Thus, if $\tilde{P}$ were not acyclical, there would have to be a cycle of the form $x_{t'} \tilde{P} x_t \hat{P} x_{t_1} \hat{P} \dots \hat{P} x_{t_n} \hat{P} x_{t'}$, which implies $(x_t, x_{t'}) \in \hat{P}$ by the definition of transitive closure, a contradiction. Hence, since $\tilde{P}$ is acyclical and $M \subseteq \tilde{P}$, there exists a ranking $\tilde{\pi} \in \Pi^D$ that represents $\tilde{P}$, i.e., $(x_{t_1}, x_{t_2}) \in \tilde{P}$ implies $\tilde{\pi}(x_{t_1}) > \tilde{\pi}(x_{t_2})$. It follows that $\tilde{\pi}$ also represents $P \setminus U$. As in the proof of Proposition 5, define $A_{\tilde{\pi}} = \{(x_{t_1}, x_{t_2}) \in P \mid \tilde{\pi}(x_{t_2}) > \tilde{\pi}(x_{t_1})\} \subseteq U$, so that

$$\begin{aligned} C(\tilde{\pi}) &= \sum_{t=1}^{T} \lambda\Big(B_t \cap \Big(\bigcup_{t' \mid \tilde{\pi}(x_{t'}) > \tilde{\pi}(x_t)} x_{t'}^{\gtrsim}\Big)\Big) = \sum_{t=1}^{T} \lambda\Big(B_t \cap \Big(\bigcup_{t' \mid (x_t, x_{t'}) \in A_{\tilde{\pi}}} x_{t'}^{\gtrsim}\Big)\Big) \\ &\leq \sum_{t=1}^{T} \lambda\Big(B_t \cap \Big(\bigcup_{t' \mid (x_t, x_{t'}) \in U} x_{t'}^{\gtrsim}\Big)\Big) = I(D). \end{aligned}$$

By Proposition 4, this implies $\tilde{\pi} \in \Pi^{D*}$, which is a contradiction because $\tilde{\pi}(x_{t'}) > \tilde{\pi}(x_t)$.

Suppose next that $(x_t, x_{t'}) \in \mathrm{trans}(P \setminus U)$ for all $U \in \mathscr{U}^*$ but, by contradiction, there exists $\pi \in \Pi^{D*}$ with $\pi(x_{t'}) > \pi(x_t)$. As shown in the proof of Proposition 5, the set $A_\pi = \{(x_t, x_{t'}) \in P \mid \pi(x_{t'}) > \pi(x_t)\}$ is a feasible edge-removal set which induces the same value of the objective in (5) as π induces in (4), so that $A_\pi \in \mathscr{U}^*$ because both problems achieve the same minimal value. Denoting $\hat{P} = P \setminus A_\pi$, it therefore holds that $(x_t, x_{t'}) \in \mathrm{trans}(\hat{P})$ by our assumption, but $(x_t, x_{t'}) \notin \hat{P}$. Hence there must be a sequence $x_t \hat{P} x_{t_1} \hat{P} \ldots \hat{P} x_{t_n} \hat{P} x_{t'}$. Since π represents $\hat{P}$, this implies $\pi(x_t) > \pi(x_{t_1}) > \ldots > \pi(x_{t_n}) > \pi(x_{t'})$, contradicting that $\pi(x_{t'}) > \pi(x_t)$.