Is Complexity the Problem?
Testing Random Choice with Heterogeneity[*]

Shuhua Si[†]

November 8, 2025

See the most recent version here.

Abstract

Economic choices are often stochastic: the same person may make a different choice when facing the same alternatives repeatedly. Standard models assume that the degree of randomness reflects the size of utility differences, but choice inconsistencies could also reflect difficulty comparing alternatives. Recent studies estimate such *comparison difficulty* (or "complexity") by fitting choice models on data collected from different subjects under a representative agent assumption. However, pooling such data while assuming homogeneous preference could violate standard models of random choice simply because of heterogeneity across subjects, even in the absence of variation in comparison difficulty. This paper develops a revealed preference framework, *collective rationalizability*, that tests for variation in comparison difficulty while incorporating heterogeneity. The framework characterizes whether violations of standard models can be explained by comparison difficulty alone, heterogeneity alone, or require both. I then provide a statistical test for collective rationalizability and apply the method to two existing experiments. In both cases, heterogeneity alone explains observed failures of stochastic transitivity well, demonstrating that comparison difficulty can be not only theoretically but also empirically confused with heterogeneity in aggregate data.

[†]Department of Economics, Columbia University. Email: ss5580@columbia.edu.

1 Introduction

Economic choices are often stochastic: the same individual may make different choices when facing the same problem repeatedly. The degree of randomness is conventionally explained by the location of alternatives on a utility scale: each option is assigned a position on a one-dimensional utility scale, and the probability of choosing one option over another depends only on the difference between their utilities. This postulate, satisfied by widely-used discrete choice models such as logit and probit, is known as *Simple Scalability* (Krantz, 1964). Yet not all random choice reflects small utility gaps. Consider choosing between health insurance plans with different monthly costs, provider networks, and covered services, or comparing job offers that differ in salary, location, and work-life balance. Even when one option is actually better than the other, the comparison itself might be inherently difficult: the stochastic choice need not reflect that the alternatives have similar values but rather that they are difficult to compare. This distinction matters for welfare, as the loss from choice mistakes could be much larger than Simple Scalability implies. As a recent body of empirical work shows convincingly, accounting for various cognitive and information frictions has important welfare implications in markets featuring complex choices (Barahona et al., 2023; Brown and Jeon, 2024; Vatter, 2025). Therefore, a central question is: How can we separately detect the difficulty of comparing alternatives from the alternatives' utility differences?

Recent studies estimate comparison difficulty by specifying functional forms and fitting them to the choice data under a representative agent assumption. The choice data consist of the proportion of total choices favoring option A over B collected from possibly heterogeneous subjects.[1] Researchers rely on such data because it is rarely feasible to obtain rich individual-level choice frequencies, where each person is observed making a large number of repeated decisions over all pairs. The problem, however, is that aggregate data might mask *individual heterogeneity*. A 60% choice share for option A over B could mean that *every* person chooses A with probability 0.6, or that 60% of individuals always choose A while the remaining always choose B, with no individual randomness, just different tastes. When aggregate data violate Simple Scalability, is it because some options are harder to compare, or because people have heterogeneous preferences?

The main contribution of this paper is a revealed preference framework, *collective rationalizability*, for testing random choice models while allowing for heterogeneity. I derive the collectively rationalizable version of Simple Scalability and two other models incorporating the difficulty of comparing alternatives (*comparison difficulty*). The characterizations identify whether violations of Simple Scalability can be explained by heterogeneity alone, comparison difficulty alone, or the two together. This framework does not assume any specific functional forms for individual preference or comparison difficulty. Building on the theoretical foundation, I then present a statistical test for collective rationalizability. Applications to data from two existing experiments (Prashnani et al., 2018; Clithero, 2018) demonstrate the practical applicability of the method. In both cases,

---

[1]See, for example, Natenzon (2019); He and Natenzon (2023); Shubatt and Yang (2024).

heterogeneity alone explains the observed deviations from Simple Scalability well.

This paper focuses on three nested classes of models proposed in the literature: Simple Scalability (Krantz, 1964), Moderate Utility (Halff, 1976), and Weak Utility (Luce and Suppes, 1965). While Simple Scalability assumes all variations in choice probabilities reflect utility differences, Moderate and Weak Utility models introduce variable comparison difficulty, with Weak Utility imposing fewer restrictions on how it varies across pairs. To disentangle preference differences from particular difficulty of comparison, existing revealed preference approaches check for *stochastic transitivity* (Luce and Suppes, 1965; Tversky and Russo, 1969; He and Natenzon, 2024): if A is chosen more frequently than B (denoted $A \succcurlyeq B$), and $B \succcurlyeq C$, then specific requirements must hold for choices between A and C. Under Simple Scalability, a larger utility difference translates into a more consistent choice, and in this example since A and C must be further apart than any intermediate pair, the choice frequency for A over C must exceed that of either of the other two pairs. But with comparison difficulty, the frequency of choice also depends on how hard A and C are to compare, and thus only weaker forms of stochastic transitivity must hold.

Existing revealed preference approaches yield theoretical tests for the presence of comparison difficulty. However, implementing them in practice is challenging,[2] because they require an *ideal dataset*: exact choice probabilities for each individual over all binary pairs. With the data typically available to researchers, these approaches face two obstacles. First, violations of Simple Scalability after aggregation might reflect *heterogeneity in preferences* rather than comparison difficulty. Second, it is important to account for *statistical uncertainty*, that whether deviations reflect genuine violations or sampling variation. This paper addresses both challenges by developing testable characterizations of collective rationalizability that explicitly allow for heterogeneity and provide uniformly valid statistical inference.

To clarify these challenges, the first one is *preference heterogeneity*: aggregate patterns may well deviate from Simple Scalability even in the absence of varying comparison difficulty across pairs. An example comes from binary choices of snack foods in Clithero (2018)'s experiment. The aggregate choice frequency of choosing KitKat over M&M's is 0.84, of choosing M&M's over Fruit Snacks is 0.58, and of choosing KitKat over Fruit Snacks is 0.61. These frequencies violate Simple Scalability. Since KitKat is preferred to M&M's and M&M's is preferred to Fruit Snacks, KitKat must be strongly preferred to Fruit Snacks: transitivity would require the frequency for KitKat over Fruit Snacks to exceed 0.84, yet we observe only 0.61. This could be taken as a sign of comparison difficulty, that comparing KitKat with Fruit Snacks is more difficult than comparing KitKat with M&M's. However, such violations can also arise purely from preference heterogeneity without any comparison difficulty. For instance, imagine that around half of the population always choose chocolate snacks while the other half always choose the fruit one, and among the two chocolate candies, most prefer KitKat. Each individual is fully rational, chooses deterministically, and satisfies

---

[2]See, for example, Coombs and Pruitt (1961); Griswold and Luce (1962), and a review by Rieskamp et al. (2006).

Simple Scalability, yet the aggregate frequencies appear to violate it.[3]

To address this, I develop a novel characterization of *collective rationalizability* for the three nested models of random choice described above (Theorem 1). The key insight is that aggregate choice patterns can be collectively rationalized by a model if and only if they can be expressed as a weighted average of individual choice rules that satisfy the model. The characterizations reveal when violations of Simple Scalability can be accommodated by heterogeneity and/or comparison difficulty.

To characterize collective rationalizability, one should consider all possible population compositions: the number of types, the preferences of each type, and their population fractions. The core insight is that collective rationalizability is about *convex combinations*: a choice rule can be collectively rationalized if and only if it belongs to the convex hull of the individual rationalizable set. Nonparametric conditions for collective rationalizability follow from a two-step strategy. First, I analytically characterize the set of individually rationalizable choice probabilities. Second, I derive collective rationalizability conditions by taking the convex hull of this individual rationalizable set. This approach generalizes to any random choice model once its individual rationalizable set is characterized.[4]

I offer two equivalent characterizations of collectively rationalizable versions of random choice models. First, building on techniques for modeling disjunctions (Balas, 1971), the model can be expressed as a system of linear inequalities where membership can be readily checked using convex programming. This technique applies broadly: given any individually rationalizable set, the same method produces testable linear conditions for its collectively rationalizable version. Second, I provide an alternative characterization by identifying a finite list of candidates for extreme points, with the convex hull comprising all convex combinations of these points. The extreme points offer geometric insights and more transparent intuitions, and allow to check how different models nest within each other. Both characterizations lend themselves readily to statistical testing.

In particular, the characterizations reveal a difference between comparison difficulty and heterogeneity. While some aggregate patterns can be rationalized by either explanation, the two are not equivalent: Weak Utility (with comparison difficulty but no heterogeneity) and *Collective* Simple Scalability (with heterogeneity but no comparison difficulty) are not nested within each other. Suppose, for example, that choice probabilities between A over B, B over C, and A over C are, in order, $(0.9, 0.9, 0.6)$. A single individual under Weak Utility can rationalize this with preferences $A \succ B \succ C$ but high difficulty comparing A and C. However, Collective Simple Scalability cannot: the combination of strong preferences (A vs B, B vs C) with near-indifference (A vs C) cannot

---

[3]Precisely, one such distribution of heterogeneous preferences that generates these aggregate frequencies is:

  23%: FruitSnacks $\succ$ KitKat $\succ$ M&M's   16% FruitSnacks $\succ$ M&M's $\succ$ KitKat

  58%: KitKat $\succ$ M&M's $\succ$ FruitSnacks   3%: KitKat $\succ$ FruitSnacks $\succ$ M&M's

[4]See, for example, Appendix A.5. I discuss collectively rationalizable Simple Scalability under multinomial choice and how it can rationalize behavioral regularities such as choice reversals.

arise from aggregating any heterogeneous population where individuals satisfy Simple Scalability and choice probabilities reflect only utility differences. As a second example, consider the choice probabilities $\left(\frac{2}{3}, \frac{2}{3}, \frac{1}{3}\right)$. This cycle ($A \succ B \succ C$ but $C \succ A$) violates transitivity and cannot be represented by Weak Utility. However, Collective Simple Scalability can rationalize it through heterogeneity: three equal-sized groups with preferences $A \succ B \succ C$, $B \succ C \succ A$, and $C \succ A \succ B$, each choosing deterministically, generate exactly this pattern.[5]

Beyond heterogeneity, the second challenge is accounting for *statistical variation*: observed aggregate patterns include sampling variation, making it unclear whether deviations from collective rationalizability reflect genuine departures or noise. Based on Theorem 1's characterization of which choice patterns are collectively rationalizable, Theorem 2 presents a uniformly valid test evaluating whether observed deviations can be attributed to sampling error, accommodating various data generating processes and random assignment schemes between subjects and choice problems.

One challenge for the statistical testing procedure is that the test statistic, defined as the distance from observed choice frequencies to the collectively rationalizable set, has a limit distribution that does not depend continuously on the null hypothesis. This occurs because the collectively rationalizable set is a polytope whose boundary includes edges and vertices where the test statistic exhibits discontinuities. Standard bootstrap fails in such cases (Andrews, 2000). As a solution, I adopt the numerical delta method from Hong and Li (2018), which guarantees uniform validity, that the test maintains its nominal size regardless of where the true parameter lies, under a broad range of data generating processes.

When either heterogeneity or comparison difficulty can explain violations of Simple Scalability in aggregate data, does the model become so flexible that it loses meaningful empirical content? Following Bronars (1987), Selten (1991), and Beatty and Crawford (2011), I compute the relative size of the predicted region for each model up to $n = 5$ options. I find that *Collective* Simple Scalability (allowing only heterogeneity) is always more restrictive than *Representative-agent* Weak Utility (allowing only comparison difficulty). This theoretical restrictiveness can be combined with each model's empirical fit to compare model performance along two dimensions: the frequency of successfully rationalizing *real* data and the frequency of rejecting *random* data. The following empirical applications demonstrate this idea.

Finally, I illustrate the practical applicability of collective rationalizability using data from two existing experiments: Clithero (2018), on binary choices among snack foods, and Prashnani et al. (2018) on perceptual judgments of image similarity. In both applications, aggregate choices violate Simple Scalability. Without accounting for heterogeneity, researchers might interpret these violations as evidence of comparison difficulty. However, the collective rationalizability framework reveals that heterogeneity alone explains the observed patterns better than comparison difficulty

---

[5]This example comes from the Condorcet paradox of social choice theory (Condorcet, 1785), which shows that transitive individual preferences can produce cycles in aggregate preference over binary pairs.

does in both ways: it rationalizes more of the actual experimental data while rejecting more randomly generated data. These applications demonstrate that heterogeneity can be confused with comparison difficulty, both theoretically and empirically, highlighting the importance of the collective rationalizability framework.

This paper provides additional results on identifying the ranking of comparison difficulty: if comparison difficulty indeed varies across pairs, can we go beyond and characterize *how* it varies? I provide necessary and sufficient conditions for rationalizability of individual choice probabilities under a given *ordering* of comparison difficulty. By applying the same technique from Balas (1971), I derive the collectively rationalizable version, characterizing whether observed aggregate patterns can be rationalized by a heterogeneous population where each individual shares a given ranking of comparison difficulty across pairs with $n = 3$ options. This characterization can help identify if everyone finds some specific two alternatives are hard to compare, even though people might disagree on which one they prefer. These difficult comparisons are precisely where people are most likely to make mistakes and choose options that don't reflect their true preferences.

Related Literature.   This paper contributes to the growing literature on cognitive foundations of economic decision-making, which examines what makes decisions difficult and how people respond to such difficulty (see Oprea (2024a) for a review). Recent work has identified various sources of difficulty (Iyengar and Lepper, 2000; Dean et al., 2022; de Lara and Dean, 2024; Shubatt and Yang, 2024; Puri, 2025) and studied behavioral responses (Woodford, 2020; Arrieta and Nielsen, 2024; Gabaix, 2025). A key challenge in this literature is separating cognitive difficulty from preferences. Existing approaches include directly eliciting cognitive uncertainty in the lab (Enke and Graeber, 2023), anchoring preferences through experimental design (Oprea, 2024b; Enke and Shubatt, 2023), and manipulating decision environments in the lab (Bernheim and Sprenger, 2020; Kirchler et al., 2017; Gerhardt et al., 2016; Porcelli and Delgado, 2009).

This paper focuses specifically on comparison difficulty: the cognitive challenge of comparing two alternatives. Previous researchers have used various terms, such as "similarity" (Tversky and Russo, 1969), "comparability" (Halff, 1976), and "complexity" (Shubatt and Yang, 2024) to discuss the same underlying phenomenon. I deliberately use "comparison difficulty" to distinguish the concept from other uses of similarity or complexity in the behavioral economics literature, and emphasize the specific type of cognitive difficulty arising from pairwise comparisons.

Within the comparison difficulty literature, one approach relies on objective benchmarks: if preferences are known, then the frequency of choosing the lower-value option indicates comparison difficulty (Tversky and Russo, 1969). Enke and Shubatt (2023) apply the idea to lottery choices by asking which lottery has higher expected value rather than which lottery subjects choose. Controlling for the expected value differences, the error rate is a sign of difficult comparisons. This paper contributes a revealed preference approach that embody comparison difficulty while incorpo-

rating heterogeneity. The framework remains agnostic about individuals' preferences and imposes no functional form assumptions on either preferences or comparison difficulty. This generality enables analysis in domains such as snack foods or images, where natural metrics are absent and experimental designs requiring objective benchmarks are difficult to implement.

Existing revealed preference approaches characterize comparison difficulty by strong, moderate, and weak stochastic transitivity (Luce and Suppes, 1965; Tversky and Russo, 1969; He and Natenzon, 2024). Recent works in random choice (Natenzon, 2019; He and Natenzon, 2023; Shubatt and Yang, 2024) model such difficulty of comparing alternatives with specific functional forms and distance-based complexity metrics. My approach addresses a key challenge faced by researchers testing stochastic transitivity or estimating models functional forms: the lack of exact choice probabilities for each individual and each pair. I modify the existing revealed preference framework to a collective rationalizability version for aggregate data. A critical advantage of this modification is that it explicitly allows for heterogeneity. Although researchers recognize that heterogeneity across subjects can theoretically confound comparison difficulty across options, this paper characterizes which aggregate patterns violating strong stochastic transitivity can be explained by comparison difficulty alone, by heterogeneity alone, or require both. I demonstrate the empirical importance of this distinction using two existing experiments.

More broadly, this paper belongs to the literature on revealed preferences in stochastic choice (Block and Marschak, 1959; Falmagne, 1978; McFadden, 1981, 2005; McFadden and Richter, 1990; Fishburn, 1992, 1998). They ask the question: are the distributions of choices observed from a population of individuals consistent with rational choice theory, i.e., *random utility maximization* (RUM). I ask the question: are the distributions of choices observed from a population of individuals consistent with different random choice models, with or without varying comparison difficulty across pairs. In this paper, response distributions may arise from taste heterogeneity, stochastic elements in individual choices, or both. Kitamura and Stoye (2018) develops a statistical test of random utility models for rational demand system (i.e. SARP/GARP). My paper develops a similar test using the recent econometrics results by Fang and Santos (2019); Hong and Li (2018).

The remainder of this paper is organized as follows. Section 2 describes the three nested classes of model and develops characterizations for collective rationalizability. Section 3 develops the statistical testing for collective rationalizability, including the statistical setup, a testing procedure that is uniformly valid, and a Monte Carlo investigation. Section 4 applies the framework to two existing experimental datasets. Section 5 concludes. Details and some additional results are reported in the Appendix.

2 Random Choice and Collective Rationalizability

This section develops a revealed preference framework for testing random choice models while allowing for heterogeneity. The analysis focuses on three nested classes of random choice models that progressively relax restrictions on how choice probabilities vary across different pairs of alternatives. The benchmark model, Simple Scalability, assumes that all variation in choice probabilities reflects only the relative locations of the two alternatives on a one-dimensional utility scale. Moderate Utility and Weak Utility progressively relax this assumption by allowing comparison difficulty to vary across different pairs. I characterize *collective rationalizability* for the three model classes: choice patterns are collectively rationalizable if they can be expressed as a weighted average of individual choices satisfying the corresponding model.

The section proceeds as follows. Section 2.1 introduces the model setup and existing revealed preference framework, stochastic transitivity. Section 2.2 characterizes collective rationalizability for the three model classes. Section 2.3 connects collective rationalizability to random utility models. Section 2.4 discusses identification of the *ranking* of comparison difficulty across pairs.

2.1 Three Nested Classes of Models and Stochastic Transitivity

Let Z be a *finite* set of choice options. A *binary menu* is a subset of Z with cardinality of 2. Let $\mathcal{A}$ denote the set of binary menus for which choice probabilities have been observed. A *binary stochastic choice rule* $\rho : \mathcal{A} \to [0,1]$ assigns to each menu (x,y) the probability $\rho(x,y)$ that x is chosen over y, with $\rho(x,y) + \rho(y,x) = 1$ for all $(x,y) \in \mathcal{A}$. In the main analysis, I assume $\mathcal{A} = Z^2$. The Appendix A.7 extends the stochastic transitivity to stochastic acyclicity when observations are incomplete, i.e., $\mathcal{A} \subsetneq Z^2$.

Definition 1 (**Random Choice Models**).

(i). *A Simple Scalability (SS) representation has the form* (Krantz, 1964)

$$\rho(x,y) = F(u(x), u(y)), \quad \forall x, y \in Z \tag{1}$$

for some utility function $u : Z \to \mathbb{R}$ and a function $F : \mathbb{R} \times \mathbb{R} \to (0,1)$ that is strictly increasing in the first argument and strictly decreasing in the second, satisfying $F(a,b) = 1 - F(b,a), \forall a, b \in \mathbb{R}$.[6]

(ii). *A Moderate Utility (MU) representation has the form* (Halff, 1976)

$$\rho(x,y) = F\left(\frac{u(x) - u(y)}{c(x,y)}\right), \quad \forall x, y \in Z \tag{2}$$

---

[6]In SS representation, it is assumed that all choice probabilities are neither 0 or 1. As discussed in McFadden (1974) and Fudenberg et al. (2015), the positivity assumption is empirically indistinguishable from allowing for the boundary values.

for some utility function $u : Z \to \mathbb{R}$, strictly increasing function $F : \mathbb{R} \to [0,1]$ satisfying $F(x) = 1 - F(-x)$, and distance metric[7] $c : Z^2 \to \mathbb{R}^+$.

(iii). A Weak Utility (WU) representation has the form (Luce and Suppes, 1965)

$$\rho(x,y) = F\left(\frac{u(x) - u(y)}{c(x,y)}\right), \quad \forall x,y \in Z \tag{3}$$

for some utility function $u : Z \to \mathbb{R}$, strictly increasing function $F : \mathbb{R} \to [0,1]$, satisfying $F(x) = 1 - F(-x)$, and semimetric[8] $c : Z^2 \to \mathbb{R}^+$.

Remark 1. Under Simple Scalability, Moderate Utility, or Weak Utility, the condition $\rho(x,y) \geq 1/2$ holds if and only if $u(x) \geq u(y)$. Define the preference relation $x \succcurlyeq y$ if and only if $\rho(x,y) \geq 1/2$.

The key postulate of Simple Scalability is that for a given individual, each alternative is assigned a location on a one-dimensional utility scale, and the degree to which choice probabilities vary between any two alternatives depends only on their relative locations. The term "Simple Scalability" reflects this reliance on scalar utility representations.

Many familiar stochastic choice rules are special cases of Simple Scalability with additional structures. For example, the logit model $\rho(x,y) = \frac{e^{u(x)/\lambda}}{e^{u(x)/\lambda} + e^{u(y)/\lambda}}$, and the probit model $\rho(x,y) = \Phi_\sigma(u(x) - u(y))$.[9] More generally, additive random utility models with i.i.d. utility shocks

$$\rho(x,y) = \Pr[u(x) + \varepsilon_x \geq u(y) + \varepsilon_y], \quad \text{where } \varepsilon_x, \varepsilon_y \text{ are i.i.d. noise,} \tag{4}$$

are examples of Simple Scalability, where stochasticity can stem from uncertainty or noisy reading regarding the true utilities. Alternatively, following Fudenberg et al. (2015), stochasticity in Simple Scalability can also reflect costly implementation of desired choices. Their additive perturbed utility representation, restricted here to binary menus,

$$\rho(x,y) = \underset{p \in (0,1)}{\arg\max} \ pu(x) + (1-p)u(y) - c(p), \tag{5}$$

where c is a convex perturbation function capturing costs from implementing more precise decisions, also yields an example of Simple Scalability.[10]

---

[7]A distance metric is a function $c : Z^2 \to \mathbb{R}^+$ satisfying (i) $c(x,y) = 0$ if and only if $x = y$; and (ii) symmetry $c(x,y) = c(y,x)$; and (iii) triangle inequality $c(x,z) \leq c(x,y) + c(y,z), \forall(x,y), (y,z), (x,z) \in \mathcal{A}$.

[8]A semimetric is a function $c : Z^2 \to \mathbb{R}^+$ satisfying (i) $c(x,y) = 0$ if and only if $x = y$; and (ii) symmetry $c(x,y) = c(y,x)$.

[9]These models have the additional structure that choice probabilities depend only on utility *differences*, corresponding to the Fechner model (Fechner, 1860) $\rho(x,y) = F(u(x) - u(y))$ for some utility function $u : Z \to \mathbb{R}$ and a function $F : \mathbb{R} \to (0,1)$ that is strictly increasing. For logit, F is the cumulative distribution function (CDF) of a Gumbel distribution with variance $1/\lambda$; for probit, F is the CDF of normal distribution with standard deviation σ. Φ_σ denotes the cumulative distribution function of the normal distribution normal distribution $N(0, \sigma)$.

[10]While Fudenberg et al. (2015) develop their model for general menu sizes, I only present the binary case.

The distinguishing feature of Moderate Utility and Weak Utility representations, compared to Simple Scalability, is the presence of $c(x, y)$ in the denominator. This term captures the difficulty of discriminating between the two options x and y, which we define as the *comparison difficulty* of the binary menu $\{x, y\}$. Crucially, while Simple Scalability assumes that differences in choice randomness reflect only the relative locations of alternatives on a one-dimensional utility scale, the comparison difficulty $c(x, y)$ can depend on any attributes of the alternatives. This allows Moderate Utility and Weak Utility to capture how non-utility features affect stochastic choice.

For instance, consider a newly-hired assistant professor buying a laptop. Even though evaluating the value of each laptop is not straightforward for her, choosing between two laptops that differ only in color should be relatively straightforward with low randomness in choice. However, choosing between laptops from different brands, with different operating systems, processing speeds, and storage capacities, becomes considerably more difficult, generating greater randomness in choice. This pair-specific noise in the comparison process constitutes what we call comparison difficulty, a dimension distinct from utility differences.

One interpretation of this stochasticity comes from Natenzon (2019), who models the choice process as Bayesian learning under imperfect information about alternatives. In his framework, the correlation between noise terms depends on the degree of "similarity" between options. When alternatives are easier to compare, the decision-maker can extract more precise signals about their true utility differences, leading to less random choices. Conversely, difficult comparisons, such as those requiring tradeoffs across multiple incommensurable attributes, produce noisier signal extraction and more randomness in choice.

The difference between Moderate Utility and Weak Utility lies in whether comparison difficulty $c(x, y)$ must satisfy the triangle inequality. Under Moderate Utility, if comparisons of $\{x, y\}$, $\{y, z\}$ are both easy, then x and z cannot be arbitrarily difficult to compare. Their comparison difficulty is bounded by the sum of the other two. This creates a coherent structure where comparison difficulties relate to each other systematically, like distances in physical space. Many existing models that derive comparison difficulty from the distance of attributes or correlation of the noise distribution belong to Moderate Utility (McFadden, 1978; Natenzon, 2019; He and Natenzon, 2023; Shubatt and Yang, 2024). Weak Utility relaxes this requirement, allowing comparison difficulties to violate the triangle inequality. It permits more complex patterns where the difficulty of comparing x and z need not be constrained by their respective difficulties with an intermediate option y. Testing whether data satisfies MU versus WU reveals whether comparison difficulty follows a distance-based structure or exhibits more complex, non-distance metric patterns.

It has been well established in the literature that Simple Scalability, Moderate Utility, and Weak Utility representations are equivalent to strong, moderate, and weak stochastic transitivity, respectively, when all binary pairs are observed.

Definition 2 (Stochastic Transitivity).

A choice rule ρ satisfies (Luce and Suppes, 1965; Tversky and Russo, 1969; He and Natenzon, 2024)

(i). Strong stochastic transitivity if $\forall x, y, z \in Z$

$$\min\{\rho(x,y), \rho(y,z)\} \geq (>)1/2 \quad \text{implies} \quad \rho(x,z) \geq (>) \max\{\rho(x,y), \rho(y,z)\}. \tag{6}$$

(ii). Moderate stochastic transitivity if $\forall x, y, z \in Z$

$$\min\{\rho(x,y), \rho(y,z)\} \geq 1/2 \quad \text{implies} \quad \begin{cases} \rho(x,z) > \min\{\rho(x,y), \rho(y,z)\} \\ \qquad\qquad \text{or} \\ \rho(x,z) = \rho(x,y) = \rho(y,z) \end{cases}. \tag{7}$$

(iii). Weak stochastic transitivity if $\forall x, y, z \in Z$

$$\min\{\rho(x,y), \rho(y,z)\} \geq 1/2 \quad \text{implies} \quad \rho(x,z) \geq 1/2. \tag{8}$$

The equivalence arises because stochastic transitivity conditions impose restrictions on how choice probabilities can vary across different triples of alternatives. Strong stochastic transitivity requires the most restrictive pattern: if $x \succcurlyeq y$ and $y \succcurlyeq z$, then x must be chosen over z with probability at least $\max\{\rho(x,y), \rho(y,z)\}$. This constraint is precisely what Simple Scalability implies when all variation in choice probabilities reflects only the relative locations on a one-dimensional scale: the distance between x and z must be at least as large as that of the intermediate pairs. Moderate and weak stochastic transitivity progressively relax this restriction by weakening the bound on $\rho(x,z)$. In particular, weak stochastic transitivity is exactly the transitivity of the preference relation: if $x \succcurlyeq y$ and $y \succcurlyeq z$, then $x \succcurlyeq z$.

2.2 Collective Rationalizability

The preceding analysis characterizes when choices are rationalizable by Simple Scalability, Moderate Utility, or Weak Utility *without* heterogeneity. In practice, however, researchers often work with data collected from possibly heterogeneous subjects but have to assume representative agent. The classic Condorcet paradox (Condorcet, 1785) illustrates this issue. Consider three individuals with different preference orderings: $x \succ y \succ z$, $y \succ z \succ x$, and $z \succ x \succ y$. If each individual chooses deterministically according to their preferences, the representative-agent choice probabilities are $\rho(x,y) = \frac{2}{3}, \rho(y,z) = \frac{2}{3}$, and $\rho(x,z) = \frac{1}{3}$, which violates even weak stochastic transitivity, despite each individual satisfying Simple Scalability.

To account for heterogeneity, I develop a framework of *collective rationalizability* that characterizes when stochastic choices can be represented as arising from a heterogeneous population of rational

individuals. I begin with the definition of collective rationalizability, then provide the characterization in Section 2.2.1, and present an illustrated example with three options in Section 2.2.2.

To formalize the idea, I distinguish between aggregate and individual choices. Let ρ now denote the aggregate stochastic choice rule, representing the population-level probability that x is chosen over y. The question of collective rationalizability asks: when can we find some N individuals, each characterized by a stochastic choice rule ρ_i satisfying a given model (Simple Scalability, Moderate Utility, or Weak Utility), such that the aggregate choice probabilities ρ can be expressed as a weighted average of these individual choice rules?

Definition 3 (**Collective Rationalizability**).

A Collective Simple Scalability representation has the form

$$\rho(x,y) = \sum_{i=1}^{N} \lambda_i \rho_i(x,y), \quad \forall x,y \in Z, \tag{9}$$

for some positive integer N, weights $\lambda_i > 0$ that sum to 1, and each ρ_i admits Simple Scalability.

Collective Moderate Utility and Collective Weak Utility representations are defined analogously.

Crucially, this definition allows us to determine whether observed violations of Simple Scalability in aggregate data can be explained by heterogeneity alone, comparison difficulty alone, or whether both are required. If aggregate choices cannot be *collectively* rationalized by Simple Scalability, it implies that no population composition *exists* that can generate the observed choice probabilities such that every individual finds all pairwise comparisons equally easy.

Remark 2. In the definition above, choices are collectively rationalizable if there exists some positive integer N representing the number of "types" of individuals, where each type admits the corresponding random choice model. I remain agnostic about the number of types. A natural question is whether the N required for rationalization could exceed the actual sample size. As the definition shows, collective rationalizability is about convex combination of individual choice rules. Let $m = \binom{n}{2}$ denote the number of binary pairs for n options. By Carathéodory's Theorem, as long as the sample size is at least $m + 1$, the number of types needed will not exceed the sample size. For instance, with $n = 6$ options, this requires a sample size of at least 16. This is typically satisfiable in practice.

Example 1. Consider observed population choice over three alternatives $\{x_1, x_2, x_3\}$ is

$$(\rho(x_1, x_2), \rho(x_2, x_3), \rho(x_1, x_3)) = (0.65, 0.65, 0.1).$$

If treated as a representative agent, it cannot even be rationalized by Weak Utility since it violates weak stochastic acyclicity: $\rho(x_1, x_2) > 0.5$, $\rho(x_2, x_3) > 0.5$ but $\rho(x_1, x_3) < 0.5$. However, it can be rationalized by Collective Weak Utility. Consider two subjects with individual stochastic choice rules as

$$\text{Subject } 1 - (\rho(x_1, x_2), \rho(x_2, x_3), \rho(x_1, x_3)) = (0.9, 0.4, 0.1),$$
$$\text{Subject } 2 - (\rho(x_1, x_2), \rho(x_2, x_3), \rho(x_1, x_3)) = (0.4, 0.9, 0.1).$$

It is easy to verify that aggregating subjects 1 and 2's choice rates with equal weights $\lambda_1 = \lambda_2 = 0.5$ gives us the aggregated choice probabilities above. Both of them satisfy weak stochastic transitivity but with different tastes: subject 1 has preference ordering $x_3 \succ x_1 \succ x_2$, while subject 2 has $x_2 \succ x_3 \succ x_1$.

Can the above example be rationalized by Collective Simple Scalability? Try constructing individual choice rules that satisfy Simple Scalability and aggregate to $(0.65, 0.65, 0.1)$, and you will find it impossible. To determine this systematically requires a characterization of collective rationalizability. More generally, Definition 3 raises a practical challenge: how can we test whether given aggregate data are collectively rationalizable without exhaustively searching over all possible combinations of individual types? The following section addresses this question.

2.2.1 Characterization of Collective Rationalizability

Given a finite set of n alternatives $Z = \{x_i\}_{i=1}^n$ and $m = \binom{n}{2}$ possible pairs, a stochastic choice rule can be identified with a vector $\boldsymbol{\rho} \in [0, 1]^m$, where entries of this *choice vector* are $\rho(x_i, x_j)$ with some arbitrary but henceforth fixed ordering.[11] The characterization of collective rationalizability is thus finding a subset of the whole space $[0, 1]^m$ that consists of all rationalizable choice vectors.

The core insight underlying the approach is that collective rationalizability concerns convex combinations: a choice rule is collectively rationalizable if and only if it can be expressed as a convex combination of individual choice rules satisfying the corresponding model. This suggests a two-step characterization strategy. First, I characterize the sets of choice vectors that are rationalizable by Simple Scalability, Moderate Utility, or Weak Utility, respectively. Second, I take the convex hull of each set to obtain the collective rationalizability characterization. This transforms the problem into checking membership in a well-defined convex set. Importantly, this approach generalizes beyond the three models studied here: as long as a characterization without heterogeneity is available, the techniques in Step 2 apply broadly to derive the collectively rationalizable version of the corresponding model.

---

[11]Throughout this paper, I assume $\rho(x_i, x_j)$ for $1 \le i < j \le n$, arranged in the lexicographical order. For example, with $n = 4$ options, a stochastic choice rule becomes the vector

$$\boldsymbol{\rho} = [\rho(x_1, x_2), \rho(x_1, x_3), \rho(x_1, x_4), \rho(x_2, x_3), \rho(x_2, x_4), \rho(x_3, x_4)]^T \in [0, 1]^6.$$

Step 1: Characterize rationalizable set without heterogeneity. Stochastic transitivity conditions characterize rationalizability without heterogeneity, but they are defined implicitly through constraints on all triplets. To apply the convex hull approach for collective rationalizability, I derive an explicit analytical characterization of the individual rationalizable set as a subset of $[0, 1]^m$. More precisely, I characterize this set as a union of *convex polytopes*, where each polytope corresponds to a specific preference ordering (or, for Simple Scalability and Moderate Utility, specific orderings of choice probabilities under a preference ordering), providing both *half-space-* and *vertex*-representations for each polytope.

Note that weak stochastic transitivity is the minimal constraint shared by all three models. To understand the geometry of the choice space, consider how we can partition the entire space $[0, 1]^m$ based on the preference relation $\succcurlyeq$ defined by $x \succcurlyeq y$ iff $\rho(x, y) \geq 1/2$. Weak stochastic transitivity is precisely the requirement that this binary relation is transitive. The entire choice space $[0, 1]^m$ can be partitioned into 2^m small hypercubes based on whether each $\rho(\cdot, \cdot) \geq \frac{1}{2}$. Since there are $n!$ different linear orders, there are $n!$ out of 2^m hypercubes that can be rationalized by Weak Utility. Each of these hypercubes is a convex polytope with explicit half-space constraints defined with the help of the following definition.

Definition 4 (Deterministic Choice Rule).

A choice vector $\boldsymbol{\rho}^* \in [0, 1]^m$ is a rational deterministic choice rule if all entries are in $\{0, 1\}$ and correspond to a linear order. Specifically, there exists a permutation σ over $\{1, 2, \ldots, n\}$ such that

$$\rho^*(x_i, x_j) = \begin{cases} 1, & \text{if } \sigma(i) < \sigma(j) \\ 0, & \text{if } \sigma(i) > \sigma(j) \end{cases}, \qquad \forall 1 \leq i < j \leq n.$$

The permutation σ induces a linear order where $x_i \succ x_j$ if and only if $\sigma(i) < \sigma(j)$.

Define the set of all rational deterministic choice rules $\boldsymbol{\rho}^*$ as R_{det}. Then $\boldsymbol{\rho}$ satisfies weak stochastic transitivity if and only if

$$\boldsymbol{\rho} \in \text{Cube}(\boldsymbol{\rho}^*) := \left\{ \boldsymbol{\rho} : |\rho(x_i, x_j) - \rho^*(x_i, x_j)| \leq \frac{1}{2}, \ \forall x_i, x_j \in Z \right\} \quad \text{for some } \boldsymbol{\rho}^* \in R_{\text{det}}. \quad (10)$$

The inequality constraints $\text{Cube}(\boldsymbol{\rho}^*)$ ensures that $\boldsymbol{\rho}$ inherits the same transitive ordering as $\boldsymbol{\rho}^*$.[12]

---

[12]The union of the sets $\text{Cube}(\boldsymbol{\rho}^*)$ characterizes the *closure* of the rationalizable set of WU. The closure is necessitated by complications at the indifference boundary $\rho(x, y) = 1/2$. For instance, $\boldsymbol{\rho} = [0.49, 0.49, 0.7]^T$ satisfies WU while $\boldsymbol{\rho} = [0.5, 0.5, 0.7]^T$ does not. Since the convex hull of finite unions of closed convex sets equals the closure of the convex hull, boundary effects remain measure-zero for collective rationalizability. Hence, throughout the paper, I work with closures of all rationalizable sets (WU, MU, SS) without loss for empirical applications.

Hence, Weak Utility rationalizability can be represented as the following subset:

$$R^{\mathrm{WU}} = \bigcup_{k=1}^{n!} \mathrm{Cube}(\boldsymbol{\rho}_k^*) := \bigcup_{k=1}^{n!} \left\{ \boldsymbol{\rho} : \boldsymbol{A}_k^{\mathrm{WU}} \boldsymbol{\rho} \leq \boldsymbol{b}_k^{\mathrm{WU}} \right\}. \tag{11}$$

For Simple Scalability, strong stochastic transitivity imposes additional constraints within each of the $n!$ hypercubes satisfying weak transitivity. Specifically, if $x \succcurlyeq y$ and $y \succcurlyeq z$, then $\rho(x,z) \geq \rho(x,y)$ and $\rho(x,z) \geq \rho(y,z)$. These constraints, inequalities of the form $\rho(x_i, x_j) \geq \rho(x_{i'}, x_{j'})$, slice away portions of each hypercube and leave a smaller convex polytope, and Simple Scalability representation is the union of these $n!$ convex polytopes.

Moderate Utility presents the most intricate geometry. Within each of the $n!$ hypercubes satisfying weak transitivity, moderate stochastic transitivity requires: if $x \succcurlyeq y$ and $y \succcurlyeq z$, then $\rho(x,z) \geq \rho(x,y)$ or $\rho(x,z) \geq \rho(y,z)$. The disjunctive nature of the constraints creates sets that are generally non-convex. To handle this, I divide each hypercube $\mathrm{Cube}(\boldsymbol{\rho}^*)$ into simplexes based on how the choice probabilities rank against each other. Each hypercube splits into $m!$ simplexes, and moderate stochastic transitivity determines which survive. Let L be the number of surviving simplexes per hypercube $\mathrm{Cube}(\boldsymbol{\rho}^*)$, which is equal for different hypercubes due to symmetry, and the Moderate Utility rationalizable set becomes a union of $n! \cdot L$ simplexes.

Therefore, the Simple Scalability and Moderate Utility representations can be written as

$$R^{\mathrm{SS}} = \bigcup_{k=1}^{n!} \left\{ \boldsymbol{\rho} : \boldsymbol{A}_k^{\mathrm{SS}} \boldsymbol{\rho} \leq \boldsymbol{b}_k^{\mathrm{SS}} \right\}, \quad R^{\mathrm{MU}} = \bigcup_{k=1}^{n!} \bigcup_{l=1}^{L} \left\{ \boldsymbol{\rho} : \boldsymbol{A}_{k,\ell}^{\mathrm{MU}} \boldsymbol{\rho} \leq \boldsymbol{b}_{k,\ell}^{\mathrm{MU}} \right\}, \tag{12}$$

where $\boldsymbol{A}$ and $\boldsymbol{b}$ represent inequalities encoding both the hypercube boundaries and the strong/moderate stochastic transitivity inequalities, with exact forms in Appendix A.1. Rationalizable set for Weak Utility, Simple Scalability, and Moderate Utility can be characterized as a union of *finitely* many convex polytopes, where the visualization under $n = 3$ is shown in Figure 1a, 1c, and 1e.

In addition to the half-space-representation, these polytopes' vertices are characterized as follows.

Lemma 1 (Vertices of Individual Rationalizability).

Every vertex of polytopes comprising R^{SS}, R^{MU}, or R^{WU} has entries taking values only in $\{0, \frac{1}{2}, 1\}$.

Therefore, vertices can be obtained by checking strong, moderate, or weak stochastic transitivity of $\boldsymbol{\rho} \in \{0, \frac{1}{2}, 1\}^m$. Let P_{SS}, P_{MU}, and P_{WU} denote the set of extreme points respectively. For an individual with a *given* preference ordering, the extreme points $\{0, \frac{1}{2}, 1\}$ emerge naturally from how we think about choice difficulty. When comparisons are easy, choices become deterministic (0 or 1 depending on the preference). When comparisons are maximally difficult, choices become essentially

random (approaching $\frac{1}{2}$). Any observed choice probability can be expressed as a combination of these extremes, and these values represent the fundamental boundaries of rationalizable choice under transitivity constraints with some given preference ordering.

Step 2: Characterize collective rationalizability with heterogeneity. Collective rationalizability is then obtained by taking the convex hull of the rationalizable set without heterogeneity.

Theorem 1 (**Characterization of Collective Rationalizability**)**.**

For a choice vector $\boldsymbol{\rho} \in [0,1]^m$, the following are equivalent:

(i). $\boldsymbol{\rho}$ is collectively rationalizable by Simple Scalability.

(ii). $\boldsymbol{\rho} \in \mathrm{conv}(P_{\mathrm{SS}})$.

(iii). The following linear system over $(\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_{n!}, \boldsymbol{x})$ is feasible:

$$\boldsymbol{\rho} = \sum_{k=1}^{n!} \boldsymbol{\rho}_k, \quad \boldsymbol{x} \in \Delta^{n!-1}, \quad \mathbf{0} \le \boldsymbol{\rho}_k \le x_k \mathbf{1}, \quad \boldsymbol{A}_k^{\mathrm{SS}} \boldsymbol{\rho}_k \le x_k \boldsymbol{b}_k^{\mathrm{SS}}, \qquad \forall k = 1, \ldots, n! \qquad (13)$$

Collective Moderate Utility and Collective Weak Utility are characterized analogously.

For the vertex-representation of collective rationalizability (i) $\iff$ (ii), I use the fact that:

$$\mathrm{conv}\left(\bigcup_{i=1}^{k} \mathrm{conv}(P_i)\right) = \mathrm{conv}\left(\bigcup_{i=1}^{k} P_i\right), \quad \forall \text{ finite collection of sets } P_i \subseteq \mathbb{R}^m. \qquad (14)$$

This means the collective rationalizability set equals the convex hull of all extreme points obtained from the each of the polytopes comprising R^{SS}, R^{MU}, or R^{WU}.

For the half-space-representation of collective rationalizability (i) $\iff$ (iii), the convex hull of a finite union of convex polytopes can be readily obtained using techniques from modeling disjunctions (Balas, 1971), by first using indicator variables $\boldsymbol{x}$ to represent the finite union, then relaxing the integrality constraints to obtain the convex hull.

Theorem 1 provides both vertex and half-space representations of the collectively rationalizable sets. This dual characterization offers geometric insights into the structure of collectively rationalizable choice and the nesting relationships among Simple Scalability, Moderate Utility, and Weak Utility, with or without heterogeneity, as well as with other stochastic choice models. It also establishes the theoretical foundation for statistical testing, which will be developed in Section 3.

2.2.2 Example: Collective Rationalizability for Three Options

To illustrate the theoretical results, consider the case of three choice options $Z = \{x_1, x_2, x_3\}$. The choice vector is $\boldsymbol{\rho} = [\rho(x_1, x_2), \ \rho(x_1, x_3), \ \rho(x_2, x_3)]^T \in [0, 1]^3$.

Inequalities of individual rationalizability. With three options, there are $3! = 6$ possible linear orders. For each preference ordering, individual rationalizability constraints create distinct regions. For instance, under $x_1 \succ x_2 \succ x_3$, the corresponding deterministic choice vector is

$$\boldsymbol{\rho}_k^* = [\rho^*(x_1, x_2), \rho^*(x_1, x_3), \rho^*(x_2, x_3)] = [1, 1, 1]^T.$$

(i). Weak Utility: $\boldsymbol{A}_k^{\mathrm{WU}} \boldsymbol{\rho} \leq \boldsymbol{b}_k^{\mathrm{WU}}$ means all choice rates $\rho(x_1, x_2), \rho(x_2, x_3), \rho(x_1, x_3) \in [\frac{1}{2}, 1]$, which are also included in the constraints for Simple Scalability and Moderate Utility.

(ii). Simple Scalability: $\boldsymbol{A}_k^{\mathrm{SS}} \boldsymbol{\rho} \leq \boldsymbol{b}_k^{\mathrm{SS}}$ *additionally* requires $\rho(x_1, x_3) \geq \max\{\rho(x_1, x_2), \rho(x_2, x_3)\}$.

(iii). Moderate Utility: The hypercube is further partitioned into 6 simplexes by the ranking of choice rates, where $L = 4$ of them satisfy moderate stochastic transitivity.

- $\boldsymbol{A}_{k,1}^{\mathrm{MU}} \boldsymbol{\rho} \leq \boldsymbol{b}_{k,1}^{MU}$ *additionally* requires $\rho(x_1, x_3) \geq \rho(x_1, x_2) \geq \rho(x_2, x_3)$;

- $\boldsymbol{A}_{k,2}^{\mathrm{MU}} \boldsymbol{\rho} \leq \boldsymbol{b}_{k,2}^{MU}$ *additionally* requires $\rho(x_1, x_3) \geq \rho(x_2, x_3) \geq \rho(x_1, x_2)$;

- $\boldsymbol{A}_{k,3}^{\mathrm{MU}} \boldsymbol{\rho} \leq \boldsymbol{b}_{k,3}^{MU}$ *additionally* requires $\rho(x_1, x_2) \geq \rho(x_1, x_3) \geq \rho(x_2, x_3)$;

- $\boldsymbol{A}_{k,4}^{\mathrm{MU}} \boldsymbol{\rho} \leq \boldsymbol{b}_{k,4}^{MU}$ *additionally* requires $\rho(x_2, x_3) \geq \rho(x_1, x_3) \geq \rho(x_1, x_2)$.

Each individual rationalizable set is the union of 6 such regions, with each region representing one preference ordering illustrated by different colors in Figures 1a, 1e, and 1c.

Vertices of individual rationalizability. Within each region, the extreme points have entries in $\{0, \frac{1}{2}, 1\}$. For example, for $x_1 \succ x_2 \succ x_3$, the extreme points are

$$P_{\mathrm{WU}} = \{(1,1,1), (\tfrac{1}{2},1,1), (1,1,\tfrac{1}{2}), (\tfrac{1}{2},1,\tfrac{1}{2}), (\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2}), (\tfrac{1}{2},\tfrac{1}{2},1), (1,\tfrac{1}{2},\tfrac{1}{2}), (1,\tfrac{1}{2},1)\},$$
$$P_{\mathrm{MU}} = \{(1,1,1), (\tfrac{1}{2},1,1), (1,1,\tfrac{1}{2}), (\tfrac{1}{2},1,\tfrac{1}{2}), (\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2}), (\tfrac{1}{2},\tfrac{1}{2},1), (1,\tfrac{1}{2},\tfrac{1}{2})\},$$
$$P_{\mathrm{SS}} = \{(1,1,1), (\tfrac{1}{2},1,1), (1,1,\tfrac{1}{2}), (\tfrac{1}{2},1,\tfrac{1}{2}), (\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2})\}.$$

Collective rationalizability. The next step is to take the convex hull of the individual rationalizable sets, and Figures 1b, 1f, and 1d visualize the result. For $n = 3$, collective Simple Scalability and Moderate Utility coincide, and the facet-defining inequalities can be written as
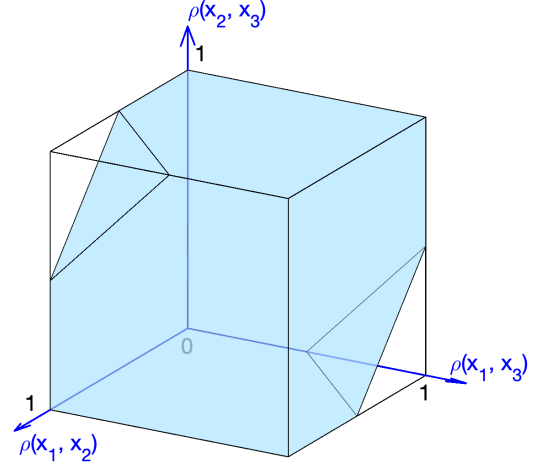
$$\text{Collective SS and MU}: \quad 1 \leq \rho(x_1, x_2) + \rho(x_2, x_3) + \rho(x_3, x_1) \leq 2, \tag{15}$$
$$\text{Collective WU}: \quad \tfrac{1}{2} \leq \rho(x_1, x_2) + \rho(x_2, x_3) + \rho(x_3, x_1) \leq \tfrac{5}{2}, \tag{16}$$
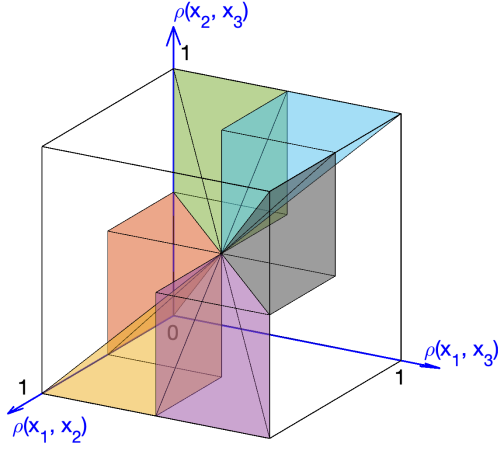
in addition to the trivial boundary inequalities $0 \leq \rho(x_i, x_j) \leq 1, \forall (i, j)$.
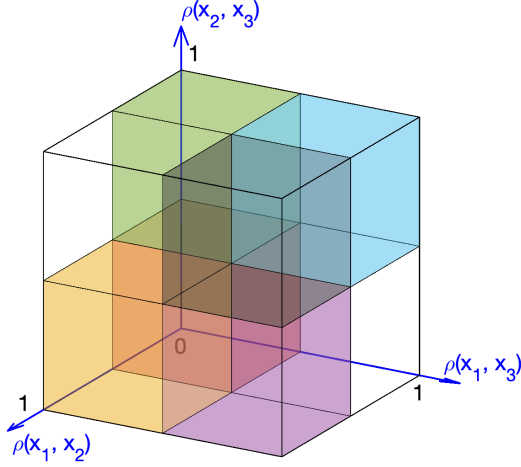
(a) Rationalizable set for Individual WU

(b) Rationalizable set for Collective WU

(c) Rationalizable set for Individual SS
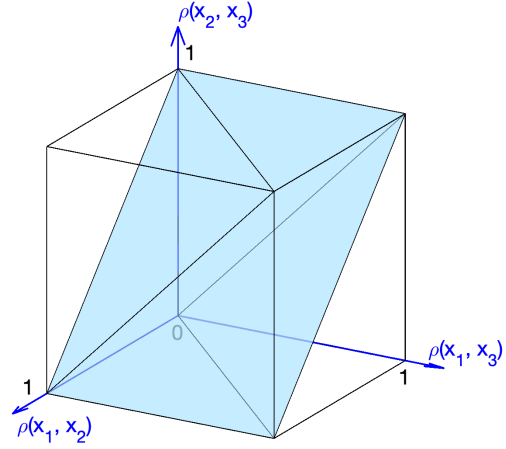
(d) Rationalizable set for Collective SS
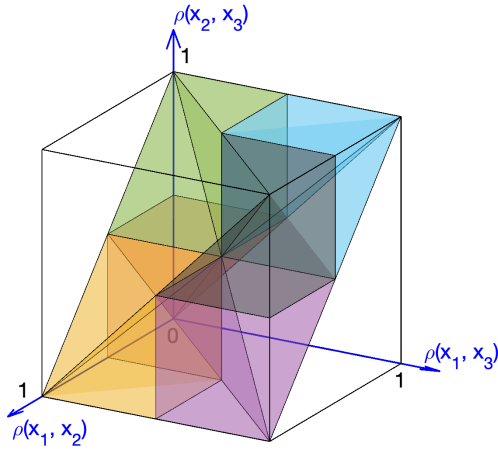
(e) Rationalizable set for Individual MU
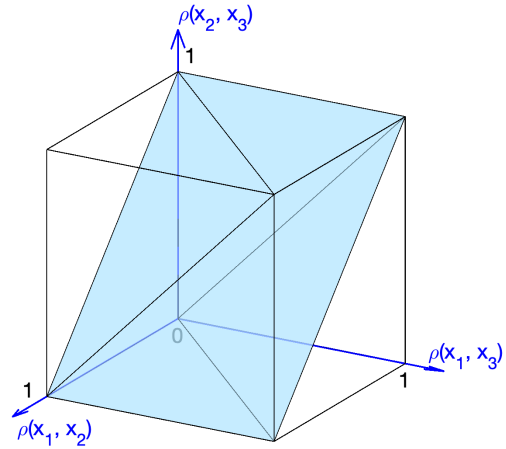
(f) Rationalizable set for Collective MU

Figure 1: Rationalizable sets for Individual and Collective WU/MU/SS models for $n = 3$. In panels (a), (c), (e), different colors represent the six regions corresponding to different preferences. MU and SS impose increasingly restrictive transitivity constraints within each region. Panels (b), (d), (f) show the corresponding collective rationalizability sets, which are the convex hulls of the individual sets. Taking the convex hull connects the previously regions into a single polytope, expanding the set of rationalizable choice patterns while allowing for heterogeneity.

Collective Rationalizability constraints capture that rational heterogeneous populations create a balance: while individuals are transitive, aggregation may produce some "cycling" or inconsistency, but not too much. The feasible intervals $[1, 2]$ or $[\frac{1}{2}, \frac{5}{2}]$ represent the natural amount of intransitivity that emerges from mixing different types, distinguishing patterns explainable by heterogeneity from those requiring additional behavioral assumptions.

Comparing Figures 1a and 1d with Figure 1c reveals two key insights. First, there are significant overlaps where violations of Simple Scalability can be explained by either heterogeneity (Collective Simple Scalability) or comparison difficulty (Weak Utility) alone. However, Collective Simple Scalability and Weak Utility are not nested. Some choice patterns can only be explained by heterogeneity, others can only be explained comparison difficulty, and still others require both to be rationalized. Interestingly under $n = 3$, the volume of the Weak Utility rationalizable set is 0.75, while that of Collective Simple Scalability is 0.67, indicating that the heterogeneity model is actually less permissive than the comparison difficulty model.

Comparing Figures 1d and 1f shows that Collective Simple Scalability coincides with Collective Moderate Utility when $n = 3$, as moderate stochastic transitivity provides no additional degrees of freedom beyond strong transitivity for three alternatives when heterogeneity is allowed.

Readers familiar with the literature will recognize that inequalities (15) are known as the "triangle inequalities" in stochastic choice, which are essentially Block-Marschak polynomials for binary choices under $n = 3$ and characterize the random utility models (Block and Marschak, 1959). This observation reveals an intriguing connection between our collective representation and random utility models. I explore the nesting relationships between different collective representations and also random utility representation in detail in the following section.

2.3 Nesting Relationships and Connections to Random Utility

I now compare the testable implications of random choice models, Simple Scalability, Moderate Utility, and Weak Utility, with or without heterogeneity, and relate them to those of random utility models. To formalize this comparison, let Z denote the finite set of alternatives and let Π denote the set of all strict preferences on Z, where each $\pi \in \Pi$ is a bijection $\pi : Z \to \{1, \ldots, |Z|\}$, which is a specific utility representation of the preference.

Definition 5 (**Random Utility**). *A Random Utility (RU) representation has the form*

$$\rho(x, y) = \sum_{\pi \in \Pi} \nu(\pi) \mathbb{1}_{\{\pi(x) \geq \pi(y)\}}, \quad \forall x, y \in Z, \tag{17}$$

where $\nu : \Pi \to \mathbb{R}^+$ is some probability distribution on Π with $\sum_{\pi \in \Pi} \nu(\pi) = 1$.

It is well known that RU and random choice representations are not nested with each other (see discussion in, for example, Rieskamp et al. (2006); Strzalecki (2025)). In fact, weak stochastic transitivity can be violated by RU,[13] while on the other side, one can construct choice vector that is rationalized by Simple Scalability but not RU, since Simple Scalability does not require the choice rates to be represented by some weighted average of strict orders.

However, once heterogeneous taste is taken into account, we can see that RU is nested within Collective Simple Scalability, the most restrictive of all three collective models. From Definition 5, a choice rule $\boldsymbol{\rho}$ is representable by RU if it corresponds to a probability measure ν over the strict preference orderings on Z, i.e., $\boldsymbol{\rho}$ is representable by RU iff $\boldsymbol{\rho} \in \text{conv}(R_{\text{det}})$, where R_{det} is the set of all rational deterministic choices from Definition 4. Since all rational deterministic choices satisfy Collective Simple Scalability, it follows that RU is nested within Collective Simple Scalability.

A more thorough analysis reveals the following relationship between the collective models and RU:

Lemma 2 (**Nesting Relations**).

(i). *RU is nested within Collective SS, with equivalence for $n \leq 6$ and strict nesting for $n \geq 7$.*

(ii). *Collective SS is nested within Collective MU, with equivalence for $n \leq 5$ and strict nesting for $n \geq 6$.*

(iii). *Collective MU is strictly nested within Collective WU for all $n \geq 3$.*

Lemma 2 follows from comparing the extreme points computed in Lemma 1. The nesting relationships can be verified by checking whether each model's extreme points can be expressed as convex combinations of the alternative model's extreme points.

Why equivalence for small n? When the number of alternatives is small, the geometry of choice probabilities is relatively simple, leading to identical constraints across models. As shown in the example for $n = 3$, Collective Simple Scalability/Moderate Utility and RU share the same facet-defining constraints: the triangle inequality (15) and the probability boundaries $[0, 1]$. This pattern persists for $n = 4$ and $n = 5$, where numerical computation reveals that all facet-defining inequalities are "triangle-type", i.e., versions of (15) applied to different triples of alternatives.[14] While these triangle inequalities remain necessary for all n, the models diverge for larger n as additional structural constraints emerge.

The complexity of larger n. For RU, its rationalizable set is a well-studied combinatorial object often referred to as the linear ordering polytope (Grötschel et al., 1985; Katthän, 2011). Beyond triplet inequalities (15), characterization for RU requires increasingly complex constraints,

---

[13] Again, recall the example of Condorcet paradox.

[14] Computed using the Quickhull algorithm (Barber et al., 1996) via `scipy.spatial.ConvexHull` in Python.

such as k-fence, Mobius ladder, and others (Grötschel et al., 1985), where the number of facet-defining inequalities of the linear ordering polytope grows exponentially with n. The complete characterization remains an open problem even for RU. Similarly, while I have established the characterizations of collective rationalizability in Theorem 1, their exact facet-defining inequalities remain to be fully provided (see Lemma 6 in the Appendix for partial results).

Economic interpretation. The equivalence between RU and Collective Simple Scalability for small n provides the following insights in the binary choice problem: behavior interpreted as arising from heterogeneous deterministic agents (RU) is observationally equivalent to behavior from heterogeneous agents who themselves face uncertainty of the alternative's utility scales or the noisy readings of them (Collective Simple Scalability). Furthermore, the equivalence between Collective Simple Scalability and Collective Moderate Utility for $n \leq 5$ implies a practical limitation: with few alternatives, we lack the power to determine whether a population's stochastic behavior arises from distance-based comparison difficulty or not.

Example 2 (RU strictly nested within Collective SS when $n = 7$).

Consider $n = 7$ and choice rates ρ defined as

$$\rho(x_i, x_j) = \begin{cases} 1 - \varepsilon & \text{if } j - i \geq 3 \\ \frac{1}{2} + \varepsilon & \text{if } 1 \leq j - i \leq 2 \end{cases},$$

with ε small enough. It is straightforward to verify that this $\boldsymbol{\rho}$ satisfies (individual) Simple Scalability and thus also collective SS. However, $\boldsymbol{\rho}$ cannot be rationalized by RU. To generate near-equal choice probabilities for nearby pairs, the population must be nearly evenly split between opposite rankings of these pairs. But the majority of the population needs to agree on the ranking of distant pairs, making it impossible to satisfy transitivity.

2.4 Identify Ranking of Comparison Difficulty

Before turning to statistical testing based on collective rationalizability, I present one additional theoretical result that goes beyond testing whether comparison difficulty varies across pairs: if comparison difficulty indeed varies across pairs, can we characterize *how* it varies? Readers primarily interested in collective rationalizability may proceed directly to Section 3.

In the absence of additional assumptions, comparison difficulty cannot be uniquely identified from binary choice data over finite sets in random choice models. Regarding rankings of comparison difficulties, He and Natenzon (2024) establishes a sufficient condition under which certain observed choices can uniquely determine the ordinal ranking of difficulty under Moderate Utility and Weak Utility representations: $\rho(x_1, x_2) > \rho(x_1, x_3) > \rho(x_2, x_3) \geq 1/2$ implies that $c(x_1, x_2) < c(x_1, x_3)$.

In this section, I develop *necessary and sufficient* conditions for rationalizing individual choice data

with a WU representation and a *given ranking of comparison difficulty*. Consider the problem: given observed choices $\boldsymbol{\rho} = (\rho(x_1, x_2), \rho(x_1, x_3), \rho(x_2, x_3))$, when can these be rationalized by WU with a specific difficulty ranking such as $c(x_1, x_2) > c(x_1, x_3) > c(x_2, x_3)$? I solve this for arbitrary numbers of alternatives in a *loop*. Then, I generalize this notion from individual rationalizability to collective rationalizability with shared ranking of comparison difficulty across heterogeneous agents.

I focus on binary choices over loops, namely individual choice rates $\{\rho(x_j, x_{j+1})\}_{j=1}^{m}$ for a set of m choice options $\{x_j\}_{j=1}^{m}$ (indices modulo m). Denote $\rho_j := \rho(x_j, x_{j+1})$ and $c_j := c(x_j, x_{j+1})$ for all $j = 1, \ldots, m$. To formalize the analysis, for each pair $j \in \{1, \ldots, m\}$, I introduce:

Choice rate magnitudes: $\mu_j := \left| \rho_j - \frac{1}{2} \right|$;

Sign of the preferences: $J^+(\boldsymbol{\rho}) := \left\{ j \mid \rho_j > \frac{1}{2} \right\}$, $\quad J^-(\boldsymbol{\rho}) := \left\{ j \mid \rho_j < \frac{1}{2} \right\}$, $\quad J^0(\boldsymbol{\rho}) := \left\{ j \mid \rho_j = \frac{1}{2} \right\}$.

The loop structure is crucial: utility differences must sum to zero throughout any cycle. This constraint interacts with the given ranking of comparison difficulty, because higher comparison difficulty and choice rate magnitudes imply larger absolute utility differences. The key insight is that a given ranking of c_j is compatible with observed choices if it aligns appropriately with the ranking of μ_j across the sets with different preference directions J^+ and J^-. To formalize this alignment of c_j and μ_j, I define the notions of *ranking function* and *ranking-dominance*.

Definition 6 (**Ranking Functions and Ranking-Dominance**).

Let # denote the cardinality of a finite set.

1. *A ranking function of a sequence of values $\{x_1, \ldots, x_m\}$ is a function $\pi : \{1, \ldots, m\} \to \{1, \ldots, m\}$*

$$\pi(j) = \#\{k \in \{1, \ldots, m\} : x_j < x_k\} + 1. \tag{18}$$

2. *Let σ, π be two ranking functions over $\{1, \ldots, m\}$. Given two index subsets $A, B \subseteq \{1, \ldots, m\}$, A ranking-dominates B with respect to (σ, π) if*

$$\#\{j \in A \mid \sigma(j) \le s, \pi(j) \le t\} \ge \#\{j \in B \mid \sigma(j) \le s, \pi(j) \le t\}, \quad \forall s, t \in \{1, \ldots, m\}, \tag{19}$$

with strict inequality for some (s, t).

Intuitively, an index j with ranking function value $\pi(j)$ means that x_j ranks the $\pi(j)$-th largest among values in the sequence. And ranking-dominance means elements in A consistently rank higher than those in B under both rankings, so that they can be paired such that the element in A always dominates. With these two definitions, the characterization in Proposition 1 shows that ranking-dominance provides the key obstruction for rationalizability under a given ranking of

comparison difficulty.

Proposition 1 (Weak Utility with a Given Ranking of Comparison Difficulty).

Consider individual choice rates over a loop of binary pairs $\boldsymbol{\rho} = \{\rho_j\}_{j=1}^m \in [0,1]^m$. Let σ be the ranking function of the choice rate magnitudes $\{\mu_j\}_{j=1}^m$, and let π be a given ranking function of comparison difficulty $\{c_j\}_{j=1}^m$. Then, $\boldsymbol{\rho}$ can be represented by WU with the given ranking π if and only if neither $J^+(\boldsymbol{\rho})$ nor $J^-(\boldsymbol{\rho})$ ranking-dominates each other with respect to (σ, π), with necessity additionally assuming that all choice magnitudes μ_j are distinct and not equal to $\frac{1}{2}$.

The intuition is that over a loop, if J^+ ranking-dominates J^-, then the positive preferences systematically have larger both choice magnitudes and comparison difficulties than negative preferences, and the required zero-sum condition on utility differences cannot be satisfied.

Given the key role of the ranking-dominance property, here we provide further intuition behind this concept, alongside Proposition 1. Consider a setup of $m = 5$ choice options forming a loop, and the following two different cases of ranking of comparison difficulty. We will discuss whether WU representation with the given ranking is possible by examining the conditions for ranking-domination.

Example a. Consider the following ranking of the choice magnitudes μ_j, and a given ranking of comparison difficulty c_j, where each j is the index for a binary pair:

$$\mu_1 > \mu_2 > \mu_3 > \mu_4 > \mu_5, \qquad c_1 > c_3 > c_5 > c_2 > c_4.$$

Thus the ranking functions σ and π are respectively

$$\sigma(1) = 1, \ \sigma(2) = 2, \ \sigma(3) = 3, \ \sigma(4) = 4, \ \sigma(5) = 5,$$
$$\pi(1) = 1, \ \pi(2) = 4, \ \pi(3) = 2, \ \pi(4) = 5, \ \pi(5) = 3.$$

Also assume the index subsets of the signs of the preferences are

$$J^+(\boldsymbol{\rho}) = \left\{j \mid \rho_j > \tfrac{1}{2}\right\} = \{1, 4, 5\}, \ J^-(\boldsymbol{\rho}) = \left\{j \mid \rho_j < \tfrac{1}{2}\right\} = \{2, 3\}.$$

Example b. Consider the following ranking of μ_j, and a given ranking of c_j

$$\mu_1 > \mu_2 > \mu_3 > \mu_4 > \mu_5, \quad c_1 > c_3 > c_2 > c_5 > c_4.$$

Thus the ranking functions σ and π are respectively

$$\sigma(1) = 1, \ \sigma(2) = 2, \ \sigma(3) = 3, \ \sigma(4) = 4, \ \sigma(5) = 5,$$
$$\pi(1) = 1, \ \pi(2) = 3, \ \pi(3) = 2, \ \pi(4) = 5, \ \pi(5) = 4.$$

Also assume the index subsets are

$$J^+(\boldsymbol{\rho}) = \left\{ j \mid \rho_j > \tfrac{1}{2} \right\} = \{1, 3, 4\}, \ \ J^-(\boldsymbol{\rho}) = \left\{ j \mid \rho_j < \tfrac{1}{2} \right\} = \{2, 5\}.$$

For the two examples described above, We illustrate the ranking functions of *Example a* in Figure 2a, and *Example b* in Figure 2b. Each colored cell corresponds to one index j in J^+ or J^-, with its coordinates on the grid being $(\sigma(j), \pi(j))$. Notice that the term $\#\{j \in J^+ \mid \sigma(j) \leq s, \pi(j) \leq t\}$ can be interpreted as the number of blue (or red for J^-) cells on the top-left $s \times t$ submatrix. The inequality condition for ranking-dominance (19) thus implies that for every top-left submatrix, there are at least as many blue cells as there are red cells.
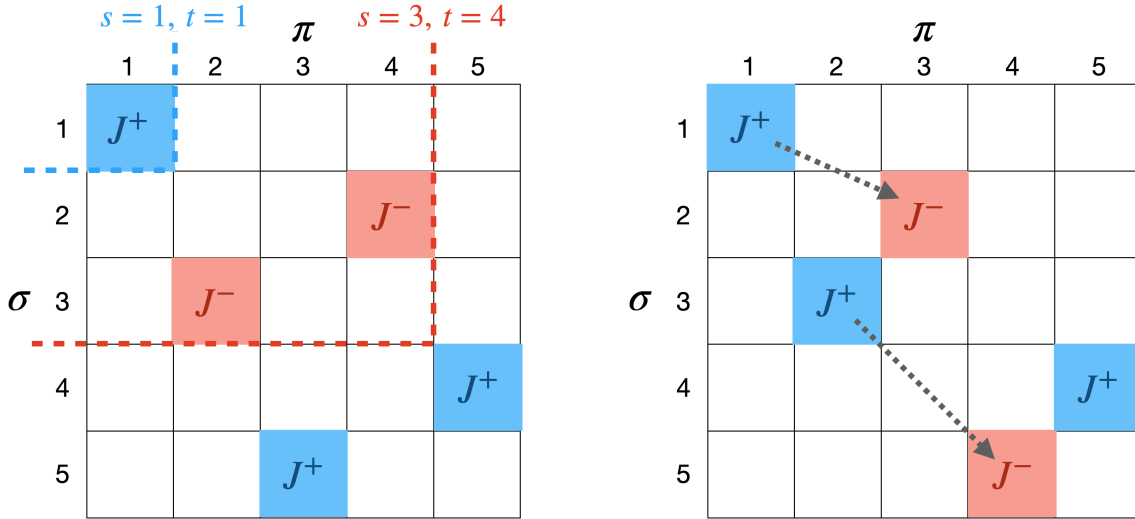
(a) Neither J^+ nor J^- ranking-dominates the other

(b) J^+ ranking-dominates J^-

Figure 2: Illustration of the ranking-domination condition.

Figure 2a shows that in *Example a*, neither J^+ nor J^- ranking-dominates the other. Here the blue dashed lines correspond to a top-left $s \times t$ submatrix with $(s, t) = (1, 1)$, which contains 1 blue and 0 red cells. Meanwhile, the red dashed lines correspond to a top-left $s \times t$ submatrix with $(s, t) = (3, 4)$, which contains 1 blue and 2 red cells. This implies that choosing $(s, t) = (1, 1)$ makes the left-hand-side of (19) strictly greater, while $(s, t) = (3, 4)$ makes the right-hand-side of (19) strictly greater, which concludes that neither A nor B ranking-dominates the other. From Proposition 1, this implies that the choice rates data *can* be represented as WU with this specific ranking of comparison difficulty.

Figure 2b on the other hand, shows that in *Example b*, J^+ ranking-dominates J^-. One can verify by enumerating all possible pairs of (s,t), to see that every top-left $s \times t$ submatrix has at least as many blue cells as red, with strict inequality holding in some cases (for example $(s,t) = (1,1)$). Another way to see ranking-dominance is from the fact that every cell in J^- has a unique corresponding cell in J^+ on its top-left direction, as indicated by the dashed arrows. From Proposition 1, this implies that the choice rates data *cannot* be represented as WU with this specific ranking of comparison difficulty.

2.4.1 Example: Ranking of Comparison Difficulty for Three Options

I now illustrate the idea in the context of three options $\{x_1, x_2, x_3\}$. Note that while our result above is limited to choice rates on a loop instead of full pairing, this distinction vanishes for the case of three options, affording us the complete analysis.[15]

Ranking of comparison difficulty without heterogeneity. Recall the analysis in Proposition 1, I need to check the ranking-dominance between the ranking functions, π for comparison difficulty c_j and σ for choice magnitude μ_j over the index subsets with different signs of preferences J^+ and J^-. Without loss of generality, consider the ranking of comparison difficulty $c(x_1, x_2) > c(x_1, x_3) > c(x_2, x_3)$. Using our simplified notation $c_j := c(x_j, x_{j+1})$, this is $c_1 > c_3 > c_2$, and thus corresponds to the ranking function $\pi(1) = 1, \pi(2) = 3, \pi(3) = 2$. The strategy is to examine each possible preference ordering (determining J^+ and J^-), then derive what constraints the magnitude ranking σ must satisfy to avoid ranking-dominance and thus ensure the compatibility with π.

As an example, consider the case of $J^+ = \{1, 2\}$, $J^- = \{3\}$, which means $\rho(x_1, x_2) > \frac{1}{2}$, $\rho(x_2, x_3) > \frac{1}{2}$, and $\rho(x_3, x_1) < \frac{1}{2}$. Since $\#J^+ = 2 > \#J^- = 1$, J^- cannot ranking-dominate J^+, which can be seen by taking $s = t = 3$ in inequality (19). Hence it remains to choose rankings σ such that J^+ does not ranking-dominate J^-, which requires

$$\#\{j \in J^+ \mid \sigma(j) \le s, \pi(j) \le t\} \ < \ \#\{j \in J^- \mid \sigma(j) \le s, \pi(j) \le t\} \tag{20}$$

holds for some $s, t \in \{1, 2, 3\}$. Since $\#J^- = 1$, (20) holds when the left side equals 0 while the right side equals 1. Given $J^- = \{3\}$ and $\pi(3) = 2$ and $\pi(1) = 1$, this forces either:

- $\sigma(3) = 1 \implies \mu_3 > \max(\mu_1, \mu_2)$, or

- $\sigma(3) = 2, \sigma(1) = 3, \sigma(2) = 1 \implies \mu_2 > \mu_3 > \mu_1$.[16]

Combined together, this simply requires $\mu_3 > \mu_1$, namely $\rho(x_1, x_3) > \rho(x_1, x_2)$.

Intuitively, the case $J^+ = \{1, 2\}$, $J^- = \{3\}$ corresponds to the preference ordering $x_1 \succ x_2 \succ x_3$,

---

[15] For a general number of choice options and full pairing, we can obtain necessary conditions for iterating over all possible loops.

[16] For simplicity, I ignore cases where $\boldsymbol{\rho}$ contains duplicate values, hence σ is a permutation. In the full analysis in Proposition 1, duplicate values of ρ are allowed.

which means $u(x_1) - u(x_3) > u(x_1) - u(x_2)$. Combined with the given requirement $c(x_1, x_3) < c(x_1, x_2)$, rationalization requires

$$\frac{u(x_1) - u(x_3)}{c(x_1, x_3)} > \frac{u(x_1) - u(x_2)}{c(x_1, x_2)} \implies \rho(x_1, x_3) > \rho(x_1, x_2).$$

Applying this analysis to all six preference orderings yields the complete characterization:

- $x_1 \succ x_2 \succ x_3$ or $x_3 \succ x_2 \succ x_1$.
  The additional rationalizability constraint is $\left|\rho(x_1, x_3) - \frac{1}{2}\right| > \left|\rho(x_1, x_2) - \frac{1}{2}\right|$. This yields a rationalizable region comprising half of the small cube corresponding to the preference ordering, depicted in green in Figure 3.

- $x_1 \succ x_3 \succ x_2$ or $x_2 \succ x_3 \succ x_1$.
  No additional constraints are imposed beyond the hypercube boundaries. The entire cube is rationalizable under these orderings, as the pair with maximal utility difference coincides with the pair of maximal comparison difficulty. These regions are depicted in blue.

- $x_3 \succ x_1 \succ x_2$ or $x_2 \succ x_1 \succ x_3$.
  The additional rationalizability constraint is $\left|\rho(x_2, x_3) - \frac{1}{2}\right| > \max\{\left|\rho(x_1, x_2) - \frac{1}{2}\right|, \left|\rho(x_1, x_3) - \frac{1}{2}\right|\}$. This yields the most restrictive rationalizable region, comprising one-third of the cube. These correspond to the case where maximal utility difference coincides with minimal comparison difficulty, depicted in orange.
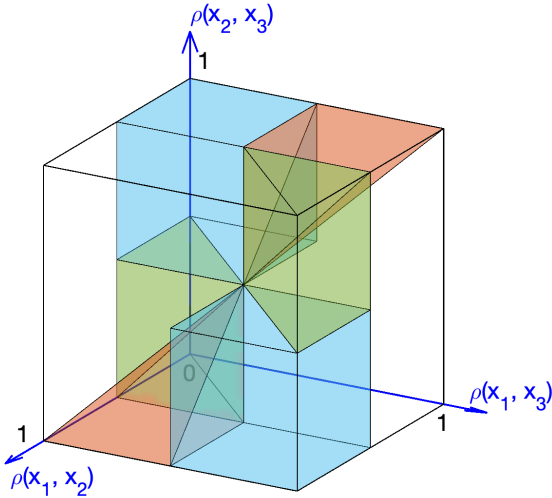

Figure 3: Individual WU-rationalizable with $c(x_1, x_2) > c(x_1, x_3) > c(x_2, x_3)$
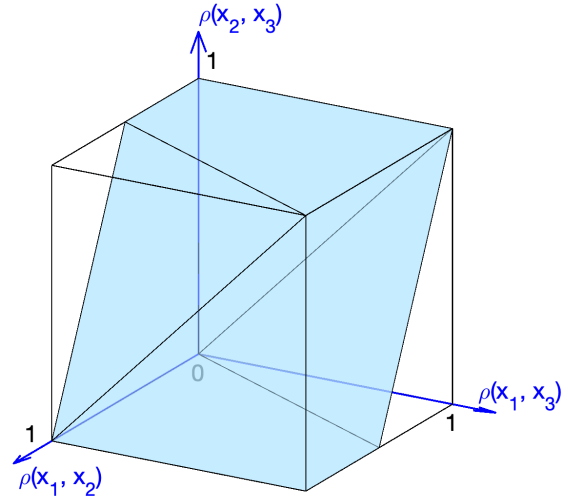
Figure 4: Collective WU-rationalizable with $c_i(x_1, x_2) > c_i(x_1, x_3) > c_i(x_2, x_3), \forall i$

Ranking of comparison difficulty with heterogeneity. From individual to collective rationalizability, I extend the given ranking of comparison difficulty to a heterogeneous population. In

particular, I allow each individual in the population to have possibly different comparison difficulty functions $c_i(x, y)$, as long as all c_i share a common ranking function.

Definition 7 (**Collective Weak Utility with a Common Ranking of Difficulty**).

A Collective WU representation with a common ranking of comparison difficulty has the form

$$\rho(x, y) = \sum_{i=1}^{N} \lambda_i F_i \left(\frac{u_i(x) - u_i(y)}{c_i(x, y)} \right), \quad \forall (x, y) \in Z^2 \tag{21}$$

with utility functions $u_i : Z \to \mathbb{R}$, semimetrics $c_i : Z^2 \to \mathbb{R}^+$, and functions F_i strictly increasing and satisfying $F_i(x) = 1 - F_i(-x)$, weights $\lambda_i \geq 0$ that sum to 1, and c_i follows the same ranking function π for any i.

It naturally follows that the characterization for collective rationalizability with a common ranking of comparison complexity is the convex hull of that of the individual rationalizability. For the previous example, the collective WU-rationalizable set under the commonly shared ranking $c(x_1, x_2) > c(x_1, x_3) > c(x_2, x_3)$ is visualized in Figure 4, with the facet-defining inequalities

$$1 \leq 2\rho(x_1, x_2) + \rho(x_2, x_3) + \rho(x_3, x_1) \leq 3. \tag{22}$$

Recall that for collective WU without ranking requirements, the characterizing condition is $\frac{1}{2} \leq \rho(x_1, x_2) + \rho(x_2, x_3) + \rho(x_3, x_1) \leq \frac{5}{2}$, which is cyclic symmetric and representing the stochastic transitivity while allowing for heterogeneity. Here with the ranking on comparison difficulty, we see an increased weight on the term $\rho(x_1, x_2)$ in the cyclic inequality. Intuitively, this is because $\rho(x_1, x_2)$ is pushed toward $\frac{1}{2}$ due to the large comparison difficulty, and dampens the information it reveals about the utility difference $u(x_1) - u(x_2)$. The increased weight compensates for this dampened information in the rationalizability.

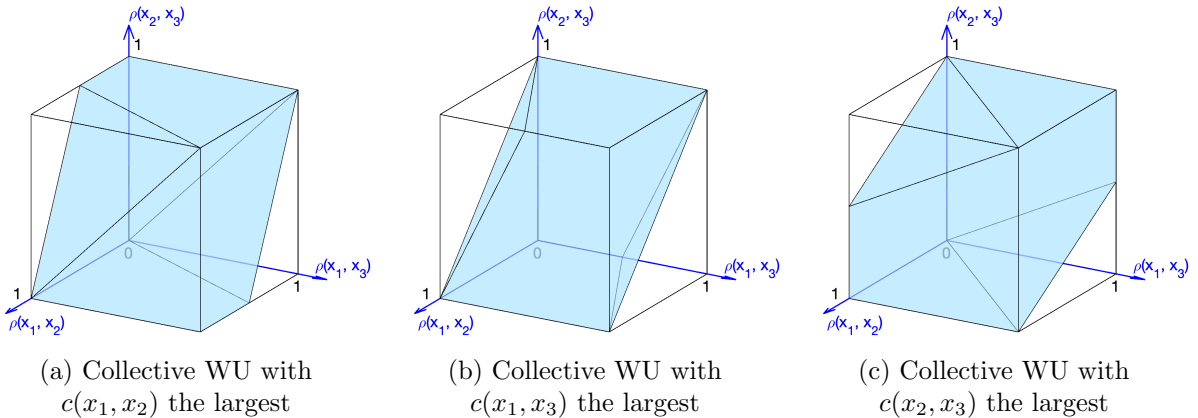| (a) Collective WU with $c(x_1, x_2)$ the largest | (b) Collective WU with $c(x_1, x_3)$ the largest | (c) Collective WU with $c(x_2, x_3)$ the largest |

Figure 5: Collective WU rationalizable with different rankings of comparison difficulty

For all 6 possible rankings of c, we see that the rationalizable set only depends on which comparison is the most difficult, as shown in Figures 5.

Furthermore, by comparing the difference between these rationalizable sets, one can identify when the pair with highest comparison difficulty can be *uniquely* determined from the aggregate choices while allowing for different tastes. I formalize this result in the following Lemma 3, with visualization in Figure 6. While Lemma 3 only states the conditions for $c(x_1, x_2) > \max(c(x_2, x_3), c(x_3, x_1))$, the other two cases can be obtained by simply re-indexing.

Lemma 3 (Collective WU with a Uniquely Identified Common Ranking of Difficulty).

Consider $n = 3$ options $\{x_1, x_2, x_3\}$. Let the triple $\boldsymbol{\rho} = (\rho(x_1, x_2), \rho(x_2, x_3), \rho(x_3, x_1)) \in [0, 1]^3$ denote the population choice probability. Assume everyone shares the ranking of comparison difficulty. $\boldsymbol{\rho}$ is represented by collective WU with the following common ranking of comparison difficulty

$$c_i(x_1, x_2) > \max\{c_i(x_2, x_3),\ c_i(x_3, x_1)\},$$

if and only if either of the following systems of linear inequalities holds:

(A)
$$\begin{cases} 2\rho_{12} + \rho_{23} + \rho_{31} \leq 3 \\ \rho_{12} + 2\rho_{23} + \rho_{31} \geq 3 \\ \rho_{12} + \rho_{23} + 2\rho_{31} \geq 3 \end{cases}$$

(B)
$$\begin{cases} 2\rho_{12} + \rho_{23} + \rho_{31} \geq 1 \\ \rho_{12} + 2\rho_{23} + \rho_{31} \leq 1 \\ \rho_{12} + \rho_{23} + 2\rho_{31} \leq 1 \end{cases}$$
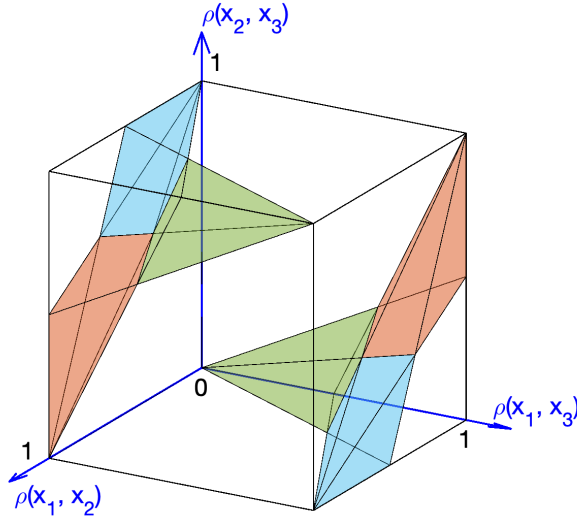

Figure 6: Collective WU representation where the pair with highest comparison difficulty can be determined. Blue: $c(x_1, x_2) > \max\{c(x_2, x_3), c(x_3, x_1)\}$. Green: $c(x_1, x_3) > \max\{c(x_1, x_2), c(x_2, x_3)\}$. Orange: $c(x_2, x_3) > \max\{c(x_1, x_2), c(x_1, x_3)\}$

28

3 Statistical Testing of Collective Rationalizability

The theoretical characterization in Theorem 1 establishes when choice patterns are collectively rationalizable while explicitly allowing for heterogeneity. This section addresses the question of whether estimated choice frequencies may, up to *sampling variation*, have arisen from a collectively rationalizable model.

Testing collective rationalizability involves testing multiple inequalities, a problem extensively studied in econometrics. The highlight of this section is twofold: first, to provide a broad class of d.g.p.s over which the resulting test based on collective rationalizability is uniformly valid, where stochastic transitivity cannot be directly tested; second, to illustrate how to overcome the main challenge that the limiting distribution of the test statistic depends discontinuously on the null hypothesis. I adopt the numerical delta method (Fang and Santos, 2019; Hong and Li, 2018) to provide uniformly valid inference, by using the fact that the target function is directionally differentiable.

The remainder proceeds as follows. Section 3.1 specifies the general class of d.g.p.s. Section 3.2 develops the test statistic and establishes uniform validity. Section 3.3 analyzes finite-sample performance through Monte Carlo simulations.

3.1 Data Generating Process

Let Z be a finite set of n choice options, generating $m := \binom{n}{2}$ binary choice problems. Consider N subjects randomly drawn from a population, where each subject faces a potentially different subset of these problems. For each subject i and binary choice problem j, researchers observe:

(i). An indicator $I_{ij} = \mathbb{1}[\text{subject } i \text{ faces problem } j]$.

(ii). The average choice frequency $\bar{\rho}_{ij}$ when $I_{ij} = 1$.

Note that complete observations are not required. Subjects may face only subsets of possible choices, and repetitions may be insufficient to estimate individual choice probabilities. These limitations prevent testing based on individual rationalizability. Instead, I construct the population choice frequency from the available data:

$$\hat{\boldsymbol{\rho}} = [\hat{\rho}_1, \ldots, \hat{\rho}_m]^T, \quad \text{where} \quad \hat{\rho}_j = \frac{\sum_{i=1}^N \bar{\rho}_{ij} I_{ij}}{\sum_{i=1}^N I_{ij}}, \quad \forall j = 1, \ldots, m. \tag{23}$$

The framework covers three key sources of variation. (i) *Heterogeneity*: aggregating choices while assuming a representative agent may produce apparently irrational patterns simply due to heterogeneous preferences, rather than irrational individuals. (ii) *Assignment variation*: subjects may face different subsets of choice problems. In experiments, this reflects time or budget constraints where each subject is assigned only a subset of questions; in observational data, this captures consumers encountering different products. (iii) *Repeated choices*: subjects' responses may exhibit varying

29

degrees of correlation when they face the same choice problem multiple times. This variation highlights the need for careful experimental design when researchers assume independent decisions to estimate individual choice probabilities.

I specify the first two sources of variation in Assumption 1, and the third in Assumption 2.

Assumption 1 (Sampling and Random Assignment).

(a) *Individual choice probabilities:* $\boldsymbol{\rho}_i \sim_{iid} F$, *where F is a distribution over $[0,1]^m$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}_0$.*

(b) *Assignments:* $\boldsymbol{K}_i \sim_{iid} G$, *where G is a distribution over $\{0, 1, \ldots, \bar{K}\}^m$ and K_{ij} denotes how many times subject i faces problem j.*

(c) *Independence:* $\boldsymbol{\rho}_i \perp \boldsymbol{K}_i$ *(no selection bias between preferences and assignments).*

Remark 3. (For Assumption 1)

(i). Assumption 1(a) is standard in the literature. (b) and (c) generalize Kitamura and Stoye (2018) to allow repeated observations. This framework encompasses various assignment mechanisms, such as random sampling with replacement, where subjects face randomly drawn problems; or lab experiments, where subjects are randomly assigned to sessions with predetermined choice problems.

(ii). The independence assumption in (c) rules out selection bias between preferences and the problems subjects face. This assumption is naturally satisfied under random assignment in experiments, but may be violated in observational settings where consumers self-select into choice situations. For instance, price-sensitive consumers may seek out discount stores and thus more frequently encounter choice problems there.

When subjects face the same choice problem multiple times, their responses may exhibit varying degrees of correlation. I accommodate the full spectrum: from independent responses (treating each as a fresh decision) to perfect correlation (always repeating the initial choice) and any intermediate pattern. The test remains valid regardless of the actual correlation structure, offering robustness in settings where the degree of correlation is unknown or may vary across subjects.

Formally, let $\hat{\rho}_{ij}^{(k)}$ denote subject i's response the kth time facing problem j, and let

$$\hat{\boldsymbol{\rho}}_{ij} = [\hat{\rho}_{ij}^{(1)}, \ldots, \hat{\rho}_{ij}^{(K_{ij})}]^T \in \{0,1\}^{K_{ij}}$$

denote the vector of all responses from subject i facing problem j. For any k and $p \in [0,1]$, let $H(k, p)$ denote a distribution over valid correlation matrices for a k-dimensional Bernoulli(p)

random vector.

Assumption 2 (**Repeated Choice Responses**).

For subject i facing problem j multiple times $(K_{ij} > 0)$:

(a) *Each response:* $\hat{\rho}_{ij}^{(k)}|\rho_{ij} \sim Bernoulli(\rho_{ij})$ *for* $k = 1, \ldots, K_{ij}$

(b) *Correlation structure:* $Cov(\hat{\boldsymbol{\rho}}_{ij}|K_{ij}, \boldsymbol{\rho}_i) = \rho_{ij}(1 - \rho_{ij})\boldsymbol{V}_{ij}$, *where* $\boldsymbol{V}_{ij} \sim H(K_{ij}, \rho_{ij})$

The observed average choice frequency for subject i on problem j is:

$$\bar{\rho}_{ij} = \begin{cases} \frac{1}{K_{ij}} \sum_{k=1}^{K_{ij}} \hat{\rho}_{ij}^{(k)} & \text{if } K_{ij} \neq 0 \\ 0 & \text{if } K_{ij} = 0 \end{cases}, \quad \forall i = 1, \ldots, N, \ \forall j = 1, \ldots, m. \tag{24}$$

Lemma 4. *Under Assumptions 1, 2, and Regularity Assumptions 3 in Appendix A.9.1*

$$\sqrt{N}\,(\hat{\boldsymbol{\rho}} - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \ \boldsymbol{\Sigma}), \tag{25}$$

where $\boldsymbol{\Sigma}$ is defined in Appendix A.10.

The asymptotic covariance $\boldsymbol{\Sigma}$ captures the above three sources of variation: population distribution ($\boldsymbol{\Sigma}_0$), assignment patterns (G), and within-subject correlation across repeated observations. In a special case where each subject faces all problems exactly once, G degenerates to a constant and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$. Under the Regularity Assumptions 3 (d) and (e), $\boldsymbol{\Sigma}$ is positive definite.

3.2 Test Statistic and Uniform Validity

Let $Q \subseteq [0, 1]^m$ denote the polytope of collective rationalizability (SS, MU, or WU). By Theorem 1, the testing problem is

$$H_0: \quad \boldsymbol{\mu} \in Q, \quad \text{against} \quad H_1: \quad \boldsymbol{\mu} \notin Q.$$

The null hypothesis holds when the mean of population distribution lies within the rationalizability polytope. The test statistic measures the distance from the estimated frequencies to this polytope.

$$T_N = \sqrt{N}\,\big\|\hat{\boldsymbol{\rho}} - \text{proj}_Q(\hat{\boldsymbol{\rho}}, \boldsymbol{\Omega})\big\|_{\Omega}, \tag{26}$$

where $\|\boldsymbol{x}\|_{\Omega} = \sqrt{\boldsymbol{x}^T \boldsymbol{\Omega} \boldsymbol{x}}$, and $\text{proj}_Q(\cdot, \boldsymbol{\Omega})$ is the projection onto Q with $\boldsymbol{\Omega}$-induced distance, defined as

$$\text{proj}_Q(\boldsymbol{x}, \boldsymbol{\Omega}) = \arg\min_{\boldsymbol{q} \in Q} \ \|\boldsymbol{x} - \boldsymbol{q}\|_{\Omega}. \tag{27}$$

Define the transformation $\phi(\boldsymbol{x}) = \|\boldsymbol{x} - \text{proj}_Q(\boldsymbol{x})\|$,[17] then the null can then be written as

$$H_0: \quad \phi(\boldsymbol{\mu}) \leq 0 \quad \text{against} \quad H_1: \quad \phi(\boldsymbol{\mu}) > 0,$$

and $T_N = \sqrt{N}\,(\phi(\hat{\boldsymbol{\rho}}))$, where from Lemma 4, $\sqrt{N}\,(\hat{\boldsymbol{\rho}} - \boldsymbol{\mu}) \to \mathbb{G}_0 := \mathcal{N}(\boldsymbol{0},\ \boldsymbol{\Sigma})$. The function ϕ measures the Euclidean distance projecting onto a convex polytope. Its limiting distribution does not vary continuously in $\boldsymbol{\mu}$ when $\boldsymbol{\mu}$ lies on the boundary of Q: intuitively, as $\boldsymbol{\mu}$ moves across edges and facets of the polytope, the distribution changes abruptly. This discontinuity invalidates the consistency of standard bootstrap methods in this setting (Andrews, 2000; Fang and Santos, 2019).[18] Fortunately, the function ϕ is *directionally differentiable* and, moreover, Lipschitz and convex. These properties allow me to adopt the numerical delta method from Hong and Li (2018), which provides *uniformly valid inference.*

Formally, the key insight is that for directionally differentiable ϕ,

$$\sqrt{N}\,(\phi(\hat{\boldsymbol{\rho}}) - \phi(\boldsymbol{\mu})) \to \phi'_{\boldsymbol{\mu}}(\mathbb{G}_0), \tag{28}$$

where $\phi'_{\boldsymbol{\mu}}$ denotes the directional derivative of ϕ at $\boldsymbol{\mu}$ (Shapiro, 1991).[19] Define

$$\mathbb{Z}_N^* = \sqrt{N}\,(\hat{\boldsymbol{\rho}}_b - \hat{\boldsymbol{\rho}}), \tag{29}$$

where $\hat{\boldsymbol{\rho}}_b$ is parameter estimates obtained using standard bootstrap. Then for positive step size $\epsilon_N \to 0, \sqrt{N}\epsilon_N \to \infty$, the pointwise validity of the inference is obtained as

$$\hat{\phi}'_N(\mathbb{Z}_N^*) := \frac{1}{\epsilon_N}\,(\phi(\hat{\boldsymbol{\rho}} + \epsilon_N \mathbb{Z}_N^*) - \phi(\hat{\boldsymbol{\rho}})) \to \phi'_{\boldsymbol{\mu}}(\mathbb{G}_0). \tag{30}$$

Moreover, for the uniform inference, ϕ being convex implies that $\frac{1}{\epsilon_N}\,(\phi(\boldsymbol{\mu} + \epsilon_N \mathbb{G}_0) - \phi(\boldsymbol{\mu}))$ first-order stochastically dominates $\phi'_{\boldsymbol{\mu}}(\mathbb{G}_0)$, and ϕ being Lipschitz implies that $\frac{1}{\epsilon_N}\,(\phi(\boldsymbol{\mu} + \epsilon_N \mathbb{G}_0) - \phi(\boldsymbol{\mu}))$ can be replaced by its feasible sample version $\hat{\phi}'_N(\mathbb{Z}_N^*)$; these together imply that the $1-\alpha$ quantile of $\hat{\phi}'_N(\mathbb{Z}_N^*)$ provides uniform size control (Romano and Shaikh, 2012).

Theorem 2 (Uniformly Valid Inference for Collective Rationalizability).

Consider the following numerical delta method with S bootstrap replications. For each $s = 1, \ldots, S$, draw $\mathbb{Z}_s$ from the distribution of $\mathbb{Z}_N^*$ defined in (29). For a given ϵ_N such that $\epsilon_N \to 0, \sqrt{N}\epsilon_N \to \infty$,

---

[17]Note that since the numerical delta method that will be introduced shortly provides a uniformly valid testing procedure, the choice of weighting matrix $\boldsymbol{\Omega}$ in the distance is immaterial: rescaling by $\boldsymbol{\Omega}$ corresponds to a linear transformation of the data and the null polytope. Hence, I set $\boldsymbol{\Omega} = I$ without loss of generality. See Appendix A.9.2 for more detailed argument.

[18]See Appendix A.9.4 for the proof of *pointwise* validity of the subsampling procedure in our case.

[19]See Appendix A.9.3 for the detailed analysis on the pointwise asymptotic distribution of T_N, which relates closely to Fang and Santos (2019)'s discussion on convex set projections.

evaluate for each s:

$$\hat{\phi}'_N(\mathbb{Z}^*_s) := \frac{1}{\epsilon_N} \left(\phi(\hat{\boldsymbol{\rho}} + \epsilon_N \mathbb{Z}^*_s) - \phi(\hat{\boldsymbol{\rho}}) \right). \tag{31}$$

Let $\hat{c}_{1-\alpha}$ denote the $1 - \alpha$ quantile of the empirical distribution of $\hat{\phi}'_N(\mathbb{Z}^*_s)$. Under Assumptions 1, 2, and 3,

$$\liminf_{N \to \infty} \inf_{\boldsymbol{\mu} \in Q} \Pr[T_N \le \hat{c}_{1-\alpha}] \ge 1 - \alpha. \tag{32}$$

3.3 Monte Carlo Simulations

The performance of the proposed testing procedure is evaluated through Monte Carlo simulations, focusing on Collective Simple Scalability rationalizability with three choice options. I construct a power curve by varying the distance between the population mean $\boldsymbol{\mu}$ and the rationalizability polytope $Q \subset \mathbb{R}^3$. Starting from a boundary point $\boldsymbol{\mu}_0 = [2/3,\ 1/3,\ 2/3]^T$ on a facet of Q, I move along the normal vector $[1, -1, 1]^T$ in both directions, setting $\boldsymbol{\mu} = \boldsymbol{\mu}_0 + d \cdot [1,\ -1,\ 1]^T$ where d controls the distance from the boundary. For $d < 0$, $\boldsymbol{\mu}$ lies inside Q (collectively rationalizable); for $d > 0$, $\boldsymbol{\mu}$ lies outside Q (not collectively rationalizable). The simulation examines nine values $d \in \{0,\ \pm 0.01,\ \pm 0.02,\ \pm 0.04,\ \pm 0.08\}$. Small deviations such as $d = 0.01$ represent near-rational behavior, while $d = 0.08$ indicates substantial departures from rationalizability.

Table 1: Population means $\boldsymbol{\mu}$ used for the Monte Carlo simulations[a]

	$\boldsymbol{\mu}_4$	$\boldsymbol{\mu}_3$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_0$	$\boldsymbol{\mu}'_1$	$\boldsymbol{\mu}'_2$	$\boldsymbol{\mu}'_3$	$\boldsymbol{\mu}'_4$
$\boldsymbol{\mu}$	$\begin{bmatrix} 0.59 \\ 0.41 \\ 0.59 \end{bmatrix}$	$\begin{bmatrix} 0.63 \\ 0.37 \\ 0.63 \end{bmatrix}$	$\begin{bmatrix} 0.65 \\ 0.35 \\ 0.65 \end{bmatrix}$	$\begin{bmatrix} 0.66 \\ 0.34 \\ 0.66 \end{bmatrix}$	$\begin{bmatrix} 0.67 \\ 0.33 \\ 0.67 \end{bmatrix}$	$\begin{bmatrix} 0.68 \\ 0.32 \\ 0.68 \end{bmatrix}$	$\begin{bmatrix} 0.69 \\ 0.31 \\ 0.69 \end{bmatrix}$	$\begin{bmatrix} 0.71 \\ 0.29 \\ 0.71 \end{bmatrix}$	$\begin{bmatrix} 0.75 \\ 0.26 \\ 0.75 \end{bmatrix}$
d	-0.08	-0.04	-0.02	-0.01	0	0.01	0.02	0.04	0.08

[a] Assume the individual choice frequency $\boldsymbol{\rho}_i \sim_{iid} F$, where F is a distribution over $[0,1]^3$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}_0 = 0.01 \times I_3$. $\boldsymbol{\mu}_0$ lies on the facet of the collective-SS rationalizable polytope Q. Define $\boldsymbol{\mu} = \boldsymbol{\mu}_0 + d \cdot [1, -1, 1]^T$. An increase of d implies $\boldsymbol{\mu}$ getting deeper inside the polytope, with $d = 0$ indicating $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ on the boundary, and $d < 0$ indicating $\boldsymbol{\mu}$ lies outside Q.

I consider 18 d.g.p.s combining nine different values of population mean $\boldsymbol{\mu}$ (Table 1) and two within-subject correlation structures: (i). Independent responses (d.g.p.s 1–9): subjects treat repeated problems as fresh decisions; (ii). Perfect correlation (d.g.p.s 10–18): subjects always repeat their initial choice. Individual choice probabilities are drawn from $\boldsymbol{\rho}_i \sim_{iid} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$ with $\boldsymbol{\Sigma}_0 = 0.01 \times I_3$. Each subject faces 10 randomly selected binary problems (with replacement). Sample sizes are $N \in \{100,\ 200,\ 500\}$. Figure 7 visualizes the population distributions for three representative cases, illustrating how individual heterogeneity generates a cloud of choice probabilities around the population mean.
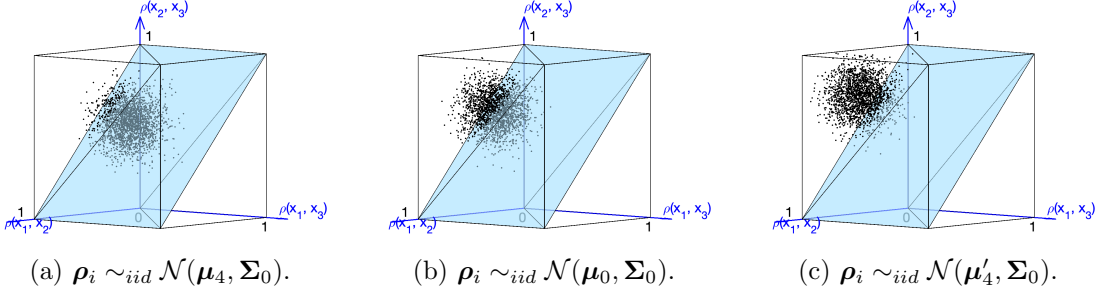
(a) $\boldsymbol{\rho}_i \sim_{iid} \mathcal{N}(\boldsymbol{\mu}_4, \boldsymbol{\Sigma}_0)$.   (b) $\boldsymbol{\rho}_i \sim_{iid} \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.   (c) $\boldsymbol{\rho}_i \sim_{iid} \mathcal{N}(\boldsymbol{\mu}_4', \boldsymbol{\Sigma}_0)$.

Figure 7: Visualization of population distribution $\boldsymbol{\rho}_i$, with 2000 sample size, under different $\boldsymbol{\mu}$. **(a).** $\boldsymbol{\mu}_4$ is in the interior of Q. **(b).** $\boldsymbol{\mu}_0$ is on its boundary. **(c).** $\boldsymbol{\mu}_4'$ is outside the interior of Q.

I implement the numerical delta method with 5000 replications per d.g.p, using bootstrap sample size $N_b = N$, and two step sizes $\epsilon_N = N^{-1/3}$ or $\epsilon_N = N^{-1/6}$, satisfying $\epsilon_N \to 0, \sqrt{N}\epsilon_N \to \infty$. The nominal size is set at $\alpha = 0.05$.

Table 2: Monte Carlo results: rejection rates[a]

N	$\boldsymbol{\mu}_4$	$\boldsymbol{\mu}_3$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_0$	$\boldsymbol{\mu}_1'$	$\boldsymbol{\mu}_2'$	$\boldsymbol{\mu}_3'$	$\boldsymbol{\mu}_4'$
d.g.p. 1–9: Each decision is independent; step size $\epsilon_N = N^{-1/3}$.									
100	0.000	0.000	0.000	0.018	0.054	0.147	0.319	0.757	0.999
200	0.000	0.000	0.001	0.010	0.048	0.214	0.495	0.950	1.000
500	0.000	0.000	0.000	0.002	0.051	0.371	0.828	1.000	1.000
d.g.p. 1–9: Each decision is independent; step size $\epsilon_N = N^{-1/6}$.									
100	0.000	0.000	0.002	0.017	0.056	0.154	0.334	0.761	0.999
200	0.000	0.000	0.001	0.008	0.049	0.216	0.499	0.948	1.000
500	0.000	0.000	0.000	0.002	0.050	0.360	0.821	1.000	1.000
d.g.p. 10–18: Each decision replicates the first encounter; step size $\epsilon_N = N^{-1/3}$.									
100	0.000	0.000	0.007	0.019	0.053	0.119	0.245	0.590	0.980
200	0.000	0.000	0.003	0.013	0.056	0.167	0.370	0.836	1.000
500	0.000	0.000	0.000	0.004	0.053	0.263	0.652	0.994	1.000
d.g.p. 10–18: Each decision replicates the first encounter; step size $\epsilon_N = N^{-1/6}$.									
100	0.000	0.000	0.006	0.020	0.056	0.137	0.260	0.602	0.981
200	0.000	0.000	0.003	0.016	0.056	0.172	0.374	0.828	1.000
500	0.000	0.000	0.000	0.003	0.050	0.261	0.656	0.995	1.000

[a] See Table 1 for definition of $\boldsymbol{\mu}$ vectors. Recall that $\boldsymbol{\mu}_0$ is on the boundary of Q. From $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}_4$, the population mean moves away from the polytope; from $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}_4'$, the population mean moves into the interior of the polytope. We keep the variance of the population distribution $\boldsymbol{\Sigma}_0$ the same across all d.g.p.s at $\boldsymbol{\Sigma}_0 = 0.01 \times I_3$. In the d.g.p.s 1–9, subjects make each decision independently; while in the d.g.p.s 10–18, subjects replicate the decision they made when first encountering the choice problem. All entries are rejection rates computed from 5000 simulations, with sample sizes $N = \{100, 200, 500\}$ and bootstrap sample size $N_b = N$. The step size for the numerical delta is $\epsilon_N = N^{-1/3}$ or $\epsilon_N = N^{-1/6}$.

Table 2 reports rejection rates under different d.g.p.s. The test maintains appropriate size control at the boundary ($\boldsymbol{\mu}_0$), with empirical rejection rates close to the nominal 5%. Power increases monotonically with distance from the polytope and sample size. In particular, comparing across d.g.p.s 1–9 and d.g.p.s 10–18, when decisions are independent across repetitions, a smaller sample size is required to achieve 95% power compared to that when repetitions are perfectly correlated. Comparing within d.g.p.s 1–9 and d.g.p.s 10–18, the step size choice $\epsilon_N = N^{-1/6}$ yields slightly higher power than $\epsilon_N = N^{-1/3}$, although overall both choices perform similarly well.

4 Empirical Application: Two Existing Experiments

The main contribution of this paper is the characterization of collective rationalizability (Theorem 1) and its corresponding statistical testing (Theorem 2). Together, they enable testing for random choice models while allowing for heterogeneity, which is especially important when choices are collected from possibly heterogeneous subjects but researchers have to assume representative agent for estimation or testing. The framework addresses the empirical challenge that researchers are often limited to finite observations of individuals making few or no repeated choices from the same choice set. In the following, I demonstrate the practical applicability of collective rationalizability by re-analyzing data from two existing experiments, and show that comparison difficulty can be not only theoretically but also empirically confused with heterogeneity in both cases.

Experiment 1: Snack food choices (Clithero, 2018)**.** This paper studies binary choices among snack foods. Subjects ($N = 31$) choose between all possible pairs from 17 snacks, yielding 136 binary comparisons per subject. Each subject faces each pair exactly once, resulting in 4,216 total observations. The within-subject design provides complete choice data across all pairs, but with only single observation per subject-pair combination.[20]

Experiment 2: Perceptual similarity of images (Prashnani et al., 2018)**.** The experiment examines perceptual judgments of image similarity. Given a reference image and two distorted versions based on it, subjects indicate which distortion appears more similar to the reference. The experiment has 40 reference images with 15 distortion types applied to each. With 40 responses collected per comparison across 4,800 binary choice problems, this yields 192,000 total observations. Importantly, the public dataset contains only aggregate choice frequencies without individual-level tracking.[21]

These datasets illustrate two key strengths of our collective rationalizability approach:

First, neither experiment has natural "ground truth" preferences. In food choices, researchers want to remain agnostic about which snacks people should prefer. For image similarity, the complexity

---

[20]In particular, I only use the Two-Alternative Forced-Choice Task (2AFC-Task) for the test. More discussion on the Yes-No Choice Task (YN-Task) in Clithero (2018) will be discussed later.

[21]In particular, I only use their test set, where 40 human responses for each pairwise comparison are collected.

of distortion types (blur, noise, rotations, etc.) makes it hard to assess the objective similarity.[22] Other existing methods (Tversky and Russo, 1969; Enke and Shubatt, 2023) require either known preferences or specially designed tasks with objective answers, neither of which can be applied here.

Second, both datasets exemplify common data limitations that prevent direct application of existing revealed-preference approaches. In Clithero (2018)'s data, single observations per subject-pair makes estimating individual probabilities nontrivial. In Prashnani et al. (2018)'s data, individual identifiers are unavailable publicly. Tests of stochastic transitivity require estimating individual choice probabilities, which is impossible in both cases. Our collective framework is designed precisely for such scenarios, testing models using aggregate patterns rather than requiring repeated individual observations.

4.1 Application 1: Snack Food Choices

I first apply the methods to the experimental data from Clithero (2018), who studies binary choices among $n = 17$ different snack foods. In their two-alternative forced choice (2AFC) task, subjects ($N = 31$) chose which of two items they would prefer to eat at the end of the experiment. Each subject faced all 136 possible binary pairs exactly once. Examples of the snacks are shown in Figure 8.

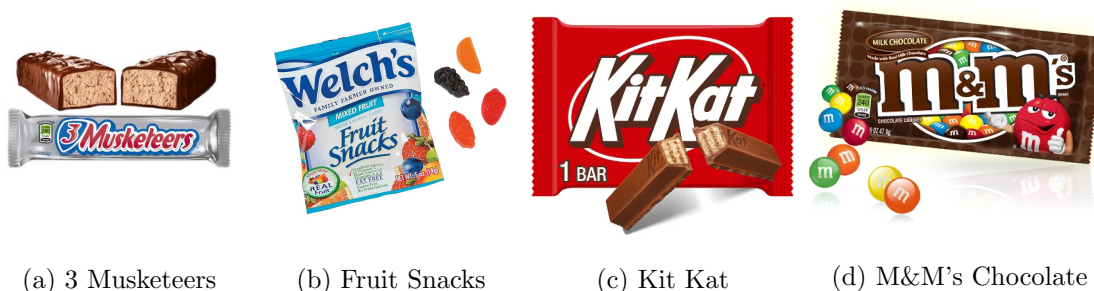(a) 3 Musketeers   (b) Fruit Snacks   (c) Kit Kat   (d) M&M's Chocolate

Figure 8: Examples of the types of snack foods involved in the experiment.

Comparison of the various models. I analyze the aggregate choice frequencies, i.e., the fraction of subjects choosing snack i over j, and examine the violation rate under the various models, SS/MU/WU and Collective SS. The results are summarized in Table 3. We see that Simple Scalability is heavily violated when treating the data as coming from a representative agent: among the 680 possible triples of snacks, 41.8% violate strong stochastic transitivity, and this number increases to 99% when we check all subsets of size 5. This raises a key question: does this violation stem from varying comparison difficulty across pairs, or can heterogeneity alone explain it?

---

[22]Actually the primary research question asked in the original paper Prashnani et al. (2018) is: what is the human perception of these different distortions, and how does it differ from existing metrics? Therefore, researchers of course do not want to impose strong assumptions about image similarities before the experiment.

Table 3: Rationalizability of the dataset under different models[a]

Choice subset size	Total subset count	Individual SS violation rate	Individual MU violation rate	Individual WU violation rate	Collective SS violation rate
3	680	0.418	0.107	0.040	0.000
4	2,380	0.856	0.324	0.130	0.000
5	6,188	0.990	0.577	0.264	0.000

[a] We enumerate all choice option subsets of size 3, 4 and 5 in the dataset with full pairwise choice rates, and report the total number of subsets, and the proportion with population choice rates violating individual SS/MU/WU and collective SS.

A researcher not accounting for heterogeneity across subjects may observe the significant reduction in violation rates after introducing comparison difficulty and conclude that comparison difficulty is the key factor. However, using the characterization developed in this paper, we see that all subsets with sizes 3, 4, and 5 are *collectively* rationalizable by Simple Scalability, which cannot be rejected with statistical power, implying that heterogeneity alone can be sufficient for explaining all violations of Simple Scalability in this case.

Examples of transitivity violation. Consider a concrete example: Fruit Snacks, M&M's, and KitKat. The aggregate choice rates are:

$$\rho(\text{KitKat}, \text{M\&M's}) = 0.839, \quad \rho(\text{M\&M's}, \text{FruitSnacks}) = 0.581, \quad \rho(\text{KitKat}, \text{FruitSnacks}) = 0.613.$$

They violate strong stochastic transitivity if assuming a representative agent: the preference for KitKat over M&M's and M&M's over Fruit Snacks should imply an even stronger preference for KitKat over Fruit Snacks, hence $\rho(\text{KitKat}, \text{FruitSnacks}) > 0.839$, yet we observe only 0.613. This pattern can be explained by *comparison difficulty* alone, that KitKat is more difficult to be compared with Fruit Snacks than with M&M's. However, it can also be explained by *heterogeneity* alone. The choice rates are *collectively* rationalized by Simple Scalability: imagine around half the subjects prefer chocolate-based snacks while half prefer fruit flavors, and among the two chocolate-based snacks, most prefer KitKat. Each individual could satisfy Simple Scalability perfectly, yet their aggregation produces the observed violation.

In another concrete example: Fruit Snacks, M&M's, and Reese's, the aggregate choice rates are:

$$\rho(\text{M\&M's}, \text{FruitSnacks}) = 0.581, \quad \rho(\text{FruitSnacks}, \text{Reese's}) = 0.581, \quad \rho(\text{Reese's}, \text{M\&M's}) = 0.581.$$

They violate even weak stochastic transitivity, which implies that violations of Simple Scalability in this example cannot be explained by comparison difficulty alone. However, they are again collectively rationalizable by Simple Scalability. In fact, one way to collectively rationalize the choice is reminiscent of the Condorcet paradox (Condorcet, 1785); the only difference is that instead of deterministic choices, each individual chooses the preferred alternative with probability 0.87.

Evidence of heterogeneity. The collective rationalizability reveals that heterogeneity alone suffices to explain all violations of Simple Scalability in this dataset. Fortunately, independent validation is available: Clithero (2018)'s experimental design includes a separate Yes-No (YN) task that allows testing for heterogeneity directly. In the YN task, subjects are asked whether they preferred each of the 17 snacks over a common reference snack (3 Musketeers candy), with 10 repetitions per snack. Although one cannot directly check stochastic transitivity using the YN task since there is no cycle structure, these repetitions help us look at individual preferences more closely. Table 4 summarizes subjects' responses over the two questions involving Fruit Snacks and M&M's.

Table 4: Summary of subjects' responses in the YN task for certain questions[a]

Fruit Snacks $\succ$ 3 Musketeers?	M&M's $\succ$ 3 Musketeers?	Number of subjects
"Yes" all 10 times	"No" all 10 times	9
"No" all 10 times	"Yes" all 10 times	6
"Yes" all 10 times	"Yes" all 10 times	7
"No" all 10 times	"No" all 10 times	4
Other response patterns		5
Total		31

[a] The majority of subjects (26 out of 31) provided consistent responses to both questions across all 10 repetitions. Among these 26 subjects, responses were distributed across all four possible Yes/No combinations, indicating substantial heterogeneity in individual preferences.

Table 4 shows that the majority of subjects provided *consistent* responses across all 10 repetitions of both questions. Among these subjects, 9 exhibited the preference ordering Fruit Snacks $\succ$ 3 Musketeers $\succ$ M&M's, while 6 others displayed the exactly opposite ordering M&M's $\succ$ 3 Musketeers $\succ$ Fruit Snacks. This finding aligns with our earlier conjecture that the violation of Simple Scalability may stem from a population with polarized preferences between Fruit Snacks and M&M's. Additionally, 11 subjects responded with either "Yes" or "No" consistently for both questions across all 10 repetitions. While these responses are inconclusive regarding the relative preference between Fruit Snacks and M&M's, they provide further evidence of substantial preference heterogeneity among subjects.

Furthermore, using the YN task data, I perform a statistical test of whether all individuals share the same preference signs (i.e., whether $\text{sgn}(\rho_i - \frac{1}{2})$ is identical across individuals). A likelihood-ratio test rejects this hypothesis ($p < 0.001$), confirming the preference heterogeneity in the population (details in Appendix A.12).

4.2 Application 2: Image Similarity Perception

I next apply the framework to data from Prashnani et al. (2018), who study how people judge similarity between images. In their experiment, subjects compare pairs of distorted images and

identify which is more similar to a reference image (Figure 9). The dataset contains 40 reference images with 15 distortion types each, yielding 4,800 binary comparisons with 40 responses per comparison (192,000 total observations).[23]

In Prashnani et al. (2018), they model these judgments using Simple Scalability with a representative agent, estimating a single "similarity score" for each distorted image.

$$\rho(A, B) = F(s_A, s_B) = \frac{1}{1 + e^{s_A - s_B}}. \tag{33}$$

While this perceptual task differs from economic choice, the framework can be applied by treating similarity judgments analogously to preferences. Each person's perception of how similar image A is to the reference constitutes their "preference" for A. The comparison difficulty between images A and B captures the "similarity of similarity": how difficult is it to compare the similarity of image A versus that of image B to the same reference image? This framework allows testing whether the representative-agent Simple Scalability model adequately captures the data, or whether heterogeneity and comparison difficulty play important roles.

Table 5 shows substantial violations of Simple Scalability if assuming a representative agent. Among size-3 subsets, approximately 30% violate strong stochastic transitivity; this rises to 70% for size-4 subsets and exceeds 90% for size-5 subsets. The representative-agent SS model clearly fails to capture the aggregate patterns.

Table 5: Rationalizability of the dataset under different models[a]

Choice subset size	Total subset count	Individual SS violation rate	Individual MU violation rate	Individual WU violation rate	Collective SS violation rate
3	22,400	0.298	0.034	0.009	0.001
4	72,800	0.700	0.117	0.032	0.003
5	174,720	0.926	0.246	0.073	0.007

[a] We enumerate all choice option subsets of size 3, 4 and 5 in the dataset with full pairwise choice rates, and report the total number of subsets, and the proportion with population choice rates violating individual SS/MU/WU and collective SS.

Remarkably, allowing for heterogeneity virtually eliminates these violations. Collective SS is violated in only 0.1% of size-3 subsets and remains below 1% even for size-5 subsets. This finding has important implications: researchers modeling perceptual judgments may not need complex models with varying comparison difficulty. Instead, acknowledging that people have heterogeneous perceptions of similarity may suffice. Here heterogeneous perceptions mean that different individuals genuinely perceive image similarity differently, and what appears as intransitivity in aggregate data might reflect diverse but internally consistent individual perceptions.

---

[23]In particular, we only use their test set, where 40 human responses for each pairwise comparison are collected.

Table 6: Statistical test result of collective Simple Scalability[a]

Choice subset size	Rejection instances at $\alpha = 0.05$
3	1
4	13
5	66

[a] I conduct the statistical testing for Collective Simple Scalability from Theorem 2, and report the number of instances where collective SS is rejected at the 0.05 significance level. To avoid some p-hacking concerns, I employ sample splitting. Using the first half of the data, I identify which choice subsets violate Collective SS based on the theoretical characterization. For these subsets that exhibit theoretical violations, I then conduct the statistical test from Theorem 2 on the second half of the sample, reporting the number of rejections at the 0.05 significance level.

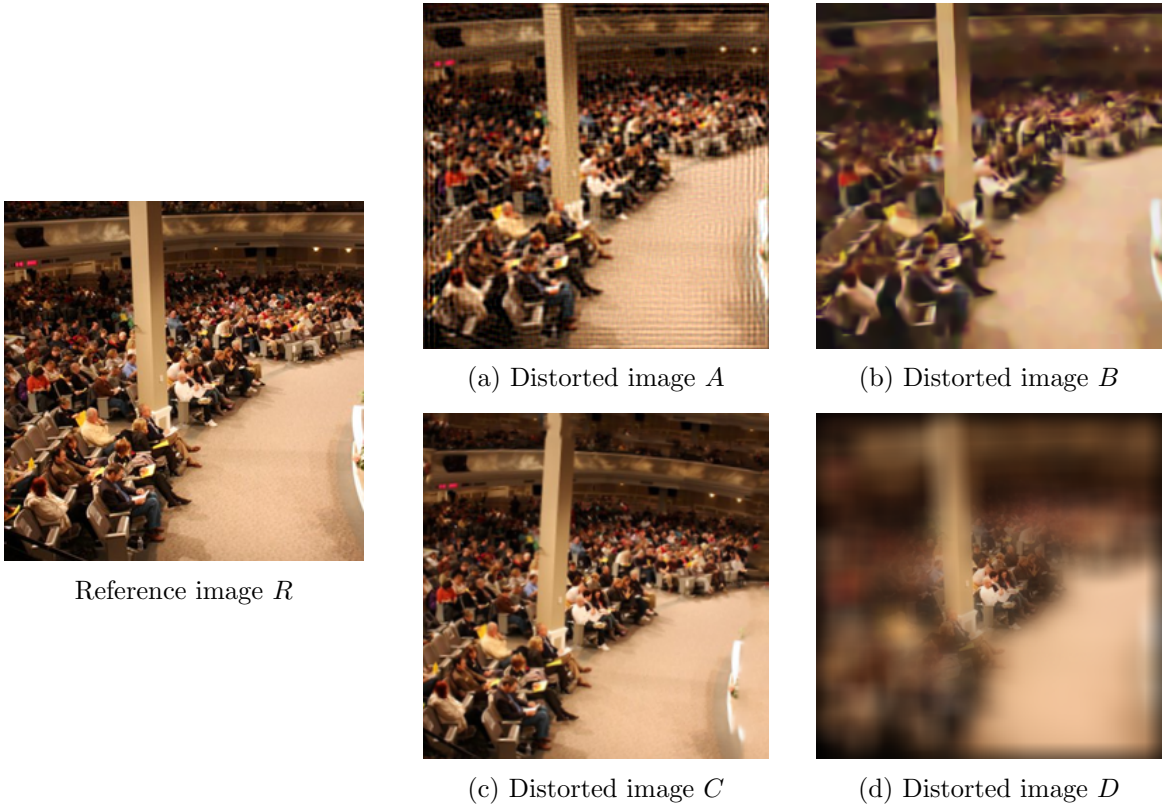Reference image R

(a) Distorted image A

(b) Distorted image B

(c) Distorted image C

(d) Distorted image D

Figure 9: Examples of reference and distorted images. In Prashnani et al. (2018)'s Amazon Mechanical Turk (MTurk) experiments, there are 40 reference images, and 15 different distortions are applied to each reference image. Subjects are asked to choose between all *pairwise comparison*: Which image is more similar to the reference? In total, there are 4800 binary choice problems, and 40 responses for each pairwise comparison are collected. Individual-level data regarding which responses come from the same individual is not available.
(a)(b). Which image, A or B, is more similar to the reference R? In this example of pairwise image comparison, 66.7% of the subjects choose image A as the more similar.
(c)(d). Which image, C or D, is more similar to the reference R? In this example of pairwise image comparison, most people (92.5%) choose image C as the more similar.

4.3 Comparison of Model Restrictiveness

At this point, one might wonder: does Collective Simple Scalability fit the empirical data well simply because it is too permissive? To explore this question, I compute the relative volumes of the rationalizable sets for each model, or equivalently, the proportion of uniformly random choice data that each model would accept.

Table 7: Relative volume of the rationalizable set of different models[a]

Choice subset size	Individual SS	Individual MU	Individual WU	Collective SS
3	0.250	0.500	0.750	0.667
4	0.008	0.092	0.375	0.250
5	0.00002	0.005	0.117	0.048

[a] Volumes computed analytically from model characterizations, except Individual MU for n=4,5 which required Monte Carlo simulation due to non-convexity.

Table 7 shows that Weak Utility (adding comparison difficulty only) is actually *more permissive* than Collective Simple Scalability (adding heterogeneity only). For $n = 3$, Collective SS accepts 66.7% of the choice space compared to WU's 75%. This gap widens dramatically as n increases: for $n = 5$, Collective SS accepts only 4.8% versus WU's 11.7%. This finding has important implications. Despite being more restrictive, Collective Simple Scalability achieves lower violation rates on the real experimental datasets than Weak Utility for both empirical applications (see, Tables 3, 5). Following the idea of Selten (1991) and Beatty and Crawford (2011), heterogeneity does a better job explaining the violation of Simple Scalability than comparison difficulty in both ways.

5 Conclusion

This paper develops a revealed preference framework for testing random choice models while explicitly allowing for heterogeneity. The central contribution, collective rationalizability, provides both theoretical characterizations and statistical tests to determine whether heterogeneity alone, comparison difficulty alone, or both required, explains violations of Simple Scalability when data is collected from possibly heterogeneous subjects. This framework addresses practical data limitations, the absence of known preferences and limited individual observations, that existing approaches often face in practice.

The empirical applications demonstrate the framework's applicability. In both the food choice and image similarity experiments, violations of Simple Scalability that appear substantial under representative-agent assumptions largely disappear when heterogeneity is properly accounted for through Collective Simple Scalability. Without this framework, these violations might be attributed to comparison difficulty. This demonstrates that not only theoretically, but also empirically, het-

erogeneity can be confused with comparison difficulty to explain the apparent intransitivities.

Several extensions would enhance the framework's theoretical foundations and empirical applicability:

Theoretical extensions to multinomial choices. Extending collective rationalizability from binary to multinomial choice with comparison difficulty would significantly broaden empirical applications. Many experiments involve choices from sets larger than two, and the framework's insights about heterogeneity versus comparison difficulty would prove valuable in these richer settings.

Identification of difficulty with new experiments. Future research could validate the test for collective rationalizability through carefully designed experiments. Building on the theoretical results for recovering the ranking of comparison difficulty developed in this paper, such experiments could empirically test different difficulty metrics in the laboratory settings.

Uniformly valid statistical testing for transitivity. Developing uniformly valid tests for individual rationalizability remains an open challenge. Such tests would enable more precise comparisons between heterogeneity-based and difficulty-based explanations of choice patterns, particularly when individual-level data is available.

Micro-foundations. The violations of Collective Simple Scalability provide valuable cases where heterogeneity alone cannot explain observations. New theories understanding when and why such violations of Collective Simple Scalability arise, and what they reveal about the underlying cognitive foundations of decision-making, is an interesting question for the future research.

References

Andrews, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, pages 399–405.

Arrieta, G. and Nielsen, K. (2024). Procedural decision-making in the face of complexity. Technical report, Working Paper.

Arts, S., Ong, Q., and Qiu, J. (2024). Measuring decision confidence. *Experimental Economics*, 27(3):582–603.

Balas, E. (1971). Intersection cuts—a new type of cutting planes for integer programming. *Operations Research*, 19(1):19–39.

Barahona, N., Otero, C., and Otero, S. (2023). Equilibrium effects of food labeling policies. *Econometrica*, 91(3):839–868.

Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483.

Beatty, T. K. M. and Crawford, I. A. (2011). How demanding is the revealed preference approach to demand? *American Economic Review*, 101(6):2782–2795.

Bernheim, B. D. and Sprenger, C. (2020). On the empirical validity of cumulative prospect theory: Experimental evidence of rank-independent probability weighting. *Econometrica*, 88(4):1363–1409.

Block, H. and Marschak, J. (1959). Random orderings and stochastic theories of response. Technical report, Cowles Foundation for Research in Economics, Yale University.

Bohren, J. A., Hascher, J., Imas, A., Ungeheuer, M., and Weber, M. (2024). A cognitive foundation for perceiving uncertainty. Technical report, National Bureau of Economic Research.

Bronars, S. G. (1987). The power of nonparametric tests of preference maximization. *Econometrica: Journal of the Econometric Society*, pages 693–698.

Brown, Z. Y. and Jeon, J. (2024). Endogenous information and simplifying insurance choice. *Econometrica*, 92(3):881–911.

Chambers, C. P. and Echenique, F. (2016). *Revealed preference theory*. Number 56. Cambridge University Press.

Chipman, J. S. (1960). Stochastic choice and subjective probability. In *Decisions, values and groups*, pages 70–95. Elsevier.

Clithero, J. A. (2018). Improving out-of-sample predictions using response times and a model of the decision process. *Journal of Economic Behavior & Organization*, 148:344–375.

Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Royal (or Royale), Paris.

Conforti, M., Cornuéjols, G., Zambelli, G., Conforti, M., Cornuéjols, G., and Zambelli, G. (2014). *Integer programming models.* Springer.

Coombs, C. and Pruitt, D. (1961). Some characteristics of choice behavior in risky situations. *Annals of the New York Academy of Sciences.*

de Clippel, G., Moscariello, P., Ortoleva, P., and Rozen, K. (2024). Caution in the face of complexity. Technical report, Mimeo.

de Lara, L. and Dean, M. (2024). Rational choice overload.

Dean, M., Ravindran, D., and Stoye, J. (2022). A better test of choice overload. *arXiv preprint arXiv:2212.03931.*

Enke, B. and Graeber, T. (2023). Cognitive uncertainty. *The Quarterly Journal of Economics*, 138(4):2021–2067.

Enke, B. and Shubatt, C. (2023). Quantifying lottery choice complexity. Technical report, National Bureau of Economic Research.

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

Fang, Z. and Santos, A. (2019). Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412.

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

Fishburn, P. C. (1978). A probabilistic expected utility theory of risky binary choices. *International Economic Review*, pages 633–646.

Fishburn, P. C. (1992). Induced binary probabilities and the linear ordering polytope: A status report. *Mathematical Social Sciences*, 23(1):67–80.

Fishburn, P. C. (1998). Stochastic utility. *Handbook of utility theory*, 1:273–319.

Frydman, C. and Jin, L. J. (2022). Efficient coding and risky choice. *The Quarterly Journal of Economics*, 137(1):161–213.

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

Fudenberg, D., Strack, P., and Strzalecki, T. (2018). Speed, accuracy, and the optimal timing of choices. *American Economic Review*, 108(12):3651–3684.

Gabaix, X. (2025). A theory of complexity aversion. *Available at SSRN*.

Georgescu-Roegen, N. (1958). Threshold in choice and the theory of demand. *Econometrica: Journal of the Econometric Society*, pages 157–168.

Gerhardt, H., Biele, G. P., Heekeren, H. R., and Uhlig, H. (2016). Cognitive load increases risk aversion. Technical report, SFB 649 Discussion Paper.

Griswold, B. J. and Luce, R. D. (1962). Choices among uncertain outcomes: A test of a decomposition and two assumptions of transitivity. *The American journal of psychology*, 75(1):35–44.

Grötschel, M., Jünger, M., and Reinelt, G. (1985). Facets of the linear ordering polytope. *Mathematical programming*, 33(1):43–60.

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

He, J. and Natenzon, P. (2023). Random choice and differentiation. Technical report, Technical report, Washington University in Saint Louis.

He, J. and Natenzon, P. (2024). Moderate utility. *American Economic Review: Insights*, 6(2):176–195.

Hey, J. D. and Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica: Journal of the Econometric Society*, pages 1291–1326.

Hong, H. and Li, J. (2018). The numerical delta method. *Journal of Econometrics*, 206(2):379–394.

Iyengar, S. S. and Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology*, 79(6):995.

Katthän, L. (2011). The linear ordering polytope via representations. *arXiv preprint arXiv:1109.5040*.

Khaw, M. W., Li, Z., and Woodford, M. (2021). Cognitive imprecision and small-stakes risk aversion. *The review of economic studies*, 88(4):1979–2013.

Khaw, M. W., Li, Z., and Woodford, M. (2022). Cognitive imprecision and stake-dependent risk attitudes. Technical report, National Bureau of Economic Research.

Kirchler, M., Andersson, D., Bonn, C., Johannesson, M., Sørensen, E. Ø., Stefan, M., Tinghög, G., and Västfjäll, D. (2017). The effect of fast and slow decisions on risk taking. *Journal of Risk and Uncertainty*, 54(1):37–59.

Kitamura, Y. and Stoye, J. (2018). Nonparametric analysis of random utility models. *Econometrica*, 86(6):1883–1909.

Krantz, D. H. (1964). *The scaling of small and large color differences*. University of Pennsylvania.

Lea, A. M. and Ryan, M. J. (2015). Irrationality in mate choice revealed by túngara frogs. *Science*, 349(6251):964–966.

Luce, R. D. (1959). *Individual choice behavior*, volume 4. Wiley New York.

Luce, R. D. and Suppes, P. (1965). Preference, utility and subjective probability. In Luce, R. D., Bush, R. R., and Galanter, E. H., editors, *Handbook of Mathematical Psychology*, volume 3, pages 249–410. Wiley, New York.

Marschak, J. (1959). Binary choice constraints on random utility indicators. Technical report, Cowles Foundation for Research in Economics, Yale University.

Martinetti, D., Montes, S., Díaz, S., and De Baets, B. (2017). On the correspondence between reciprocal relations and strongly complete fuzzy relations. *Fuzzy Sets and Systems*, 322:19–34.

Mattsson, L.-G. and Weibull, J. W. (2002). Probabilistic choice and procedurally bounded rationality. *Games and Economic Behavior*, 41(1):61–78.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*.

McFadden, D. (1978). Modeling the choice of residential location. *Transportation Research Record*, (673).

McFadden, D. (1981). Econometric models of probabilistic choice. *Structural analysis of discrete data with econometric applications*, 198272.

McFadden, D. and Richter, M. K. (1990). Stochastic rationality and revealed stochastic preference. *Preferences, Uncertainty, and Optimality, Essays in Honor of Leo Hurwicz, Westview Press: Boulder, CO*, pages 161–186.

McFadden, D. L. (2005). Revealed stochastic preference: a synthesis. *Economic Theory*, 26(2):245–264.

Natenzon, P. (2019). Random choice and learning. *Journal of Political Economy*, 127(1):419–457.

Ok, E. A. and Tserenjigmid, G. (2023). Measuring stochastic rationality. *arXiv preprint arXiv:2303.08202*.

Oprea, R. (2024a). Complexity and its measurement. Unpublished manuscript.

Oprea, R. (2024b). Decisions under risk are decisions under complexity. *American Economic Review*, 114(12):3789–3811.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., and Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214.

Plonsky, O., Apel, R., Erev, I., Ert, E., and Tennenholtz, M. (2018). When and how can social scientists add value to data scientists? a choice prediction competition for human decision making. *Unpublished Manuscript*.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*, volume 3 of *Springer Series in Statistics*. Springer, New York, NY.

Porcelli, A. J. and Delgado, M. R. (2009). Acute stress modulates risk taking in financial decision making. *Psychological science*, 20(3):278–283.

Prashnani, E., Cai, H., Mostofi, Y., and Sen, P. (2018). Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817.

Puri, I. (2025). Simplicity and risk. *The Journal of Finance*, 80(2):1029–1080.

Richter, M. K. (1966). Revealed preference theory. *Econometrica: Journal of the Econometric Society*, pages 635–645.

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

Romano, J. P. and Shaikh, A. M. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *The Annals of Statistics*, 40(6):2798–2822.

Selten, R. (1991). Properties of a measure of predictive success. *Mathematical social sciences*, 21(2):153–167.

Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1):169–186.

Shubatt, C. and Yang, J. (2024). Tradeoffs and comparison complexity. *arXiv preprint arXiv:2401.17578*.

Soltani, A., De Martino, B., and Camerer, C. (2012). A range-normalization model of context-dependent choice: a new model and evidence. *PLoS computational biology*, 8(7):e1002607.

Strzalecki, T. (2025). Stochastic choice theory. *Cambridge Books*.

Thurstone, L. (1927a). A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Thurstone, L. L. (1927b). Psychophysical analysis. *The American journal of psychology*, 38(3):368–389.

Tversky, A. (1969). Intransitivity of preferences. *Psychological review*, 76(1):31.

Tversky, A. (1972a). Choice by elimination. *Journal of mathematical psychology*, 9(4):341–367.

Tversky, A. (1972b). Elimination by aspects: A theory of choice. *Psychological review*, 79(4):281.

Tversky, A. and Russo, J. E. (1969). Substitutability and similarity in binary choices. *Journal of Mathematical psychology*, 6(1):1–12.

Vaart, A. v. d. and Wellner, J. A. (1997). Weak convergence and empirical processes with applications to statistics. *Journal of the Royal Statistical Society-Series A Statistics in Society*, 160(3):596–608.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Vatter, B. (2025). Quality disclosure and regulation: Scoring design in medicare advantage. *Econometrica*, 93(3):959–1001.

Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, 12(1):579–601.

A Appendix

A.1 Half-space Representation for Rationalizability without Heterogeneity

We first consider $\boldsymbol{A}_k$, $\boldsymbol{b}_k$ for SS, which requires us to write the inequality constraints from strong stochastic acyclicity into the standard linear inequality form $\boldsymbol{A\rho} \leq \boldsymbol{b}$. Let $\pi_k : [1, \dots, n] \to [1, \dots, n]$, $k = 1, \dots, n!$ be the different permutations over the set $[1, \dots, n]$. Also, write $\pi_k(u)$ alternatively as π_k^u for conciseness. The inequality constraints can thus be written as

$$d(\pi_k^u \pi_k^v, \pi_k^u \pi_k^w; \boldsymbol{\rho}) \leq c(\pi_k^u \pi_k^v, \pi_k^u \pi_k^w) \quad \forall 1 \leq u < v < w \leq n, \ k = 1, \dots, n! \tag{34}$$

$$d(\pi_k^v \pi_k^w, \pi_k^u \pi_k^w; \boldsymbol{\rho}) \leq c(\pi_k^v \pi_k^w, \pi_k^u \pi_k^w) \quad \forall 1 \leq u < v < w \leq n, \ k = 1, \dots, n! \tag{35}$$

over all triplets of rankings (u, v, w) and all $n!$ permutations. The quantities d and c are defined as

$$d(xy, zw; \boldsymbol{\rho}) = \begin{cases} \rho(x, y) - \rho(z, w) & \text{if } x < y \text{ and } z < w \\ -\rho(y, x) - \rho(z, w) & \text{if } x > y \text{ and } z < w \\ \rho(x, y) + \rho(w, z) & \text{if } x < y \text{ and } z > w \\ -\rho(y, x) + \rho(w, z) & \text{if } x > y \text{ and } z > w \end{cases}, \quad c(xy, zw) = \begin{cases} 0 & \text{if } x < y \text{ and } z < w \\ -1 & \text{if } x > y \text{ and } z < w \\ 1 & \text{if } x < y \text{ and } z > w \\ 0 & \text{if } x > y \text{ and } z > w \end{cases}$$

$$\tag{36}$$

in such a way that $d(xy, zw; \boldsymbol{\rho}) \leq c(xy, zw)$ is equivalent to the expression $\rho(x, y) \leq \rho(z, w)$, but rewritten to follow the $\frac{n(n-1)}{2}$-dimensional vector representation of $\boldsymbol{\rho}$, namely only containing values $\rho(x_i, x_j)$ with indices $i < j$. For a given set of values (x, y, z, w), $d(xy, zw; \boldsymbol{\rho})$ is a linear function of $\boldsymbol{\rho}$, whereas $c(xy, zw)$ is a constant. Therefore, these are linear inequalities of the form $\boldsymbol{A\rho} \leq \boldsymbol{b}$. For each given permutation π_k, stacking all such linear inequalities together for all triplets (u, v, w) gives the expression for $\boldsymbol{A}_k$ and $\boldsymbol{b}_k$.

We then consider $\boldsymbol{A}_{k,\ell}$, $\boldsymbol{b}_{k,\ell}$ for MU. Recall that for MU we need to break down the entire space of choice rates into simplexes based on different $\boldsymbol{\rho}^* \in R_{\text{deta}}$ and different orderings of the choice magnitudes in $\boldsymbol{\rho}$.

Let Π denote all permutations over the set $\{1, 2, \dots, n\}$. For each of the $n!$ possible permutations $\pi \in \Pi$, we correspond it to one strict preference over the n options, representing $x_{\pi(1)} \succ \cdots \succ x_{\pi(n)}$. In fact, we can choose any strict preference as the "canonical preference" (WLOG we choose $x_1 \succ \cdots \succ x_n$), and then any strict preference σ can be converted to this canonical preference by re-mapping the indices $1, \dots, n$ to $\pi(1), \dots, \pi(n)$.

Furthermore, let $\mathcal{T}$ denote all permutations over the set $\{1, 2, \dots, \frac{n(n-1)}{2}\}$. For each of the $(\frac{n(n-1)}{2})!$ possible permutations $\tau \in \mathcal{T}$, we correspond it to one ordering over the $\frac{n(n-1)}{2}$ pairwise choice rates. representing $\left|\rho(x_{\tau_1(1)}, x_{\tau_2(1)}) - \frac{1}{2}\right| \leq \cdots \leq \left|\rho(x_{\tau_1(\frac{n(n-1)}{2})}, x_{\tau_2(\frac{n(n-1)}{2})}) - \frac{1}{2}\right|$. Here we identify each value in $\{1, 2, \dots, \frac{n(n-1)}{2}\}$ with an index pair in $\{(i, j) \mid 1 \leq i < j \leq n\}$ in the same way that $\boldsymbol{\rho} \in \mathbb{R}^{\frac{n(n-1)}{2}}$ corresponds to $\rho(x_i, x_j)$, and define $(\tau_1(k), \tau_2(k))$ as the index pair corresponding to $\tau(k)$ for $k = 1, \dots, \frac{n(n-1)}{2}$.

Not all $\tau \in \mathcal{T}$ satisfy the MUM constraints. In fact, under the canonical preference we chose above, we have $\rho(i,j) \geq \frac{1}{2}$ for all $1 \leq i < j \leq n$, and hence the choice rate ordering reduces to $\rho(x_{\tau_1(1),\tau_2(1)}) \leq \cdots \leq \rho(x_{\tau_1(\frac{n(n-1)}{2}),\tau_2(\frac{n(n-1)}{2})})$. In order for τ to satisfy MUM, it must satisfy $\tau^{-1}(i,k) \geq \min\left(\tau^{-1}(i,j),\ \tau^{-1}(j,k)\right)$ for all triplets $1 \leq i < j < k \leq n$. Denote this set of τ satisfying MUM under the canonical preference order as $\mathcal{T}^*$, with $L = |\mathcal{T}^*|$.

Now we are ready to characterize the system of inequalities under any given strict preference σ and choice rate ranking τ. We first characterize the conditions under the canonical preference, and then generalize it to arbitrary canonical preference via index re-mapping.

For the canonical strict preference $x_1 \succ \cdots \succ x_n$, we have a total of L polytopes in $\mathbb{R}^{\frac{n(n-1)}{2}}$, each of which corresponds to one $\tau \in \mathcal{T}^*$ and is a simplex. the conditions are

$$\frac{1}{2} \leq \rho(i,j) \leq 1, \qquad \forall 1 \leq i < j \leq n,$$

$$\rho(\tau_1(k),\tau_2(k)) \leq \rho(\tau_1(k+1),\tau_2(k+1)), \qquad \forall k = 1,\ldots,\frac{n(n-1)}{2} - 1.$$

With a general strict preference π, we essentially need to replace each $i = 1,\ldots,n$ with $\pi(i)$. The caveat is that this will lead to $\rho(i,j)$ terms where $i > j$, which do not directly exist in the $\boldsymbol{\rho}$ vector and need to converted into $1 - \rho(j,i)$. In fact, let $\rho[i,j]$ denote the "unordered" indexing of ρ, such that $\rho[i,j] = \rho[j,i] = \rho(i,j)$ for $i < j$, and we can write $\rho(i,j) = \operatorname{sgn}(j-i) \cdot \rho[i,j] + \mathbb{1}[i > j]$. With this notation, we can perform the index remapping with π as

$$\frac{1}{2} \leq \operatorname{sgn}(\pi(j) - \pi(i)) \cdot \rho[\pi(i),\pi(j)] + \mathbb{1}[\pi(i) > \pi(j)] \leq 1, \qquad \forall 1 \leq i < j \leq n,$$

and

$$\operatorname{sgn}(\pi(\tau_2(k)) - \pi(\tau_1(k))) \cdot \rho[\pi(\tau_1(k)),\pi(\tau_2(k))] + \mathbb{1}[\pi(\tau_1(k)) > \pi(\tau_2(k))] \leq$$
$$\operatorname{sgn}(\pi(\tau_2(k+1)) - \pi(\tau_1(k+1))) \cdot \rho[\pi(\tau_1(k+1)),\pi(\tau_2(k+1))] + \mathbb{1}[\pi(\tau_1(k+1)) > \pi(\tau_2(k+1))],$$
$$\forall k = 1,\ldots,\frac{n(n-1)}{2} - 1.$$

For each $\pi \in \Pi$ and $\tau \in \mathcal{T}^*$, we have one set of inequality conditions above, in the standard linear inequality form $\boldsymbol{A}\boldsymbol{\rho} \leq \boldsymbol{b}$, which corresponds to one simplex. Since $|\Pi| = n!$ and $|\mathcal{T}^*| = L$, this gives the form of all $\boldsymbol{G}_{k,\ell}$ and $\boldsymbol{b}_{k,\ell}$.

A.2 Proof of Lemma 1

Proof. We first show the following lemma.

Lemma 5. *Let P be a convex polytope with affine dimension n, defined as*

$$P = \{\boldsymbol{x} \in [0,1]^n \mid x_i \leq x_j, \ \forall (i,j) \in E\}, \tag{37}$$

where $E \subset \{1, 2, \ldots, n\}^2$ is some collection of index pairs. Then any extreme point of P must have all entries taking value in $\{0, 1\}$.

Proof. We know that at any extreme point $\boldsymbol{x} = [x_1, \ldots, x_n]^T$ of a convex polytope $P \subset \mathbb{R}^n$ with affine dimension n, there exist n linearly independent facet-defining inequalities of P that are attain equality. For a polytope defined as (37), the facet-defining inequalities must take the following two forms. Type-A constraints are those of the form $x_i \geq 0$ or $x_i \leq 1$. Type-B constraints are those of the form $x_i \leq x_j$. We now consider the n inequalities constraints that attain equality at extreme point $\boldsymbol{x}$.

A type-B constraint that attains equality implies $x_i = x_j$. Thus, all the type-B constraints that attain equality collectively partitions the index set $\{1, 2, \ldots, n\}$ into some equivalence classes, which we denote as $I_1, \ldots, I_K$. In other words, whenever two distinct indices i, j belong to the same equivalence class, it means they are connected by some type-B constraints and their values must be equal. Note that none of the constraints contain indices that belong to different equivalence class. This allows us to refer to every constraint as "corresponding to" an equivalence class.

For each equivalence class I_k, there can at least be $|I_k|$ linearly independent constraints corresponding to it, because this is the total number of variables concerning any such constraints. Also, this maximum number of $|I_k|$ linearly independent constraints is only achievable if there exists at least one type-A constraint, because if instead all constraints were type-B, the maximum possible rank of the constraints would be $|I_k| - 1$, since there is a degree of freedom in the x_i's values that cannot be pinned down by any type-B constraint.

In total across all equivalence class, we can have a maximum of $\sum_{k=1}^{K} |I_k| = n$ linearly independent constraints, with equality obtainable only when every equivalence class has at least one corresponding type-A constraint. This means that the indices in every equivalence class is pinned down to be equal to either 0 or 1, by the corresponding type-A constraint. Therefore, all entries of $\boldsymbol{x}$ must take value in $\{0, 1\}$. □

Now we prove the result of Lemma 1 regarding the collective WU/MU/SS polytopes. First consider

the specific rational deterministic choice vector $\boldsymbol{\rho}^* = [0, 0, \ldots, 0]^T \in \mathbb{R}^{\frac{n(n-1)}{2}}$, for which we have

$$\text{Cube}(\boldsymbol{\rho}^*) = [0, \frac{1}{2}]^{\frac{n(n-1)}{2}}. \tag{38}$$

As discussed previously, we can partition this small hypercube $\text{Cube}(\boldsymbol{\rho}^*)$ into a total of $\left(\frac{n(n-1)}{2}\right)!$ simplexes, each corresponding to one ordering of the $\frac{n(n-1)}{2}$ entries of $\boldsymbol{\rho}$. Each such simplex is a polytope that satisfies the condition of Lemma 5 (after scaling by 2), and thus it follows that every extreme point of these simplexes must have all entries taking value in $\{0, \frac{1}{2}\}$.

For any other rational deterministic choice vector $\boldsymbol{\rho}^*$, due to symmetry, its corresponding $\text{Cube}(\boldsymbol{\rho}^*)$ and the simplexes can be obtained by taking that of $[0, 0, \ldots, 0]^T$ and re-mapping the indices $\{1, \ldots, n\}$ according to the permutation that defines $\boldsymbol{\rho}^*$. Some entries of ρ will get reversed (from $\rho(x_i, x_j)$ to $\rho(x_j, x_i) = 1 - \rho(x_i, x_j)$) in this process. Therefore, the resulting extreme points of all simplexes from $\text{Cube}(\boldsymbol{\rho}^*)$ will have all entries taking value in $\{0, \frac{1}{2}, 1\}$.

Finally, because these simplexes constitute the smallest unit of the space that individual WU/MU/SS has implications over, we know that the rationalizable set for individual WU/MU/SS must be a finite union of such simplexes. Hence, the extreme points of the polytope (of individual rationalizability) will only take value $\{0, \frac{1}{2}, 1\}$.

$\square$

A.3 Proof of Theorem 1

Proof. For all three models, the main part of the proof for (i) $\iff$ (ii) has already been stated the main text, using the key fact that for any finite collection of sets $S_1, S_2, \ldots, S_k \subseteq \mathbb{R}^d$,

$$\text{conv}\left(\bigcup_{i=1}^{k}\text{conv}(S_i)\right) = \text{conv}\left(\bigcup_{i=1}^{k}S_i\right). \tag{39}$$

Furthermore, recall that the main text has provided characterization of the closure of the rationalizable set for individual WU/MU/SS. Here we again aim to characterize the closure of the rationalizable set of collective WU/MU/SS, noting that the collective characterization is the convex hull of the *original* (i.e. non-closure) individual WU/MU/SS. Hence we need to use the fact that for a finite number of convex sets, the closure of convex hull is equal to the convex hull of closure. In other words, for convex sets $C_1, \ldots, C_K$,

$$\text{cl}\left(\text{conv}\left(\bigcup_{k=1}^{K}C_k\right)\right) = \text{conv}\left(\bigcup_{k=1}^{K}\text{cl}\,(C_k)\right). \tag{40}$$

This implies that taking the convex hull of the closure of individual rationalizability set gives the closure of the collective rationalizability set.

It remains to prove (i) $\iff$ (iii). The proof for all three models follow the same argument, so we will illustrate WU as an example.

As discussed in the main text, the rationalizable set for individual WU can be represented as the union of $n!$ small hypercubes. This union can be modeled as $\boldsymbol{\rho}$ satisfying the following feasibility problem:

$$\begin{aligned}
\sum_{k=1}^{n!}\boldsymbol{\rho}_k &= \boldsymbol{\rho} \\
-\boldsymbol{\rho}_k &\leq x_k\left(\frac{1}{2}\mathbf{1}-\boldsymbol{\rho}_k^*\right) \quad k=1,\ldots,n! \\
\boldsymbol{\rho}_k &\leq x_k\left(\frac{1}{2}\mathbf{1}+\boldsymbol{\rho}_k^*\right) \quad k=1,\ldots,n! \\
\mathbf{0} &\leq \boldsymbol{\rho}_k \leq x_k\mathbf{1} \qquad k=1,\ldots,n! \\
\sum_{k=1}^{n!}x_k &= 1 \\
\boldsymbol{x} &\in \{0,1\}^{n!}.
\end{aligned} \tag{41}$$

Here we define $n!$ auxiliary indicator variables $x_k \in \{0,1\}$, indicating whether $\boldsymbol{\rho}$ actually belongs to the k-th hypercube. We also define $n!$ copies of the $\boldsymbol{\rho}$ vector as $\boldsymbol{\rho}_k$, each lying in one hypercube and corresponding to one indicator x_k.

With the help of techniques from integer programming models and modeling disjunction (Balas,

1971), the convex hull of solutions to the linear system (41) is simply obtained by dropping the integrality restriction. Hence, we have $\boldsymbol{\rho} \in \text{conv}\left(\bigcup_{k=1}^{n!} \text{cl}\left(\text{Cube}(\boldsymbol{\rho}_k^*)\right)\right)$ iff

$$
\begin{aligned}
\sum_{k=1}^{n!} \boldsymbol{\rho}_k &= \boldsymbol{\rho} \\
-\boldsymbol{\rho}_k &\leq x_k \left(\frac{1}{2}\mathbf{1} - \boldsymbol{\rho}_k^*\right) \quad k = 1, \ldots, n! \\
\boldsymbol{\rho}_k &\leq x_k \left(\frac{1}{2}\mathbf{1} + \boldsymbol{\rho}_k^*\right) \quad k = 1, \ldots, n! \\
\mathbf{0} \leq \boldsymbol{\rho}_k &\leq x_k \mathbf{1} \qquad\quad k = 1, \ldots, n! \\
\sum_{k=1}^{n!} x_k &= 1 \\
\boldsymbol{x} &\in [0, 1]^{n!}.
\end{aligned}
\tag{42}
$$

This is exactly the form in Theorem 1 (iii).

The derivation for MU and SS are entirely analogous, with the only difference being the different inequality constraints that characterize individual rationalizability. Therefore we avoid repetition here.

$\square$

A.4 Extra Result on Theorem 1

Lemma 6. *For a choice vector $\boldsymbol{\rho} \in [0, 1]^m$, the following are equivalent:*

1. *$\boldsymbol{\rho}$ is rationalizable by collective WU.*

2. *The following inequality holds for any $\boldsymbol{\gamma} \in \mathbb{R}^{\frac{n(n-1)}{2}}$:*

$$(\boldsymbol{\rho} - \frac{1}{2}\mathbf{1})^T \boldsymbol{\gamma} \geq -\frac{1}{2} \sum_{j\,:\,\mathrm{sgn}(\gamma_j)=\mathrm{sgn}(\rho_k^*-\frac{1}{2})} |\gamma_j| \tag{43}$$

   where $k \in \{1, \ldots, n!\}$ is such that

$$k = \arg\min_{k'} \sum_{j\,:\,\mathrm{sgn}(\rho_{k'j}^*-\frac{1}{2})\neq\mathrm{sgn}(\gamma_j)} |\gamma_j|. \tag{44}$$

Proof. This Lemma provides an additional characterization for Collective WU, aimed at deriving its exact facet-defining inequalities. This is currently a partial result in that endeavor.

Writing all inequality constraints from (c) into a full matrix form, we want to check whether there exists a vector $\boldsymbol{v}$ such that

$$\boldsymbol{A}\boldsymbol{v} \leq \boldsymbol{b}$$
$$\boldsymbol{A}'\boldsymbol{v} = \boldsymbol{b}'$$

where

$$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{\rho}_1 \\ \vdots \\ \boldsymbol{\rho}_{n!} \\ x_1 \\ \vdots \\ x_{n!} \end{bmatrix}, \quad \boldsymbol{A}' = \left[\begin{array}{ccc|c} \boldsymbol{I}_{\frac{n(n-1)}{2}} & \cdots & \boldsymbol{I}_{\frac{n(n-1)}{2}} & \\ \hline & & & \mathbf{1}_{n!}^T \end{array} \right], \quad \boldsymbol{b}' = \begin{bmatrix} \boldsymbol{\rho} \\ 1 \end{bmatrix} \tag{45}$$

$$
\boldsymbol{A} = \left[\begin{array}{c|c}
\boldsymbol{I}_{n!\frac{n(n-1)}{2}} & \begin{matrix} -\frac{1}{2}\mathbf{1}_{\frac{n(n-1)}{2}} - \boldsymbol{\rho}_1^* & & \\ & \ddots & \\ & & -\frac{1}{2}\mathbf{1}_{\frac{n(n-1)}{2}} - \boldsymbol{\rho}_{n!}^* \end{matrix} \\
\hline
-\boldsymbol{I}_{n!\frac{n(n-1)}{2}} & \begin{matrix} -\frac{1}{2}\mathbf{1}_{\frac{n(n-1)}{2}} + \boldsymbol{\rho}_1^* & & \\ & \ddots & \\ & & -\frac{1}{2}\mathbf{1}_{\frac{n(n-1)}{2}} + \boldsymbol{\rho}_{n!}^* \end{matrix} \\
\hline
-\boldsymbol{I}_{n!\frac{n(n-1)}{2}} & \\
\hline
\boldsymbol{I}_{n!\frac{n(n-1)}{2}} & \begin{matrix} -\mathbf{1}_{\frac{n(n-1)}{2}} & & \\ & \ddots & \\ & & -\mathbf{1}_{\frac{n(n-1)}{2}} \end{matrix} \\
\hline
& -\boldsymbol{I}_{n!} \\
\hline
& \boldsymbol{I}_{n!}
\end{array}\right], \quad
\boldsymbol{b} = \left[\begin{array}{c} \mathbf{0} \\ \hline \mathbf{0} \\ \hline \mathbf{0} \\ \hline \mathbf{0} \\ \hline \mathbf{0} \\ \hline \mathbf{1}_{n!} \end{array}\right]
\tag{46}
$$

From Farka's lemma, this is equivalent to there not existing $\boldsymbol{\lambda} \geq \mathbf{0} \in \mathbb{R}_+^{n!(2n(n-1)+2)}$ and $\boldsymbol{\lambda}' \in \mathbb{R}^{\frac{n(n-1)}{2}+1}$ that satisfies

$$
\boldsymbol{A}^T \boldsymbol{\lambda} + \boldsymbol{A}'^T \boldsymbol{\lambda}' = \mathbf{0} \tag{47}
$$

$$
\boldsymbol{b}^T \boldsymbol{\lambda} + \boldsymbol{b}'^T \boldsymbol{\lambda}' < 0. \tag{48}
$$

Let

$$
\boldsymbol{\lambda} = \begin{bmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \boldsymbol{\lambda}_3 \\ \boldsymbol{\lambda}_4 \\ \boldsymbol{\lambda}_5 \\ \boldsymbol{\lambda}_6 \end{bmatrix}, \quad \boldsymbol{\lambda}' = \begin{bmatrix} \lambda'_1 \\ \lambda'_2 \end{bmatrix} \tag{49}
$$

where $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3, \boldsymbol{\lambda}_4 \in \mathbb{R}^{n!\frac{n(n-1)}{2}}$, $\boldsymbol{\lambda}_5, \boldsymbol{\lambda}_6 \in \mathbb{R}^{n!}$, $\boldsymbol{\lambda}'_1 \in \mathbb{R}^{\frac{n(n-1)}{2}}$, $\lambda'_2 \in \mathbb{R}$. Further, for $i = 1, 2, 3, 4$ write $\boldsymbol{\lambda}_i = [\boldsymbol{\lambda}_{ik}]_{k=1}^{n!}$ with $\boldsymbol{\lambda}_{ik} \in \mathbb{R}^{\frac{n(n-1)}{2}}$, and for $i = 5, 6$ write $\boldsymbol{\lambda}_i = [\lambda_{ik}]_{k=1}^{n!}$ with $\lambda_{ik} \in \mathbb{R}$.

The conditions (47) (48) can be written as

$$\boldsymbol{\lambda}_{1k} - \boldsymbol{\lambda}_{2k} - \boldsymbol{\lambda}_{3k} + \boldsymbol{\lambda}_{4k} + \boldsymbol{\lambda}_1' = \mathbf{0}, \quad \forall 1 \leq k \leq n! \qquad (50)$$

$$\left(-\frac{1}{2}\mathbf{1} - \boldsymbol{\rho}_k^*\right)^T \boldsymbol{\lambda}_{1k} + \left(-\frac{1}{2}\mathbf{1} + \boldsymbol{\rho}_k^*\right)^T \boldsymbol{\lambda}_{2k} - \mathbf{1}^T \boldsymbol{\lambda}_{4k} - \lambda_{5k} + \lambda_{6k} + \lambda_2' = 0, \quad \forall 1 \leq k \leq n! \qquad (51)$$

$$\mathbf{1}^T \boldsymbol{\lambda}_6 + \boldsymbol{\rho}^T \boldsymbol{\lambda}_1' + \lambda_2' < 0 \qquad (52)$$

We now simplify the system above by variable elimination.

$\boldsymbol{\lambda}_{3k}$ only appears in (50), and additionally we require $\boldsymbol{\lambda}_{3k} \geq \mathbf{0}$, hence we can eliminate $\boldsymbol{\lambda}_{3k}$ and (50) is equivalent to

$$\boldsymbol{\lambda}_{1k} - \boldsymbol{\lambda}_{2k} + \boldsymbol{\lambda}_{4k} + \boldsymbol{\lambda}_1' \geq \mathbf{0}, \quad \forall 1 \leq k \leq n! \qquad (53)$$

λ_{5k} only appears in (51), and additionally we require $\lambda_{5k} \geq 0$, hence we can eliminate λ_{5k} and (51) is equivalent to

$$\left(-\frac{1}{2}\mathbf{1} - \boldsymbol{\rho}_k^*\right)^T \boldsymbol{\lambda}_{1k} + \left(-\frac{1}{2}\mathbf{1} + \boldsymbol{\rho}_k^*\right)^T \boldsymbol{\lambda}_{2k} - \mathbf{1}^T \boldsymbol{\lambda}_{4k} + \lambda_{6k} + \lambda_2' \geq 0, \quad \forall 1 \leq k \leq n! \qquad (54)$$

By comparing λ_2' in conditions (52) and (54), we can eliminate λ_2' and combine these two conditions into the equivalent form

$$\left(-\frac{1}{2}\mathbf{1} - \boldsymbol{\rho}_k^*\right)^T \boldsymbol{\lambda}_{1k} + \left(-\frac{1}{2}\mathbf{1} + \boldsymbol{\rho}_k^*\right)^T \boldsymbol{\lambda}_{2k} - \mathbf{1}^T \boldsymbol{\lambda}_{4k} + \lambda_{6k} > \mathbf{1}^T \boldsymbol{\lambda}_6 + \boldsymbol{\rho}^T \boldsymbol{\lambda}_1', \quad \forall 1 \leq k \leq n! \qquad (55)$$

Now the conditions have been transformed to (53) (55). λ_{6k} only appears in (55), and additionally we require $\lambda_{6k} \geq 0$, hence we can eliminate λ_{6k} and (55) is equivalent to

$$\left(-\frac{1}{2}\mathbf{1} - \boldsymbol{\rho}_k^*\right)^T \boldsymbol{\lambda}_{1k} + \left(-\frac{1}{2}\mathbf{1} + \boldsymbol{\rho}_k^*\right)^T \boldsymbol{\lambda}_{2k} - \mathbf{1}^T \boldsymbol{\lambda}_{4k} > \boldsymbol{\rho}^T \boldsymbol{\lambda}_1', \quad \forall 1 \leq k \leq n! \qquad (56)$$

By comparing $\boldsymbol{\lambda}_{4k}$ in (53) and (56), we can eliminate $\boldsymbol{\lambda}_{4k}$ by setting $\boldsymbol{\lambda}_{4k} = \max\left(\boldsymbol{\lambda}_{2k} - \boldsymbol{\lambda}_{1k} - \boldsymbol{\lambda}_1', \mathbf{0}\right)$ in (55). Hence the system of conditions (50)-(52) is equivalent to the following one inequality constraint:

$$\left(-\frac{1}{2}\mathbf{1} - \boldsymbol{\rho}_k^*\right)^T \boldsymbol{\lambda}_{1k} + \left(-\frac{1}{2}\mathbf{1} + \boldsymbol{\rho}_k^*\right)^T \boldsymbol{\lambda}_{2k} - \mathbf{1}^T \max\left(\boldsymbol{\lambda}_{2k} - \boldsymbol{\lambda}_{1k} - \boldsymbol{\lambda}_1', \mathbf{0}\right) > \boldsymbol{\rho}^T \boldsymbol{\lambda}_1', \quad \forall 1 \leq k \leq n! \qquad (57)$$

Define a change of variables $\boldsymbol{\alpha}_k = \boldsymbol{\lambda}_{1k} + \boldsymbol{\lambda}_{2k}$ and $\boldsymbol{\beta}_k = \boldsymbol{\lambda}_{2k} - \boldsymbol{\lambda}_{1k}$, $\forall 1 \leq k \leq n!$. The constraints

$\boldsymbol{\lambda}_{1k}, \boldsymbol{\lambda}_{2k} \geq \mathbf{0}$ is equivalent to $\boldsymbol{\alpha}_k \geq |\boldsymbol{\beta}_k|$. Plugging this into (57), and also rename $\boldsymbol{\lambda}'_1$ as $\boldsymbol{\gamma}$, we get

$$-\frac{1}{2}\mathbf{1}^T \boldsymbol{\alpha}_k + \boldsymbol{\rho}_k^{*T} \boldsymbol{\beta}_k - \mathbf{1}^T \max(\boldsymbol{\beta}_k - \boldsymbol{\gamma},\ \mathbf{0}) > \boldsymbol{\rho}^T \boldsymbol{\gamma}, \quad \forall 1 \leq k \leq n! \tag{58}$$

We can eliminate $\boldsymbol{\alpha}_k$ by taking $\boldsymbol{\alpha}_k = |\boldsymbol{\beta}_k|$, and the condition is equivalent to

$$-\frac{1}{2}\mathbf{1}^T|\boldsymbol{\beta}_k| + \boldsymbol{\rho}_k^{*T} \boldsymbol{\beta}_k - \mathbf{1}^T \max(\boldsymbol{\beta}_k - \boldsymbol{\gamma},\ \mathbf{0}) > \boldsymbol{\rho}^T \boldsymbol{\gamma} \quad \forall 1 \leq k \leq n! \tag{59}$$

Let β_{kj} be the j-th entry of $\boldsymbol{\beta}_k$, $1 \leq j \leq \frac{n(n-1)}{2}$. The LHS of (59) is a piecewise linear function of each individual β_{kj}, we can find the β_{kj} values that maximizes LHS as

$$\beta_{kj} = \begin{cases} \max(\gamma_j, 0) & \text{if } \rho_{kj}^* = 1 \\ \min(\gamma_j, 0) & \text{if } \rho_{kj}^* = 0 \end{cases} \tag{60}$$

This further simplifies to

$$\beta_{kj} = \begin{cases} \gamma_j & \text{if } \operatorname{sgn}(\rho_{kj}^* - \frac{1}{2}) = \operatorname{sgn}(\gamma_j) \\ 0 & \text{otherwise} \end{cases} \tag{61}$$

We can eliminate $\boldsymbol{\beta}_k$ by plugging this maximizer into (59), and get

$$-\frac{1}{2}\sum_{J_k^{++}} \gamma_j + \frac{1}{2}\sum_{J_k^{--}} \gamma_j + \sum_{J_k^{-+}} \gamma_j + \sum_{J_k^{++} \cup J_k^{--}} \rho_{kj}^* \gamma_j > \sum_j \rho_j \gamma_j, \quad \forall 1 \leq k \leq n! \tag{62}$$

where we define index subsets

$$J_k^{s_1 s_2} = \left\{ 1 \leq j \leq \frac{n(n-1)}{2} \ \middle|\ \operatorname{sgn}(\gamma_j) = s_1,\ \operatorname{sgn}(\rho_{kj}^* - \tfrac{1}{2}) = s_2 \right\}, \quad s_1, s_2 \in \{+, -\}. \tag{63}$$

Because $\rho_{kj}^* \in \{0, 1\}$ are known given the index subset $J_k^{s_1 s_2}$ (determined by the sign s_2), (62) can further be organized into

$$\sum_{j \in J_k^{++}} (\tfrac{1}{2} - \rho_j)\gamma_j + \sum_{j \in J_k^{--}} (\tfrac{1}{2} - \rho_j)\gamma_j + \sum_{j \in J_k^{+-}} (-\rho_j)\gamma_j + \sum_{j \in J_k^{-+}} (1 - \rho_j)\gamma_j > 0, \quad \forall 1 \leq k \leq n! \tag{64}$$

Up to now, we have shown that a vector $\boldsymbol{\rho} \in \mathbb{R}^{\frac{n(n-1)}{2}}$ lies in $\overline{\operatorname{conv}(S)}$ (i.e. is population-WST) iff for any $\boldsymbol{\gamma} \in \mathbb{R}^{\frac{n(n-1)}{2}}$, there exists $1 \leq k \leq n!$ such that

$$\sum_{j \in J_k^{++}} \tfrac{1}{2}\gamma_j + \sum_{j \in J_k^{--}} \tfrac{1}{2}\gamma_j + \sum_{j \in J_k^{-+}} \gamma_j \leq \sum_j \rho_j \gamma_j. \tag{65}$$

In the following, we characterize which k among the $n!$ possible choices minimizes LHS of (65), which is the only k we will need to consider.

The admissible values of $\boldsymbol{\rho}_k^*$ is a strict subset of $\{0,1\}^{\frac{n(n-1)}{2}}$, but for now imagine a relaxation where we allow any $\boldsymbol{\rho}^* \in \{0,1\}^{\frac{n(n-1)}{2}}$. In this case, we can determine the optimal ρ_j^* for each index j. For a given j, if $\gamma_j > 0$ then LHS of (65) is minimized when $\rho_j^* = 0$; if $\gamma_j < 0$ then LHS of (65) is minimized when $\gamma_j > 0$; if $\gamma_j = 0$ then ρ_j^* does not matter. In summary, the optimal $\boldsymbol{\rho}^* \in \{0,1\}^{\frac{n(n-1)}{2}}$ that minimizes LHS of (65) should satisfy $\operatorname{sgn}(\rho_j^* - \frac{1}{2}) = -\operatorname{sgn}(\gamma_j)$ whenever $\gamma_j \neq 0$.

Now we return to the actual scenario where $\boldsymbol{\rho}^*$ is only allowed to take a more restricted set of values $P^* = \{\boldsymbol{\rho}_k^*\}_{k=1}^{n!} \subsetneq \{0,1\}^{\frac{n(n-1)}{2}}$. For a given $\boldsymbol{\rho}_k^*$, for each j where $\operatorname{sgn}(\rho_{kj}^* - \frac{1}{2})$ and $-\operatorname{sgn}(\gamma_j)$ differ, LHS of (65) increases by $\frac{1}{2}|\gamma_j|$. The optimal k is one that minimizes this total increase, which is

$$k^* = \arg\min_k \sum_{j \,:\, \operatorname{sgn}(\rho_{kj}^* - \frac{1}{2}) = \operatorname{sgn}(\gamma_j)} |\gamma_j|. \tag{66}$$

We know that $\boldsymbol{\rho}^* \in P^* \iff \mathbf{1} - \boldsymbol{\rho}^* \in P^*$, and can let $\widehat{k}$ be such that $\boldsymbol{\rho}_{\widehat{k}}^* = \mathbf{1} - \boldsymbol{\rho}_{k^*}^*$, namely the "reverse ordering" of k^*. Then we have

$$\widehat{k} = \arg\min_k \sum_{j \,:\, \operatorname{sgn}(\rho_{kj}^* - \frac{1}{2}) \neq \operatorname{sgn}(\gamma_j)} |\gamma_j|. \tag{67}$$

This has a very intuitive interpretation. For a given vector $\boldsymbol{\gamma}$, imagine making minimal changes (in the L_1 sense) to its elements such that the sign pattern aligns with some WST sign pattern. $\boldsymbol{\rho}_k^*$ is the WST ordering "nearest" to $\boldsymbol{\gamma}$. Also notably, $\widehat{k}$ depends only on $\boldsymbol{\gamma}$ and is independent of $\boldsymbol{\rho}$.

Using the fact that k^* and $\widehat{k}$ give exactly the reverse signs, we can rewrite the condition as follow. A choice rate $\boldsymbol{\rho} \in \mathbb{R}^{\frac{n(n-1)}{2}}$ is collective WUM-rationalizable iff for any $\boldsymbol{\gamma} \in \mathbb{R}^{\frac{n(n-1)}{2}}$, we have

$$(\boldsymbol{\rho} - \frac{1}{2}\mathbf{1})^T \boldsymbol{\gamma} \geq -\frac{1}{2} \sum_{j \,:\, \operatorname{sgn}(\gamma_j) = \operatorname{sgn}(\rho_k^* - \frac{1}{2})} |\gamma_j|, \tag{68}$$

where $k \in \{1, \ldots, n!\}$ is such that

$$k = \arg\min_{k'} \sum_{j \,:\, \operatorname{sgn}(\rho_{k'j}^* - \frac{1}{2}) \neq \operatorname{sgn}(\gamma_j)} |\gamma_j|. \tag{69}$$

$\square$

A.5   Additional Results on General Non-Binary Menus

So far, I have focused on stochastic choice over binary menus in Z^2 over a choice set Z. Meanwhile, some of the random choice models we discussed can also be generalized to non-binary menus, namely the family of all subsets 2^Z. Here I take Simple Scalability as an example, and discuss the application of collective rationalizability to Simple Scalability over general menus.

A natural generalization of Simple Scalability to general menus is given in (Tversky, 1972a). Let $\rho(x, A)$ denote the probability of choosing $x \in A$ in a menu $A \subseteq Z$. Let u and F_a, $2 \leq a \leq n$, be real-valued functions, then a choice rule satisfies Simple Scalability if for any choice option in any choice subset $A = \{x, y, \ldots, z\} \subseteq Z$, we have

$$\rho(x, A) = F_a(u(x), u(y), \ldots, u(z)), \tag{70}$$

with F_a symmetric in its last $a - 1$ arguments, and satisfying: F_a is strictly increasing in the first argument and strictly decreasing in the remaining $a - 1$ arguments provided $\rho(x, A) \neq 0, 1$. If $\rho(x, A)$ is 0 or 1, F_a is non-decreasing in the first argument and non-increasing in the others.

Generalized to multiple choices, Simple Scalability asserts that the alternatives can be scaled so that choice probability is represented as a monotonic function of the respective scale values. Simple Scalability with multiple choices is equivalent to the *order-independence* property (Tversky, 1972a), which requires that for any two menus $A, B \subseteq Z$, $\forall x, y \in B \backslash A$, $\forall z \in A$, we have

$$\rho(x, B) \geq \rho(y, B) \iff \rho(z, A \cup \{x\}) \leq \rho(z, A \cup \{y\}), \tag{71}$$

provided the choice probabilities on the two sides of either inequality are not 0 or 1. This characterization provides a directly testable condition for Simple Scalability over general menus.

Thus, following the procedure developed in this paper, one can obtain the collective rationalizability characterization by taking the convex hull of the individual rationalizable set. Below, I provide a concrete example for the case of three options $Z = \{x, y, z\}$.

With the inclusion of general non-binary menus, a choice rule ρ now corresponds to a 5-dimensional choice vector:

$$\boldsymbol{\rho} = [\rho(x, \{x, y\}), \ \rho(x, \{x, z\}), \ \rho(y, \{y, z\}), \ \rho(x, \{x, y, z\}), \ \rho(y, \{x, y, z\})]^T. \tag{72}$$

Similar to the characterization of individual SS with binary menus, here the rationalizable set can again be written as the union of $3! = 6$ convex polytopes, each of which is 5-dimensional. For example, for the convex polytope corresponding to the preference ordering $x \succ y \succ z$, in addition to the constraints $\rho(x, \{x, z\}) \geq \rho(x, \{x, y\})$ and $\rho(x, \{x, z\}) \geq \rho(y, \{y, z\})$ from strong stochastic transitivity, we furthermore have constraints $\rho(x, \{x, y, z\}) \geq \rho(y, \{x, y, z\})$ and $\rho(y, \{x, y, z\}) \geq 1 - \rho(x, \{x, y, z\}) - \rho(y, \{x, y, z\})$. Finally, applying the convex hull operation on the half-space

representation, one can readily obtain the characterization for collective SS over general menus.

Interestingly, aggregating choices from heterogeneous individuals under representative agent assumption can generate *choice reversal*, even when each individual satisfies Simple Scalability.

Example 3. Consider two individuals with choice probabilities

$$\boldsymbol{\rho}_1 = [\rho_1(x, \{x,y\}), \ \rho_1(x, \{x,z\}), \ \rho_1(y, \{y,z\}), \ \rho_1(x, \{x,y,z\}), \ \rho_1(y, \{x,y,z\})]^T = [0.2, 0.6, 0.9, 0.3, 0.6]^T,$$

$$\boldsymbol{\rho}_2 = [\rho_2(x, \{x,y\}), \ \rho_2(x, \{x,z\}), \ \rho_2(y, \{y,z\}), \ \rho_2(x, \{x,y,z\}), \ \rho_2(y, \{x,y,z\})]^T = [0.7, 0.6, 0.4, 0.8, 0.05]^T.$$

It is straightforward to verify that both individuals satisfy Simple Scalability, with Subject 1 having the preference ordering $y \succ x \succ z$, while Subject 2 $x \succ z \succ y$. However, aggregating their choice probabilities with equal weights yields

$$\rho(y, \{x,y\}) > 0.5, \ \ \rho(x, \{x,y,z\}) > 0.5.$$

This is a choice reversal that y is chosen more often than x in the binary menu $\{x,y\}$, but x is chosen more than 50% in the menu $\{x,y,z\}$.

Importantly, this choice reversal violates regularity: adding alternative z increases x's choice probability from 0.45 to 0.55. Such regularity violations cannot arise from heterogeneity in random utility models where individuals choose deterministically, but can emerge from aggregating stochastic choices that individually satisfy Simple Scalability with heterogeneous tastes.

A.6   Proof of Proposition 1

Proof. The loop choice can be rationalized by WU iff we can find the comparison difficulty $\{c_j\}_{j=1}^m$ and a function F such that

$$\sum_{j\in J^+} c_j F^{-1}(\rho_j) = -\sum_{j'\in J^-} c_{j'} F^{-1}(\rho_{j'}). \tag{73}$$

Here F^{-1} can be any strictly increasing function with $F^{-1}(\frac{1}{2}) = 0$ and $|F^{-1}(x)| \geq |F^{-1}(y)| \iff |x - \frac{1}{2}| \geq |y - \frac{1}{2}|$. Hence (73) can be rewritten as

$$\sum_{j\in J^+} c_j G(\mu_j) = \sum_{j'\in J^-} c_{j'} G(\mu_{j'}), \tag{74}$$

where G can be any strictly increasing function with $G(0) = 0$.

We first prove part 1 of the result. Because J^+ and J^- do not ranking-dominate each other, it follows that either strict inequality in (19) in both directions holds for some (s, t), or that equality holds for all (s, t). That latter immediately implies that J^+ and J^- have the identical set of values, and hence an arbitrary function G with arbitrary constant $\{c_j\}_{j=1}^m$ sequence will satisfy (74). In the following, we consider the other case where strict inequality in both directions must hold for some (s, t).

Pick an arbitrary function G (strictly increasing with $G(0) = 0$) and an arbitrary sequence of values of $\{c_j\}_{j=1}^m$ that satisfies the ordinal ranking π. If this arbitrary choice does not satisfy equality in (74), suppose we have $\sum_{j\in J^+} c_j G(\mu_j) > \sum_{j'\in J^-} c_{j'} G(\mu_{j'})$. Our assumptions imply that there exist some $s, t \in \{1, \ldots, m\}$ such that

$$\#\{j \in J^+ \mid \sigma(j) \leq s, \pi(j) \leq t\} \;<\; \#\{j \in J^- \mid \sigma(j) \leq s, \pi(j) \leq t\}. \tag{75}$$

Consider the following operation: for a given a constant $M > 0$, increase all values c_j with $\pi(j) \leq t$ by M, and increase all function values $G(\mu_j)$ with $\sigma(j) \leq s$ by M. The former preserves the ranking π of $\{c_j\}_{j=1}^m$, and the latter can be done in a way that preserves the monotonicity and continuity of G because σ is the ranking function of $\{\mu_j\}$. Define

$$\delta(M) = \sum_{j\in J^+} \big(c_j + M\mathbb{1}[\pi(j) \leq t]\big)\big(G(\mu_j) + M\mathbb{1}[\sigma(j) \leq s]\big) -$$
$$\sum_{j\in J^-} \big(c_j + M\mathbb{1}[\pi(j) \leq t]\big)\big(G(\mu_j) + M\mathbb{1}[\sigma(j) \leq s]\big)$$

as the difference between the two sides of (74) after the operation. We have $\delta(0) > 0$ by our assumption. As M becomes sufficiently large, $\delta(M)$ is dominated by the quadratic term of M, and

(75) implies that the coefficient of the quadratic term is negative. Hence for sufficiently large M we have $\delta(M) < 0$, and by continuity there must exist some M such that $\delta(M) = 0$. Therefore the resulting modified $\{c_j\}_{j=1}^m$ and G satisfied the desired equation (74), and hence the data can be rationalized by RUM with the ranking of comparison difficulty π.

The argument is identical if instead the initial arbitrary choice $\{c_j\}_{j=1}^m$ gives the opposite sign.

We then prove part 2 of the result. Assume the choice rates $\boldsymbol{\rho}$ can be rationalized by a sequence of complexities $\{c_j\}_{j=1}^m$ following comparison difficulty ranking π and a strictly increasing function G with $G(0) = 0$. The rationalization condition (74) can be rewritten as

$$\Delta = \sum_{j=1}^m c_j G(\mu_j)\delta_j = 0 \tag{76}$$

where we define

$$\delta_j = \mathbb{1}[j \in J^+] - \mathbb{1}[j \in J^-], \quad j = 1, \ldots, m. \tag{77}$$

Now, assume that one set ranking-dominates the other — without loss of generality, assume J^+ ranking-dominates J^- regarding (σ, π).

Because we assumed the choice magnitudes μ_j are distinct except $\frac{1}{2}$, and because G is strictly increasing, the values $G(\mu_j)$ are also distinct and follow the same ranking. These values $G(\mu_j)$ can be sorted in strictly descending order: $y_1 > y_2 > \cdots > y_m > 0$. Each value y_l corresponds to a unique problem index j such that $y_l = G(\mu_j)$. The ranking function σ for magnitudes is thus equivalent to this ordering, i.e., $\sigma(j) = l$ if $G(\mu_j) = y_l$.

The complexities $\{c_j\}_{j=1}^m$ however may contain ties. Let the set of *distinct* comparison difficulty values be $C_1 > C_2 > \cdots > C_K > 0$, where $K \leq m$. For each distinct value C_k, let $I_k = \{j \mid c_j = C_k\}$ be the set of indices corresponding to that comparison difficulty tier. The comparison difficulty ranking π is respected, meaning if $j \in I_k$ and $j' \in I_{k'}$, then $\pi(j) < \pi(j')$ if and only if $k < k'$.

We can rewrite the sum Δ in (76) by grouping terms by their comparison difficulty tier:

$$\Delta = \sum_{k=1}^K \sum_{j \in I_k} C_k G(\mu_j)\delta_j = \sum_{k=1}^K C_k Z_k = 0, \tag{78}$$

where we define $Z_k = \sum_{j \in I_k} G(\mu_j)\delta_j$. Let $W_k = \sum_{i=1}^k Z_i$ be the cumulative sum of the Z_i terms.

We apply Abel's summation formula (summation by parts) to the sum $\sum C_k Z_k$:

$$\Delta = \sum_{k=1}^{K} C_k(W_k - W_{k-1}) \quad \text{(with } W_0 = 0)$$
$$= C_1 W_1 + C_2(W_2 - W_1) + \cdots + C_K(W_K - W_{K-1})$$
$$= W_1(C_1 - C_2) + W_2(C_2 - C_3) + \cdots + W_{K-1}(C_{K-1} - C_K) + W_K C_K = 0. \tag{79}$$

By construction, each comparison difficulty difference $(C_k - C_{k+1})$ is strictly positive, and $C_K > 0$.

Next, we analyze the sign of the partial sums W_k. The term W_k represents the net $G(\mu)$-weighted difference over all choice problems whose comparison difficulty c_j is at or above tier k (i.e., $c_j \geq C_k$). Let $S_k = \bigcup_{i=1}^{k} I_i$ be this set of indices. Then

$$W_k = \sum_{j \in S_k} G(\mu_j)\delta_j. \tag{80}$$

To analyze this sum, we apply summation by parts again, this time over the strictly ordered magnitudes $y_l = G(\mu_j)$. We re-order the summation by magnitude rank $l = 1, \ldots, m$:

$$W_k = \sum_{l=1}^{m} y_l \cdot \delta_{j_l} \cdot \mathbb{1}[j_l \in S_k] \quad \text{(where } G(\mu_{j_l}) = y_l)$$
$$= \sum_{l=1}^{m} \left(\sum_{t=l}^{m-1} (y_t - y_{t+1}) + y_m \right) \delta_{j_l} \cdot \mathbb{1}[j_l \in S_k]$$
$$= \sum_{t=1}^{m-1} (y_t - y_{t+1}) \left(\sum_{l=1}^{t} \delta_{j_l} \cdot \mathbb{1}[j_l \in S_k] \right) + y_m \left(\sum_{l=1}^{m} \delta_{j_l} \cdot \mathbb{1}[j_l \in S_k] \right). \tag{81}$$

The inner sum $\sum_{l=1}^{t} \delta_{j_l} \cdot \mathbb{1}[j_l \in S_k]$ is the sum of δ_j over all items whose comparison difficulty is in the top k tiers and whose magnitude is in the top t ranks. This is precisely the cumulative difference from the definition of ranking-domination. Let's define this as $D_{k,t}$:

$$D_{k,t} = \#\{j \in J^+ \mid \pi(j) \leq \pi_k, \sigma(j) \leq t\} - \#\{j \in J^- \mid \pi(j) \leq \pi_k, \sigma(j) \leq t\}, \tag{82}$$

where π_k is the rank corresponding to comparison difficulty tier C_k. The ranking-domination assumption states that $D_{k,t} \geq 0$ for all k, t, and there exists at least one pair (k^*, t^*) for which $D_{k^*,t^*} > 0$.

Substituting this into (81), we get:

$$W_k = \sum_{t=1}^{m-1} (y_t - y_{t+1})D_{k,t} + y_m D_{k,m}. \tag{83}$$

Since magnitudes are distinct, we have $y_t - y_{t+1} > 0$, and also $y_m > 0$ from our assumption that

choice magnitudes are not equal to $\frac{1}{2}$. As $D_{k,t} \geq 0$ for all t, every term in this sum is non-negative. Thus, $W_k \geq 0$ for all $k = 1, \ldots, K$.

Now we find our contradiction. From (79), we have a sum of non-negative terms that equals zero:

$$\underbrace{W_1}_{\geq 0} \underbrace{(C_1 - C_2)}_{>0} + \cdots + \underbrace{W_{K-1}}_{\geq 0} \underbrace{(C_{K-1} - C_K)}_{>0} + \underbrace{W_K}_{\geq 0} \underbrace{C_K}_{>0} = 0.$$

This can only be true if every term is zero, which implies $W_k = 0$ for all $k = 1, \ldots, K$.

However, the ranking-domination assumption guarantees the existence of a pair (k^*, t^*) such that $D_{k^*, t^*} > 0$. Let's examine W_{k^*}:

$$W_{k^*} = \sum_{t=1}^{m-1} (y_t - y_{t+1}) D_{k^*, t} + y_m D_{k^*, m}.$$

This is a sum of non-negative terms. The term corresponding to t^* will contain the factor $D_{k^*, t^*} > 0$ multiplied by a strictly positive coefficient. Therefore, the sum must be strictly positive, i.e., $W_{k^*} > 0$.

We have deduced from the rationalization assumption that $W_{k^*} = 0$, and from the ranking-domination assumption that $W_{k^*} > 0$, which leads to a contradiction. Therefore, if J^+ ranking-dominates J^-, the data cannot be rationalized. The same logic applies if J^- ranking-dominates J^+. This completes the proof. $\square$

A.7 Additional Results on Rationalizability with Incomplete Data

It has been well established in the literature that Simple Scalability, Moderate Utility, and Weak Utility representations are equivalent to *strong, moderate, and weak stochastic transitivity* respectively when all binary choices are observed (Definition 2). However, testing stochastic transitivity requires complete data, i.e., observing all $\binom{n}{2}$ binary pairs for n options. With just 20 options, this means 190 binary pairs for each individual, a requirement hard to meet in practice.

This part provides necessary and sufficient conditions that characterize the three models SS/MU/WU when only a subset of choice pairs $\mathcal{A} \subseteq Z^2$ is observed. The key insight is extending stochastic transitivity, which examines all triples of options, to *stochastic acyclicity*, which examines sequences of any length in the observed data. We define this as a *loop*.

Formally, consider a sequence of $n \geq 3$ distinct options $(x_1, x_2, \ldots, x_n)$. A *loop* is the collection of n binary pairs, consisting of all adjacent pairs (x_k, x_{k+1}) for $1 \leq k \leq n-1$, as well as the pair (x_1, x_n) to "close the loop". The definition of stochastic transitivity can be naturally extended to such loops, which we define as stochastic acyclicity as follows:

Definition 8 (Stochastic Acyclicity).

A fully observed loop $\{(x_k, x_{k+1})\}_{k=1}^n$ (with indices taken modulo n) satisfies

1. *Strong stochastic acyclicity if $\forall k \in \{1, \ldots, n\}$*

$$\min_{1 \leq l \leq n, \, l \neq k} \rho(x_l, x_{l+1}) \geq (>) \frac{1}{2} \quad \text{implies} \quad \rho(x_{k+1}, x_k) \geq (>) \max_{1 \leq l \leq n, \, l \neq k} \rho(x_l, x_{l+1}). \quad (84)$$

2. *Moderate stochastic acyclicity if $\forall k \in \{1, \ldots, n\}$*

$$\min_{1 \leq l \leq n, \, l \neq k} \rho(x_l, x_{l+1}) \geq \frac{1}{2} \quad \text{implies} \quad \begin{cases} \rho(x_{k+1}, x_k) > \min_{1 \leq l \leq n, \, l \neq k} \rho(x_l, x_{l+1}) \\ \qquad\qquad \text{or} \\ \rho(x_{k+1}, x_k) = \rho(x_l, x_{l+1}), \ \forall 1 \leq l \leq n, \ l \neq k \end{cases}. \quad (85)$$

3. *Weak stochastic acyclicity if $\forall k \in \{1, \ldots, n\}$*

$$\min_{1 \leq l \leq n, \, l \neq k} \rho(x_l, x_{l+1}) \geq \frac{1}{2} \quad \text{implies} \quad \rho(x_{k+1}, x_k) \geq \frac{1}{2}. \quad (86)$$

The constraints in the definition above are essentially the same as those in Definition 2, except only restricted to loops that are fully observed.

With this, we have the following characterization for rationalizability with incomplete data:

Lemma 7. *Let $\rho : \mathcal{A} \to [0,1]$ be a binary stochastic choice rule observed on $\mathcal{A} \subseteq Z^2$.*

1. *ρ is represented by WU iff every fully observed loop satisfies weak stochastic acyclicity.*

2. *ρ is represented by MU iff every fully observed loop satisfies moderate stochastic acyclicity.*

3. *ρ is represented by SS only if every fully observed loop satisfies strong stochastic acyclicity.*

Proof of this lemma is postponed to A.8.

Lemma 7 establishes that for Weak Utility and Moderate Utility, checking stochastic acyclicity over all observed loops provides necessary and sufficient conditions for rationalizability, with incomplete data. The proof constructs an extension of the observed choice probabilities to all unobserved pairs that preserves the acyclicity property, after which the characterization of stochastic transitivity for complete data apply. For WU, such extension is straightforward by applying the congruence theorem (Richter, 1966) to the preference relation $x \succcurlyeq y$ iff $\rho(x, y) \geq 1/2$. For MU, the extension is not trivial, since ρ is instead a reciprocal relation (Martinetti et al., 2017),[24] and the results of extending acyclicity to a complete rational preference relation cannot be directly applied here. I show the sufficiency by constructing a directed graph, where two vertices x and y have a directed edge iff $(x, y) \in \mathcal{A}$ and $\rho(x, y) > \frac{1}{2}$, and then show that unobserved frequencies can be incrementally assigned while preserving moderate acyclicity, therefore extending ρ from $\mathcal{A}$ to Z^2.

For Simple Scalability, however, the characterization is only necessary. The strict inequality requirements in strong stochastic acyclicity (84) create the obstacle. For a concrete counter-example, consider four choice options x_1, x_2, x_3, x_4 with observed pairs $\rho(x_1, x_2) = \rho(x_2, x_3) = \rho(x_4, x_3) = 0.7$ and $\rho(x_1, x_4) = 0.5$. The only fully observed loop satisfies strong stochastic acyclicity. Yet assigning any value to the unobserved $\rho(x_1, x_3)$ creates a contradiction: the loop $\{(x_1, x_2), (x_2, x_3), (x_1, x_3)\}$ requires $\rho(x_1, x_3) > 0.7$, while the loop $\{(x_4, x_1), (x_1, x_3), (x_3, x_4)\}$ requires $\rho(x_1, x_3) \leq 0.7$. Therefore such extension is not possible, and ρ cannot represented by SS. In the meantime, we note that when all observed choice frequencies differ from 0.5, sufficiency is restored.

This characterization enables practical testing on realistic datasets where complete observation is infeasible. By extending the equivalence between Simple Scalability, Moderate Utility, and Weak Utility representations and stochastic transitivity to incomplete data through stochastic acyclicity, Lemma 7 provides a theoretical foundation for empirical testing with partial observations.

Remark 4 (Relation to other notions of acyclicity in stochastic choice).

This characterization differs from related work on stochastic acyclicity in several ways. A close example is Ok and Tserenjigmid (2023), where they develop some notion of stochastic transitivity and acyclicity for choices with a (possibly incomplete) collection of binary menus. In particular,

---

[24]A reciprocal relation Q has entries $Q(i, j) \in [0, 1]$ and satisfy $Q(i, j) + Q(j, i) = 1$.

the λ-acyclicity defined in their paper is equivalent to weak stochastic transitivity when $\lambda = 1$; furthermore, satisfying λ-acyclicity for all $\lambda \in (0,1]$ is equivalent to a slightly weaker notion of moderate stochastic transitivity, as defined in Chipman (1960). However, Ok and Tserenjigmid (2023)'s approach did not account for the nuance between strict inequality and equality as the MU model requires.[25] My approach analyzes the reciprocal relation ρ directly rather than converting to binary relations, which both provides the precise characterization for MU and reveals why sufficiency fails for SS.

---

[25]Here we adopt the same formulation used in Georgescu-Roegen (1958), Fishburn (1978) and He and Natenzon (2024), although with finite data the two formulations are empirically indistinguishable.

A.8 Proof of Lemma 7

Proof. In the following proof, we use WSA, MSA and SSA as abbreviation for weak/moderate/strong stochastic acyclicity.

Necessity of the transitivity conditions are immediate, and a counter-example is provided in the main text for SSA, so here we only discuss the sufficiency for WSA and MSA. As we discussed, literature has shown that characterizing a choice rule $\rho : Z^2 \to [0,1]$ by the three random choice models is equivalent to ρ satisfying the corresponding transitivity (He and Natenzon, 2024). Therefore, it suffices to show that, for any $\rho : \mathcal{A} \to [0,1]$ where no cycle within $\mathcal{A}$ violates WSA/MSA, ρ can always be extended to $Z^2 \to [0,1]$ that maintains the same property.

We will start by proving the result for WSA, as the proof for MSA builds upon it.

Let $\rho : \mathcal{A} \to [0,1]$ be a choice rule observed on subset $\mathcal{A} \subseteq Z^2$ such that every cycle within $\mathcal{A}$ satisfies WSA. We will show it can be extended to $\bar{\rho} : Z^2 \to [0,1]$ that satisfies WSA.

We first assume there exist no pairs with $\rho(x,y) = \frac{1}{2}$, which is a case we will revisit later. Define a directed graph G, such that each node corresponds to one choice option in Z, and two nodes (x,y) are connected by a directed edge $x \to y$ iff $(x,y) \in \mathcal{A}$ and $\rho(x,y) > \frac{1}{2}$. The assumption that every cycle in $\mathcal{A}$ satisfies WSA implies that G is a directed acyclic graph (DAG), and hence allows for topological sorting. For any choice problem $(x,y) \in Z^2 \backslash \mathcal{A}$, let $\bar{\rho}(x,y) = 1$ if x comes before y in the topological sorting, and $\bar{\rho}(x,y) = 1$ otherwise. By nature of topological sorting, adding the newly assigned pairs with $\bar{\rho}(x,y) = 1$ to the DAG, the resulting graph is still a DAG, and hence we obtain an extension $\bar{\rho}$ that still satisfies WSA. Now consider the case when some pairs in $\mathcal{A}$ have $\rho(x,y) = \frac{1}{2}$. For this, we assign the elements of Z into a set of equivalence classes $\{V_i\}_{i=1}^I$ such that $\forall (x,y) \in \mathcal{A}$, x and y belong to the same equivalence class iff $\rho(x,y) = \frac{1}{2}$. To show that this is well-defined, consider the following construction process: start with each $x \in Z$ as its own equivalence class, and iteratively combine two classes V_i and V_j if $\exists x \in V_i, y \in V_j$ such that $(x,y) \in \mathcal{A}$ and $\rho(x,y) = \frac{1}{2}$. When this process terminates, we can ensure that any pair $(x,y) \in \mathcal{A}$ with $\rho(x,y) = \frac{1}{2}$ belong to the same equivalence class, so it remains to verify that any two options x, y from the same equivalence class V_i with $(x,y) \in \mathcal{A}$ must satisfy $\rho(x,y) = \frac{1}{2}$. In fact, the previous construction process implies there exist $z_1, \ldots, z_l \in V_k$ such that $(x, z_1, \ldots, z_l, y)$ is a fully observed cycle with $\rho(x, z_1) = \rho(z_1, z_2) = \cdots = \rho(z_l, y) = \frac{1}{2}$. From our assumption, the cycle $(x, z_1, \ldots, z_l, y)$ satisfies WSA, which implies that we must have $\rho(x,y) = \frac{1}{2}$ (by applying WSA twice in different directions).

Having constructed these equivalence classes, we can further show that for any two given equivalence classes V_i and V_j, for any $x \in V_i, y \in V_j, (x,y) \in \mathcal{A}$, the values $\rho(x,y) - \frac{1}{2}$ must share the same sign. If otherwise, assume there exist $(x,y), (x',y') \in (V_i \times V_j) \cap \mathcal{A}$ with $\rho(x,y) > \frac{1}{2}$ and $\rho(x',y') < \frac{1}{2}$. Note that we cannot have them equal to $\frac{1}{2}$ because they belong to different equivalence classes, and also we can allow $x = x'$ or $y = y'$. Then, there must exist $z_1, \ldots, z_k \in V_i$ and $w_1, \ldots, w_l \in V_j$ such that

$(x, z_1, \ldots, z_k, x', y', w_1, \ldots, w_l, y)$ is a fully observed cycle, and hence $\rho(x, y) > \frac{1}{2}$ and $\rho(x', y') < \frac{1}{2}$ contradict WSA for this cycle.

With the above argument, we can construct a DAG G similar to above, with the only difference being that we merge each equivalence class into one single node. The construction is still well-defined because we showed that any directed edge between two equivalence classes much share the same direction. The rest follows from the original case, with the additional construction that $\bar{\rho}(x, y) = \frac{1}{2}$ whenever (x, y) belong to the same equivalence class.

Therefore, we conclude the desired result for WSA.

The case of MSA and SSA are more complicated than WSA, intuitively because MSA and SSA pose constraints not only on the sign of $\rho(x, y) - \frac{1}{2}$, but on the magnitude as well. An alternative viewpoint is that, while WSA can be expressed as a complete binary relation R where xRy iff $\rho(x, y) \geq \frac{1}{2}$, MSA and SSA are instead general reciprocal relations. Hence we will build on the proof of WSA, and further show that the magnitude constraints from MSA/SSA can be satisfied.

First consider MSA. Since MSA implies WSA, all results for WSA hold, and in particular we can still construct a DAG G where $x \to y$ iff $(x, y) \in \mathcal{A}$ and $\rho(x, y) > \frac{1}{2}$. The merging of nodes into equivalence classes relies on WSA, which also holds under MSA, so the case with $\rho(x, y) = \frac{1}{2}$ can be handled as well. Hence, through this process we can assign signs to every pair of $(x, y) \in Z^2 \backslash \mathcal{A}$ such that WSA is satisfied.

It remains to assign specific values to those $\rho(x, y)$ with $\rho(x, y) > \frac{1}{2}$ or $\rho(x, y) < \frac{1}{2}$ that preserves MSA. In particular we will show that, if $\rho : \mathcal{A} \to [0, 1]$ is such that every cycle in $\mathcal{A}$ satisfies MSA, then for any given $(x, y) \in Z^2 \backslash \mathcal{A}$, we can assign a value to $\rho(x, y)$ (and by nature $\rho(y, x)$ as well), such that the extended $\bar{\rho}$ defined on $\mathcal{A} \cup \{(x, y), (y, x)\}$ still satisfies MSA. If this holds, then we can repeat this process iteratively, and assign values to every pair in $Z^2 \backslash \mathcal{A}$ and still preserve MSA.

Let $(x, y) \in Z^2 \backslash \mathcal{A}$ be an arbitrarily given unobserved pair that we aim to assign value to. The WSA process has already determined the sign of $\rho(x, y) - \frac{1}{2}$, so without loss of generality assume $\rho(x, y) > \frac{1}{2}$. Now consider what constraints MSA poses on the value of $\rho(x, y)$, which only apply in the case of fully observed cycles (after (x, y) and (y, x) have been assigned and thus added to $\mathcal{A}$). Consider an arbitrary fully observed cycle $(z_1, \ldots, z_{n-1}, z_n)$, where we let $z_1 = x$, $z_n = y$. We also assume this cycle contains no ρ values equal to $\frac{1}{2}$, because in that case we can merge them into one node. Let p be the number of values greater than $\frac{1}{2}$ within the set $C = \{\rho(z_k, z_{k+1})\}_{k=1}^{n-1}$. There are the following cases:

1. If $p = n - 1$, then MSA requires either (a.) $\rho(x, y) > \min_{1 \leq k \leq n-1} \rho(z_k, z_{k+1})$ or (b.) $\rho(x, y) = \rho(z_k, z_{k+1}) = c, \ \forall k = 1 \ldots, n - 1$.
2. If $2 \leq p \leq n - 2$, then MSA has no requirement on $\rho(x, y)$.
3. If $p = 1$, which we assume is $\rho(z_r, z_{r+1}) > \frac{1}{2}$ (where $1 \leq r \leq n - 1$), then MSA requires either

(a.) $\rho(z_r, z_{r+1}) > \min\{\rho(x,y)\} \cup \{1 - \rho(z_k, z_{k+1})\}_{1 \le k \le n-1, \ l \ne r}$ or (b.) $\rho(x,y) = \rho(z_k, z_{k+1}) = c, \ \forall k = 1, \ldots, n-1.$

4. If $p = 0$, this is impossible because it contradicts with the sign assignment $\rho(x,y) > \frac{1}{2}$ during the WSA step.

Among the four cases above, case 1 poses lower bounds for $\rho(x,y)$, case 3 poses lower bounds for $\rho(x,y)$, and both cases 1 and 3 also allow for a special case of equality. The only possibility of incompatible constraints are if an upper bound and a lower bound have no common admissible set, which we now show cannot happen.

Assume cycle $(x, z_2, \ldots, z_{n-1}, y)$ belongs to case 1, and another cycle $(x, w_2, \ldots, w_{m-1}, y)$ belongs to case 3 with $\rho(w_r, w_{r+1}) > \frac{1}{2}$. The key is to notice that $(x, z_2, \ldots, z_{n-1}, y, w_{m-1}, \ldots, w_2)$ is also a fully observed cycle of length $m + n - 2$, and it contains $m + n - 3$ positive values. This one negative value is precisely $\rho(w_{r+1}, w_r)$. Hence MSA requires that either (a.) $\rho(w_r, w_{r+1}) > \min \{\rho(z_k, z_{k+1})\}_{1 \le k \le n-1} \cup \{1 - \rho(w_l, w_{l+1})\}_{1 \le l \le m-1, \ l \ne p}$ or (b.) $\rho(z_k, z_{k+1}) = 1 - \rho(w_l, w_{l+1}) = \rho(w_r, w_{r+1}) = c$ for some constant c, $\forall 1 \le k \le n-1$, $1 \le l \le m-1$, $l \ne r$. In case (a.), we either have (a1.) $\rho(w_r, w_{r+1}) > \min\{\rho(z_k, z_{k+1})\}_{1 \le k \le n-1}$ or (a2.) $\rho(w_r, w_{r+1}) > \min\{1 - \rho(w_l, w_{l+1})\}_{1 \le l \le m-1, \ l \ne p}$. Case (a1.) is precisely the compatibility requirement needed for the upper and lower bounds on $\rho(x,y)$. Case (a2.) implies that the upper bound on $\rho(x,y)$ does not exist because the inequality has already been satisfied by another $1 - \rho$ value. In case (b.), this would force $\rho(x,y) = c$, but this value must satisfy all other MSA constraints, because any other fully observed cycle that (x,y) belongs to can replace (x,y) with a sequence $\{(x, z_2), \ldots, (z_{k-1}, z_k)\}$ where every ρ value equals c. Since the cycle after replacement must satisfy MSA, so will the cycle with (x,y) satisfy MSA.

Hence, we have shown that all constraints posed by MSA on $\rho(x,y)$ must be compatible, and therefore we must be able to assign a value to $\rho(x,y)$ that preserves MSA. Therefore, we conclude the desired result for MSA. $\qquad \square$

A.9   Additional Results on Statistical Testing

A.9.1   Regularity Conditions for Data Generating Processes

Define index subsets based on whether the distribution is degenerate:

$$J = \{j = 1, \ldots, m \mid \mu_j \notin \{0, 1\}\},$$
$$J' = \{j = 1, \ldots, m \mid \mu_j \in \{0, 1\}\}.$$

When $j \in J'$, ρ_{ij} is degenerate distributed, which follows that $\hat{\rho}_{ij}$ and hence $\hat{\rho}_j$ are also degenerate, equaling 0 or 1 almost surely. In other words, the subset J' consists of indices where all individuals choose one specific option with probability 1. This is the set of "trivial" questions, and will contain no variation in the data. Hence for the purpose of analysis this subset can be omitted, as it contributes nothing to the data distribution. In the rest of this subsection, we will only focus on the index subset J for clarity of exposition, as if $J' = \varnothing$. Importantly this is with the understanding that when $J' \neq \varnothing$, we simply add degenerate constant 0 or 1 dimensions to the asymptotic distribution of $\hat{\rho}$, which does not change the result of statistical testing.

Assumption 3 (Regularity Conditions).

(a) $\Pr[K_{ij} > 0] > 0, \ \forall j.$

(b) $\boldsymbol{V}_{ij}$ and $\boldsymbol{V}_{i'j'}$ are independent for $i \neq i'$ or $j \neq j'$.

(c) $\hat{\boldsymbol{\rho}}_{ij} | \boldsymbol{\rho}_i$ and $\hat{\boldsymbol{\rho}}_{i'j'} | \boldsymbol{\rho}_{i'}$ are conditionally independent for $i \neq i'$ or $j \neq j'$.

(d) The distribution F is non-degenerate.

(e) The distribution of $[I_{i1}, \ldots, I_{im}]^T$ is non-degenerate.

Remark 5 (For Regularity Conditions).

(i). Assumption 3(a) rules out the case where the j-th problem will have zero observations with probability 1, in which case we can simply drop that problem index.

(ii). Assumption 3(c) combined with Assumption 1 (c) implies that $\hat{\boldsymbol{\rho}}_{ij}$ and $\hat{\boldsymbol{\rho}}_{i'j'}$ are (unconditionally) independent when $i \neq i'$.

(iii). Assumption 3(d) is a non-degeneracy requirement over the distributions F, where non-degeneracy of a random vector means that there exist no nonzero linear combination of its entries that is almost surely constant. Intuitively, here it requires that there exist no deterministic relation among choice rates of different problems, except for problems where all choice rates are constant 0 or 1. This condition is equivalent to that $\boldsymbol{\Sigma}_0$ is invertible. Also

note that having $J' \neq \varnothing$ will inevitably lead to degeneracy of F on those dimensions, and as we discussed above, we omit those dimensions here and handle separately.

(iv). Assumption 3(e) is a non-degeneracy requirements over the distributions G, namely requiring there exists no nonzero linear combination of its entries that is almost surely constant. Intuitively, this requires there exists no deterministic relation among whether a subject faces various problems.

A.9.2 Normalizing $\boldsymbol{\Omega}$ to $\boldsymbol{I}$

Recall Equation 26, I use the statistic $T_N = \sqrt{N} \left\| \hat{\boldsymbol{\rho}} - \text{proj}_Q(\hat{\boldsymbol{\rho}}, \boldsymbol{\Omega}) \right\|_{\boldsymbol{\Omega}}$, where $\|\boldsymbol{x}\|_{\boldsymbol{\Omega}} = \sqrt{\boldsymbol{x}^T \boldsymbol{\Omega} \boldsymbol{x}}$, and $\text{proj}_Q(\cdot, \boldsymbol{\Omega})$ is the projection onto Q with $\boldsymbol{\Omega}$-induced distance, defined as $\text{proj}_Q(\boldsymbol{x}, \boldsymbol{\Omega}) = \arg\min_{\boldsymbol{q} \in Q} \|\boldsymbol{x} - \boldsymbol{q}\|_{\boldsymbol{\Omega}}$.

Note that the expression with $\boldsymbol{\Omega}$-induced norm can be transformed into one with standard Euclidean norm by redefining the variables.

By letting $\hat{\boldsymbol{\rho}}' = \boldsymbol{\Omega}^{1/2} \hat{\boldsymbol{\rho}}$ and $Q' = \boldsymbol{\Omega}^{1/2} Q = \{\boldsymbol{\Omega}^{1/2} \boldsymbol{x} \mid \boldsymbol{x} \in Q\}$. We have

$$\text{proj}_Q(\hat{\boldsymbol{\rho}}, \boldsymbol{\Omega}) = \arg\min_{\boldsymbol{q} \in Q} \|\hat{\boldsymbol{\rho}} - \boldsymbol{q}\|_{\boldsymbol{\Omega}} = \boldsymbol{\Omega}^{-1/2} \arg\min_{\boldsymbol{q}' \in Q'} \|\hat{\boldsymbol{\rho}}' - \boldsymbol{q}'\| = \boldsymbol{\Omega}^{-1/2} \text{proj}_{Q'}(\hat{\boldsymbol{\rho}}'),$$

where $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^T \boldsymbol{x}}$ and $\text{proj}_Q(x)$ are the Euclidean norm and projection. Hence T_n can be rewritten as

$$T_n = \sqrt{n} \left\| \hat{\boldsymbol{\rho}}' - \text{proj}_{Q'}(\hat{\boldsymbol{\rho}}') \right\|, \text{ where } \sqrt{n}(\hat{\boldsymbol{\rho}}' - \boldsymbol{\Omega}^{1/2} \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \, \boldsymbol{\Omega}^{1/2} \boldsymbol{\Sigma} \boldsymbol{\Omega}^{1/2}).$$

Hence without loss of generality we only need to consider the case $\boldsymbol{\Omega} = \boldsymbol{I}$.

A.9.3 Pointwise Asymptotic Distribution of Statistic

Lemma 8. *Under Assumption 1, 2, and 3,*

$$T_N \xrightarrow{d} T, \tag{87}$$

where the limiting distribution satisfies either one of the following:

i. $T = 0$ with probability 1, or

ii. T^2 is the mixture between a point mass at 0 and a mixture of generalized χ^2 distributions, i.e.

$$\Pr[T = 0] = \pi_0, \tag{88}$$

$$T^2 \mid T > 0 \stackrel{d}{=} \boldsymbol{Z}^T \boldsymbol{\Lambda} \boldsymbol{Z}, \tag{89}$$

where $\pi_0 \in [0, 1/2]$ is the probability mass at 0, and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_m)$ is positive semi-definite, and $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Remark 6 (For Lemma 8)**.**

(i). The intuition behind the asymptotic distribution of T_N hinges crucially on the location of the true population mean $\boldsymbol{\mu}$ relative to the polytope Q. If $\boldsymbol{\mu}$ lies strictly in the interior of the polytope, the limiting distribution of T_N degenerates to a point mass at zero. If $\boldsymbol{\mu}$ lies on the boundary of Q, the limiting distribution becomes non-degenerate: it is characterized as a mixture of a point mass at zero and a mixture of χ^2 distribution. The non-trivial portion of this distribution emerges from the possibility of projecting onto different facets, edges, and vertices of the polytope.

(ii). The probability mass at zero never exceeds $1/2$, which makes simulating a critical value via subsampling pointwise valid. However, since the distribution of T_N does not vary continuously in $\boldsymbol{\mu}$, it is unclear whether uniform validity is obtained via subsampling.

Proof. Recall that $\hat{\boldsymbol{\rho}}$ has asymptotic distribution

$$\sqrt{N}\,(\hat{\boldsymbol{\rho}} - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\mathbf{0},\ \mathbf{\Sigma}),$$

or

$$\hat{\boldsymbol{\rho}} \approx \boldsymbol{\mu} + \frac{1}{\sqrt{N}} \mathbf{\Sigma}^{\frac{1}{2}} \mathbf{Z}, \tag{90}$$

where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ follows standard normal distribution.

We will now derive the asymptotic distribution of T_N. Let C be the hypercube $[0, 1]^m$, and Q is a m-dimensional polytope with $Q \subset C$. We will consider different cases of $\boldsymbol{\mu} \in Q$ under the null hypothesis, and show that the asymptotic distribution of T_N^2 is either a constant 0, or a mixture between χ^2 distributions and a point mass at 0 where the weight of the point mass does not exceed $\frac{1}{2}$. Interestingly, the following analysis relates closely to Fang and Santos (2019)'s discussion on convex set projections.

Case I. $\boldsymbol{\mu} \in \mathrm{int}(Q) \cap \mathrm{int}(C)$. As $N \to \infty$, we have $\Pr[\hat{\boldsymbol{\rho}} \in Q] \to 1$. Because $\hat{\boldsymbol{\rho}} \in Q$ implies $\pi_Q(\hat{\boldsymbol{\rho}}) = \hat{\boldsymbol{\rho}}$ and thus $T_N = 0$, this means $\Pr[\hat{\boldsymbol{\rho}} = 0] \to 1$ and thus T_N converges to 0 in probability.

Case II. $\boldsymbol{\mu} \in \partial Q \cap \mathrm{int}(C)$. In this case, the distribution of T_N is more complicated and depends on the face of Q that $\boldsymbol{\mu}$ lies on. We provide a characterization below. As discussed in the main text, we only need to consider the case of $\mathbf{\Omega} = \mathbf{I}$.

We now analyze the distribution of $T_N^2 = N \left\| \hat{\boldsymbol{\rho}} - \mathrm{proj}_Q(\hat{\boldsymbol{\rho}}) \right\|^2$, the squared projection distance of $\hat{\boldsymbol{\rho}}$ onto the polytope Q. Consider the normal cone $N_Q(\boldsymbol{\mu})$ and tangent cone $T_Q(\boldsymbol{\mu})$ at $\boldsymbol{\mu}$.

As $n \to \infty$, $\hat{\rho}$ lies close to μ, and thus we have $\text{proj}_Q(\hat{\rho}) = \mu + \text{proj}_{T_Q(\mu)}(\hat{\rho} - \mu)$, which is the first-order expansion of the projection operator. Therefore, we have

$$\hat{\rho} - \text{proj}_Q(\hat{\rho}) = \hat{\rho} - (\mu + \text{proj}_{T_Q(\mu)}(\hat{\rho} - \mu)) = \text{proj}_{N_Q(\mu)}(\hat{\rho} - \mu),$$

by Moreau's Decomposition Theorem.

Now from the asymptotic distribution of $\hat{\rho}$ as $\hat{\rho} \approx \mu + \frac{1}{\sqrt{N}}\Sigma^{\frac{1}{2}}Z$ where $Z \sim \mathcal{N}(0, I)$, it follows that the asymptotic distribution of $T_N^2 = N\left\|\hat{\rho} - \text{proj}_Q(\hat{\rho})\right\|^2$ is equal to the distribution $\|\text{proj}_{N_Q(\mu)}(\Sigma^{\frac{1}{2}}Z)\|^2$. Note that since $\mu \in \partial Q$, the normal cone $N_Q(\mu)$ is always non-trivial, i.e. $N_Q(\mu) \neq \{0\}$.

We now show the distribution of T_N is a mixture of square root of generalized chi-squared distributions and a point mass at 0. Note that the projection operator $\text{proj}_{N_Q(\mu)}$ is piecewise linear, namely that the entire space $\mathbb{R}^m$ can be partitioned into subsets $R_1, \ldots, R_k$ such that

$$\text{proj}_{N_Q(\mu)}(x) = P_k x \quad \text{for } x \in R_k,$$

where P_k are projection matrices, thus symmetric and idempotent. Each of these regions correspond to the projected point lying on a different face of Q. In particular, one of these regions is exactly the tangent cone $T_Q(\mu)$ and has $P_k = 0$, giving the point mass at 0.

For the other regions where $P_k \neq 0$, we can write

$$T_N^2 = \|P_k \Sigma^{\frac{1}{2}} Z\|^2 = Z^T \Sigma^{\frac{1}{2}} P_k \Sigma^{\frac{1}{2}} Z.$$

The matrix $\Sigma^{\frac{1}{2}} P_k \Sigma^{\frac{1}{2}}$ is symmetric and thus has eigenvalue decomposition $U\Lambda U^T$ where U is an orthogonal matrix and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m)$. Because $UZ \overset{d}{=} Z \sim \mathcal{N}(0, I)$, we have

$$T_N^2 = Z^T U\Lambda U^T Z \overset{d}{=} Z^T \Lambda Z = \sum_{j=1}^{m} \lambda_j Z_j^2,$$

which is a generalized chi-squared distribution.

Hence, $\|\text{proj}_{N_Q(\mu)}(\Sigma^{\frac{1}{2}} Z)\|^2$ follows a mixture distribution of generalized χ^2 distributions and a point mass at 0. Since Σ is full ranked under Assumption 3 (d) and (e), and hence $\Sigma^{\frac{1}{2}} Z$ is non-degenerate, the point mass at 0 corresponds only to when $\Sigma^{\frac{1}{2}} Z$ lies in the polar cone of $N_Q(\mu)$, namely the tangent cone $T_Q(\mu)$. For any closed convex cone $C' \subsetneq \mathbb{R}^m$, we have $\Pr[Z \in C'] \leq \frac{1}{2}$, with equality iff C' is a half-space.

Take any point $A \neq C'$, and apply the Separating Hyperplane Theorem to the two sets C'

and A, namely that, there exists $\boldsymbol{a} \neq 0$ such that

$$C' \subseteq \{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{a}^T \boldsymbol{x} \leq b\}. \tag{91}$$

It is impossible for any point $\boldsymbol{x} \in C'$ to have $\boldsymbol{a}^T \boldsymbol{x} > 0$, because otherwise conic properties would then imply that $\boldsymbol{a}^T \boldsymbol{x}$ has no upper bound. Hence we know that $C' \subseteq \{\boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{a}^T \boldsymbol{x} \leq 0\}$, i.e., the cone is contained in some closed half-space H. Then we have

$$\Pr[\boldsymbol{Z} \in C'] \leq \Pr[\boldsymbol{Z} \in H] = \Pr[\boldsymbol{a}^T \boldsymbol{Z} \geq 0] = \Pr\left[\frac{\boldsymbol{a}^T \boldsymbol{Z}}{\|\boldsymbol{a}\|} \geq 0\right] = \Pr[\mathcal{N}(0,1) \geq 0] = \frac{1}{2}. \tag{92}$$

Case III. $\boldsymbol{\mu} \in Q \cap \partial C$. $\boldsymbol{\mu}$ being on the boundary of C implies that certain dimensions of $\boldsymbol{\mu}$ equal 0 or 1, and the distributions of $\boldsymbol{\rho}_i$ and $\hat{\boldsymbol{\rho}}_i$ are degenerate on those dimensions. In this case, we can project Q and C onto the non-degenerate dimensions, obtaining polytope Q' and hypercube C'. Other quantities such as $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are marginalized over the non-degenerate dimensions as $\boldsymbol{\mu}'$ and $\boldsymbol{\Sigma}'$. This reduces the problem to a case with smaller m and $\boldsymbol{\mu}' \in \text{int}(C)$ that is already discussed above.

$\square$

A.9.4 Pointwise Validity of Subsampling

According to Assumption 1 (c), $(\bar{\boldsymbol{\rho}}_1, \boldsymbol{I}_1), \ldots, (\bar{\boldsymbol{\rho}}_N, \boldsymbol{I}_N)$ is a sample of n independent and identically distributed random variables, where

$$\bar{\boldsymbol{\rho}}_i = [\bar{\rho}_{i1}, \ldots, \bar{\rho}_{ij}]^T, \ \boldsymbol{I}_i = [I_{i1}, \ldots, I_{ij}]^T, \ \forall i = 1, \ldots, N. \tag{93}$$

Consider a subsample size of $b < N$, with $b/N \to 0$ and $b \to \infty$ as $N \to \infty$. Label the total of $N_b = \binom{N}{b}$ size-b subsets from 1 to $\binom{n}{b}$, and let $T_{N,b,i}$ denote the same statistic evaluated on the i-th subset. Let $G_N(x)$ denote the cumulative distribution function of the sampling distribution of T_N, namely

$$G_N(x) = \Pr[T_N \leq x]. \tag{94}$$

Let $\hat{G}_{N,b}$ denotes the subsampling approximation of the distribution above, namely

$$\hat{G}_{N,b}(x) = \frac{1}{N_b} \sum_{i=1}^{N_b} \mathbb{1}[T_{N,b,i} \leq x]. \tag{95}$$

Define $g_{N,b}(1-\alpha)$ as the $1-\alpha$ quantile of $\hat{G}_{N,b}(x)$, specifically

$$g_{N,b}(1-\alpha) = \inf\left\{x \,\middle|\, \hat{G}_{N,b}(x) \geq 1-\alpha\right\}. \tag{96}$$

Lemma 9 (Pointwise Validity of Subsampling). *Assume $b/N \to 0$ and $b \to \infty$ as $N \to \infty$. Under Assumption 1, 2, and 3, $G(x)$ is continuous at $g(1 - \alpha)$, $\forall \alpha < 0.5$. Hence,*

$$\inf_{\boldsymbol{\mu} \in Q} \lim_{N \to \infty} \Pr[T_N \leq g_{N,b}(1 - \alpha)] = 1 - \alpha. \tag{97}$$

Proof. Recall $(\bar{\boldsymbol{\rho}}_1, \boldsymbol{I}_1), \ldots, (\bar{\boldsymbol{\rho}}_N, \boldsymbol{I}_N)$ is a sample of N independent and identically distributed random variables. Let $G_N(x)$ denote the cumulative distribution function of the sampling distribution of T_N, and $G(x)$ the corresponding limiting cumulative distribution function, whose $1 - \alpha$ quantile is $g(1 - \alpha)$. According to Lemma 8,

1. Under some $\boldsymbol{\mu} \in Q$, $G_N(x)$ converges in distribution to $G(x) = \mathbb{1}\{x \geq 0\}$, i.e. the limiting distribution T is a degenerate distribution equals 0 with probability 1. Since $g_{N,b}(1 - \alpha) \geq 0$,

$$\lim_{N \to \infty} \Pr\{T_N \leq g_{N,b}(1 - \alpha)\} = 1. \tag{98}$$

2. Under some $\boldsymbol{\mu} \in Q$, $G_n(x)$ converges in distribution to $G(x)$, a mixture between a point mass at 0 and chi-squared distributions, with the point mass at 0 having weight no greater than $\frac{1}{2}$.

   The validity of this subsampling test process is established in Politis et al. (1999), Theorem 2.2.1 and Theorem 2.6.1. Since the weight of the point mass 0 never exceeds $\frac{1}{2}$, it follows that $G(x)$ is continuous at $g(1 - \alpha)$, for any $\alpha < \frac{1}{2}$. In practice, we usually choose $\alpha = 0.05$. Therefore

$$g_{N,b}(1 - \alpha) \longrightarrow g(1 - \alpha), \tag{99}$$
$$\text{and} \quad \lim_{N \to \infty} \Pr[T_N \leq g_{N,b}(1 - \alpha)] = 1 - \alpha. \tag{100}$$

Combine Case (a) and (b),

$$\inf_{\boldsymbol{\mu} \in Q} \lim_{N \to \infty} \Pr[T_N \leq g_{N,b}(1 - \alpha)] = 1 - \alpha. \tag{101}$$

$\square$

A.9.5 Remarks on Testing Without Individual Data

The numerical delta method with bootstrap procedure described above requires knowing which responses come from the same individual. In some scenarios, such individual-level data may be unavailable and we only have access to population-level aggregate response. In this subsection, we discuss a possible alternative bootstrap procedure, and conditions under which is valid.

Consider the scenario where n individuals respond to m binary choice problems following Assumptions 1, 2, and 3. Due to the randomness in $\boldsymbol{K}_i$, each individual may respond each problem a

different number of times. Assume there is no information on which responses come from which individuals, and we only observe a collection of N_j responses for each problem $j = 1, \ldots, m$, with $N_0 = \sum_{j=1}^{m} N_j$. In this case, one natural alternative bootstrap procedure would be to treat the N_0 responses as if each comes from a different individual who responded to only 1 problem. Then bootstrap proceeds as normal over this "assumed" population.

It can be shown that under some additional assumptions, the critical value obtained from this alternative bootstrap is still valid. This is formalized as follows. We first state the additional assumption required for this equivalence:

Assumption 4. *Following the setup of Assumption 1,*

(a). $\bar{K} = 1$.

(b). *For all $j \neq j'$, either $\Pr[K_{ij} > 0, \ K_{ij'} > 0] = 0$ or $(\Sigma_0)_{jj'} = 0$.*

The first assumption $\bar{K} = 1$ means that each individual responds to each problem at most once. This assumption naturally holds in a wide range of real-world scenarios, such as polls and elections. In lab experiments, participants are more likely to be asked to respond to the same problem multiple times, but that is also when individual-level data is more likely to be available. When this assumption is violated and an individual may provide multiple responses to the same problem, these responses provide less information about the true population μ than we assumed, leading to under-estimation of the true covariance, and an over-confidence in testing.

The second assumption is more delicate. It implies that for any two choice problems that may be played by the same individual, their responses must be uncorrelated over the population. This is a rather strong assumption, especially for choice problems that share a common choice option. For problems that share no common choice options, this assumption is relatively more justifiable despite still being demanding. The good news is that even when this assumption is violated, so long as the cross covariance $(\Sigma_0)_{jj'}$ are small, the distortion introduced will also small, as we shall see that adopting this alternative bootstrap procedure is equivalent to mis-specifying the covariance matrix by dropping the off-diagonal terms.

The proposed alternative testing procedure corresponds to the following alternative d.g.p.:

Assumption 5 (Alternative d.g.p.).

(a). *There is a set of n choice options, forming $m := \binom{n}{2}$ binary choice problems.*

(b). *Let F be a non-degenerate distribution over $[0, 1]^m$, with mean μ and covariance Σ_0.*

(c). *Let $\mathcal{J}$ be a distribution over the discrete set $\{1, 2, \ldots, m\}$.*

(d). Randomly sampling N' individuals from a large population facing choice problems:

    I. $(\boldsymbol{\rho}_i, J_i) \sim_{\text{iid}} F \times \mathcal{J}$.

    II. J_i denotes the problem index that individual i responds to. Note that each individual only responds to one problem.

    III. $\boldsymbol{\rho}_i$ and J_i are independent.

(e). $\hat{\rho}_{ij} | \rho_{ij} \sim \text{Bernoulli}(\rho_{ij})$.

Under this alternative d.g.p., an estimator $\hat{\boldsymbol{\rho}}^{\text{alt}}$ can be similarly constructed as

$$\hat{\rho}_j^{\text{alt}} = \frac{\sum_{i=1}^{N'} \hat{\rho}_{ij} \mathbb{1}[J_i = j]}{\sum_{i=1}^{N'} \mathbb{1}[J_i = j]}, \qquad j = 1, \ldots, m. \tag{102}$$

The correspondence between the two d.g.p.s are established by the following result:

Lemma 10. *Consider the two d.g.p.s defined by Assumptions 1-4 and Assumption 5. Let $p_j = \Pr[K_{ij} > 0]$ and $q_j = \Pr[J_i = j]$. If $q_j = p_j / (\sum_{j'=1}^{m} p_{j'})$ and $N' = N \left(\sum_{j=1}^{N} p_j \right)$, then the two estimators $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\rho}}^{\text{alt}}$ defined by (23) and (102) have the same asymptotic distribution. Furthermore, their bootstrap estimators also have the same asymptotic distribution.*

Proof. Denote $q_j = \Pr[J_i = j]$. Also we adopt the same notations $p_j = \Pr[K_{ij} > 0]$ and $p_{jj'} = \Pr[K_{ij} > 0, K_{ij'} > 0]$ from previous analysis.

Recall the estimator $\hat{\boldsymbol{\rho}}^{\text{alt}}$ under the alternative d.g.p.

$$\hat{\rho}_j^{\text{alt}} = \frac{\sum_{i=1}^{N_0} \hat{\rho}_{ij} \mathbb{1}[J_i = j]}{\sum_{i=1}^{N_0} \mathbb{1}[J_i = j]}, \qquad j = 1, \ldots, m. \tag{103}$$

Following very similar derivation of the asymptotic distribution in Lemma 4, we have

$$\sqrt{N_0}(\hat{\boldsymbol{\rho}}^{\text{alt}} - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^{\text{alt}}), \tag{104}$$

where

$$\Sigma_{jj}^{\text{alt}} = \frac{1}{q_j} \mu_j (1 - \mu_j), \quad \forall j \tag{105}$$

$$\Sigma_{jj'}^{\text{alt}} = 0, \qquad\qquad \forall j \neq j'. \tag{106}$$

On the other hand, for the original estimator $\hat{\boldsymbol{\rho}}$, its asymptotic covariance $\boldsymbol{\Sigma}$ when $\bar{K} = 1$ reduces

to

$$\Sigma_{jj} = \frac{1}{p_j}\mu_j(1 - \mu_j), \quad \forall j \tag{107}$$

$$\Sigma_{jj'} = \frac{p_{jj'}}{p_j p_{j'}}(\Sigma_0)_{jj'}, \quad \forall j, j'. \tag{108}$$

From Assumption 4 (b), the off-diagonal terms $\Sigma_{jj'}$ vanish. Also, the correspondence between the two d.g.p. implies that $N'q_j = Np_j$. Therefore, the two estimators $\hat{\boldsymbol{\rho}}^{\text{alt}}$ and $\hat{\boldsymbol{\rho}}$ have exactly the same asymptotic distribution.

From here, it immediately follows that the bootstrap estimators also share the same asymptotic distributions, which is equal to that of the original estimators. $\square$

A.10 Proof of Lemma 4

Define

$$S_{Nj} = \frac{1}{N} \sum_{i=1}^{N} \bar{\rho}_{ij} I_{ij}, \quad T_{Nj} = \frac{1}{N} \sum_{i=1}^{N} I_{ij}, \quad \forall j = 1, \ldots, m,$$

and we have $\hat{\rho}_j = S_{Nj}/T_{Nj}$. Multivariate Central Limit Theorem gives

$$\sqrt{N} \left(\begin{bmatrix} S_{N1} \\ T_{N1} \\ \vdots \\ S_{Nm} \\ T_{Nm} \end{bmatrix} - \boldsymbol{u} \right) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \ \boldsymbol{S}),$$

where

$$\boldsymbol{u} = \begin{bmatrix} \mathbb{E}[\bar{\rho}_{i1} I_{i1}] \\ \mathbb{E}[I_{i1}] \\ \vdots \\ \mathbb{E}[\bar{\rho}_{im} I_{im}] \\ \mathbb{E}[I_{im}] \end{bmatrix}$$

and

$$\boldsymbol{S} = \begin{bmatrix} \mathrm{Var}(\bar{\rho}_{i1} I_{i1}) & \mathrm{Cov}(\bar{\rho}_{i1} I_{i1}, I_{i1}) & \ldots & \mathrm{Cov}(\bar{\rho}_{i1} I_{i1}, \bar{\rho}_{im} I_{im}) & \mathrm{Cov}(\bar{\rho}_{i1} I_{i1}, I_{im}) \\ \mathrm{Cov}(I_{i1}, \bar{\rho}_{i1} I_{i1}) & \mathrm{Var}(I_{i1}) & \ldots & \mathrm{Cov}(I_{i1}, \bar{\rho}_{im} I_{im}) & \mathrm{Cov}(I_{i1}, I_{im}) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathrm{Cov}(\bar{\rho}_{im} I_{im}, \bar{\rho}_{i1} I_{i1}) & \mathrm{Cov}(\bar{\rho}_{im} I_{im}, I_{i1}) & \ldots & \mathrm{Var}(\bar{\rho}_{im} I_{im}) & \mathrm{Cov}(\bar{\rho}_{im} I_{im}, I_{im}) \\ \mathrm{Cov}(I_{im}, \bar{\rho}_{i1} I_{i1}) & \mathrm{Cov}(I_{im}, I_{i1}) & \ldots & \mathrm{Cov}(I_{im}, \bar{\rho}_{im} I_{im}) & \mathrm{Var}(I_{im}) \end{bmatrix}.$$

Consider the function $\phi : \mathbb{R}^{2m} \to \mathbb{R}^m$ defined as

$$\phi \left(\begin{bmatrix} x_1 \\ y_1 \\ \vdots \\ x_m \\ y_m \end{bmatrix} \right) = \begin{bmatrix} x_1/y_1 \\ \vdots \\ x_m/y_m \end{bmatrix}.$$

Its Jacobian $\boldsymbol{J} \in \mathbb{R}^{2m \times m}$ is

$$\boldsymbol{J} = \begin{bmatrix} 1/y_1 & & \\ -x_1/y_1^2 & & \\ & \ddots & \\ & & 1/y_m \\ & & -x_m/y_m^2 \end{bmatrix}.$$

Assuming ϕ is differentiable at $\boldsymbol{u}$ (checked in later analysis), from the Delta method, we have

$$\sqrt{n}\left(\hat{\boldsymbol{\rho}} - \phi(\boldsymbol{u})\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \ \boldsymbol{J}(\boldsymbol{u})^T \boldsymbol{S} \boldsymbol{J}(\boldsymbol{u})\right) \tag{109}$$

where $\boldsymbol{J}(\boldsymbol{u})$ is the Jacobian at the point $\boldsymbol{u}$.

We now derive the expression of $\boldsymbol{u}$, $\boldsymbol{S}$ and $\boldsymbol{J}$, and verify that ϕ is differentiable at $\boldsymbol{u}$.

We first compute the conditional mean and covariance of $\bar{\rho}_{ij}|\rho_{ij}, K_{ij}$. We have

$$\mathbb{E}[\bar{\rho}_{ij}|\rho_{ij}, K_{ij}] = \frac{1}{K_{ij}} \sum_{k=1}^{K_{ij}} \mathbb{E}\left[\hat{\rho}_{ij}^{(k)}\middle|\rho_{ij}, K_{ij}\right] = \frac{1}{K_{ij}} \sum_{k=1}^{K_{ij}} \rho_{ij} = \rho_{ij}.$$

Also,

$$\begin{aligned}
\mathrm{Var}[\bar{\rho}_{ij}|\rho_{ij}, K_{ij}] &= \mathrm{Var}\left[\frac{1}{K_{ij}} \sum_{k=1}^{K_{ij}} \hat{\rho}_{ij}^{(k)}\middle|\rho_{ij}, K_{ij}\right] \\
&= \frac{1}{K_{ij}^2} \sum_{1 \le k_1, k_2 \le K_{ij}} \mathrm{Cov}\left(\hat{\rho}_{ij}^{(k_1)}, \hat{\rho}_{ij}^{(k_2)}\middle|\rho_{ij}, K_{ij}\right) \\
&= \frac{1}{K_{ij}^2} \sum_{1 \le k_1, k_2 \le K_{ij}} \mathbb{E}_{\boldsymbol{V}_{ij} \sim H(K_{ij}, \rho_{ij})}\left[\mathrm{Cov}\left(\hat{\rho}_{ij}^{(k_1)}, \hat{\rho}_{ij}^{(k_2)}\middle|\rho_{ij}, K_{ij}, \boldsymbol{V}_{ij}\right)\right] \\
&= \frac{1}{K_{ij}^2} \sum_{1 \le k_1, k_2 \le K_{ij}} \mathbb{E}_{\boldsymbol{V}_{ij} \sim H(K_{ij}, \rho_{ij})}\left[\rho_{ij}(1 - \rho_{ij})(V_{ij})_{k_1 k_2}\right] \\
&= \rho_{ij}(1 - \rho_{ij}) v_{K_{ij}, \rho_{ij}}^2,
\end{aligned}$$

where we define

$$v_{k,p}^2 = \frac{1}{k^2} \sum_{1 \le k_1, k_2 \le k} \mathbb{E}_{\boldsymbol{V} \sim H(k,p)}\left[V_{k_1 k_2}\right], \quad \forall k = 1, \dots, \bar{K} = \mathbb{E}_{\boldsymbol{V} \sim H(k,p)}\left[\boldsymbol{1}^T \boldsymbol{V} \boldsymbol{1}\right], \ \forall p \in [0, 1].$$

Note we have $v_{k,p}^2 \ge 0$ because $\boldsymbol{V} \sim H(k,p)$ is positive semi-definite. Furthermore, for $j \ne j'$, we

have

$$\mathbb{E}[\bar{\rho}_{ij}\bar{\rho}_{ij'}|\rho_{ij},\rho_{ij'},K_{ij},K_{ij'}] = \mathbb{E}\left[\frac{1}{K_{ij}K_{ij'}}\left(\sum_{k=1}^{K_{ij}}\hat{\rho}_{ij}^{(k)}\right)\left(\sum_{k'=1}^{K_{ij'}}\hat{\rho}_{ij'}^{(k')}\right)\Bigg|\rho_{ij},\rho_{ij'},K_{ij},K_{ij'}\right]$$

$$= \frac{1}{K_{ij}K_{ij'}}\sum_{k=1}^{K_{ij}}\sum_{k'=1}^{K_{ij'}}\mathbb{E}\left[\hat{\rho}_{ij}^{(k)}\hat{\rho}_{ij'}^{(k')}\Big|\rho_{ij},\rho_{ij'},K_{ij},K_{ij'}\right]$$

$$= \frac{1}{K_{ij}K_{ij'}}\sum_{k=1}^{K_{ij}}\sum_{k'=1}^{K_{ij'}}\mathbb{E}\left[\hat{\rho}_{ij}^{(k)}\Big|\rho_{ij},K_{ij}\right]\mathbb{E}\left[\hat{\rho}_{ij'}^{(k')}\Big|\rho_{ij'},K_{ij'}\right] \text{ (Assumption 2 (d))}$$

$$= \frac{1}{K_{ij}K_{ij'}}\sum_{k=1}^{K_{ij}}\sum_{k'=1}^{K_{ij'}}\rho_{ij}\rho_{ij'}$$

$$= \rho_{ij}\rho_{ij'},$$

where we used the conditional independence of $\hat{\rho}_{ij}^{(k)}$ and $\hat{\rho}_{ij'}^{(k')}$ given $\rho_{ij},\rho_{ij'},K_{ij},K_{ij'}$.

Define $p_j = \mathbb{E}[I_{ij}] = \Pr[K_{ij} > 0]$ and $p_{jj'} = \mathbb{E}[I_{ij}I_{ij'}] = \Pr[K_{ij} > 0, K_{ij'} > 0]$, which are both determined by the distribution G.

To compute $\boldsymbol{u}$, we have

$$\mathbb{E}[\bar{\rho}_{ij}I_{ij}] = \mathbb{E}_{\rho_{ij},K_{ij}}[\mathbb{E}[\bar{\rho}_{ij}|\rho_{ij},K_{ij}]\cdot I_{ij}]$$

$$= \mathbb{E}_{\rho_{ij},K_{ij}}[\rho_{ij}I_{ij}]$$

$$= p_j\mu_j.$$

$$\mathbb{E}[I_{ij}] = p_j.$$

To compute $\boldsymbol{S}$, we have

$$\mathrm{Var}[\bar{\rho}_{ij}I_{ij}] = \mathbb{E}[\bar{\rho}_{ij}^2 I_{ij}] - \mathbb{E}[\bar{\rho}_{ij}I_{ij}]^2$$

$$= \mathbb{E}_{\rho_{ij},K_{ij}}\left[\mathbb{E}[\bar{\rho}_{ij}^2|\rho_{ij},K_{ij}]\cdot I_{ij}\right] - p_j^2\mu_j^2$$

$$= \mathbb{E}_{\rho_{ij},K_{ij}}\left[\left(\mathrm{Var}(\bar{\rho}_{ij}|\rho_{ij},K_{ij}) + \mathbb{E}[\bar{\rho}_{ij}|\rho_{ij},K_{ij}]^2\right)\cdot I_{ij}\right] - p_j^2\mu_j^2$$

$$= \mathbb{E}_{\rho_{ij},K_{ij}}\left[\left(\rho_{ij}(1-\rho_{ij})v_{K_{ij},\rho_{ij}}^2 + \rho_{ij}^2\right)\cdot I_{ij}\right] - p_j^2\mu_j^2$$

$$= \mathbb{E}_{\rho_{ij},K_{ij}}\left[\rho_{ij}(1-\rho_{ij})\cdot v_{K_{ij},\rho_{ij}}^2 I_{ij}\right] + p_j\mathbb{E}[\rho_{ij}^2] - p_j^2\mu_j^2$$

$$= \bar{v}_j^2 + p_j\left((\Sigma_0)_{jj} + \mu_j^2\right) - p_j^2\mu_j^2$$

$$= p_j(1-p_j)\mu_j^2 + \bar{v}_j^2 + p_j(\Sigma_0)_{jj}.$$

where we define $\bar{v}_j^2 = \mathbb{E}_{\rho_{ij},K_{ij}}\left[\rho_{ij}(1-\rho_{ij})\cdot v_{K_{ij},\rho_{ij}}^2 I_{ij}\right] \geq 0$, and use the fact that $\mathbb{E}[\rho_{ij}^2] = \mathrm{Var}(\rho_{ij}) +$

$\mathbb{E}[\rho_{ij}]^2 = (\Sigma_0)_{jj} + \mu_j^2$. Also,

$$\text{Var}[I_{ij}] = \mathbb{E}[I_{ij}^2] - \mathbb{E}[I_{ij}]^2 = p_j(1 - p_j).$$

and

$$\text{Cov}(\bar{\rho}_{ij} I_{ij}, I_{ij}) = \mathbb{E}[\bar{\rho}_{ij} I_{ij}^2] - \mathbb{E}[\bar{\rho}_{ij} I_{ij}] \cdot \mathbb{E}[I_{ij}] = p_j(1 - p_j)\mu_j.$$

For the terms where $j \neq j'$,

$$
\begin{aligned}
\text{Cov}(\bar{\rho}_{ij} I_{ij}, \bar{\rho}_{ij'} I_{ij'}) &= \mathbb{E}[\bar{\rho}_{ij} \bar{\rho}_{ij'} I_{ij} I_{ij'}] - \mathbb{E}[\bar{\rho}_{ij} I_{ij}] \cdot \mathbb{E}[\bar{\rho}_{ij'} I_{ij'}] \\
&= \mathbb{E}_{\rho_{ij}, \rho_{ij'}, K_{ij}, K_{ij'}} \left[\mathbb{E}[\bar{\rho}_{ij} \bar{\rho}_{ij'} | \rho_{ij}, \rho_{ij'}, K_{ij}, K_{ij'}] \cdot I_{ij} I_{ij'} \right] - p_j p_{j'} \mu_j \mu_{j'} \\
&= \mathbb{E}_{\rho_{ij}, \rho_{ij'}, K_{ij}, K_{ij'}} \left[\rho_{ij} \rho_{ij'} \cdot I_{ij} I_{ij'} \right] - p_j p_{j'} \mu_j \mu_{j'} \\
&= p_{jj'}(\mu_j \mu_{j'} + (\Sigma_0)_{jj'}) - p_j p_{j'} \mu_j \mu_{j'} \\
&= (p_{jj'} - p_j p_{j'})\mu_j \mu_{j'} + p_{jj'}(\Sigma_0)_{jj'}. \\
\text{Cov}(\bar{\rho}_{ij} I_{ij}, I_{ij'}) &= \mathbb{E}[\bar{\rho}_{ij} I_{ij} I_{ij'}] - \mathbb{E}[\bar{\rho}_{ij} I_{ij}] \cdot \mathbb{E}[I_{ij'}] \\
&= \mathbb{E}_{\rho_{ij}, \rho_{ij'}, K_{ij}, K_{ij'}} \left[\mathbb{E}[\bar{\rho}_{ij} | \rho_{ij}, K_{ij}] \cdot I_{ij} I_{ij'} \right] - p_j p_{j'} \mu_j \\
&= \mathbb{E}_{\rho_{ij}, \rho_{ij'}, K_{ij}, K_{ij'}} \left[\rho_{ij} \cdot I_{ij} I_{ij'} \right] - p_j p_{j'} \mu_j \\
&= (p_{jj'} - p_j p_{j'})\mu_j. \\
\text{Cov}(I_{ij}, I_{ij'}) &= \mathbb{E}[I_{ij} I_{ij'}] - \mathbb{E}[I_{ij}] \cdot \mathbb{E}[I_{ij'}] \\
&= p_{jj'} - p_j p_{j'}.
\end{aligned}
$$

Furthermore, the Jacobian $J(u)$ at point u is

$$
J(u) = \begin{bmatrix}
1/p_1 & & \\
-\mu_1/p_1 & & \\
& \ddots & \\
& & 1/p_m \\
& & -\mu_m/p_m
\end{bmatrix}.
$$

Note that our assumption $p_j > 0, \forall j$ implies that ϕ is differentiable at u. Hence the Jacobian is well defined, and the Delta method applies. This gives the expressions of u, S and J.

Finally, we plug the expressions of u, S and J into Equation 109.

The asymptotic mean of $\hat{\rho}$ is $\phi(u) = \mu$.

Denote the asymptotic variance of $\hat{\rho}$ as $\Sigma = J^T S J$. From the block diagonal structure of J, the

diagonal entries of $\boldsymbol{\Sigma}$ are

$$\Sigma_{jj} = \begin{bmatrix} 1/p_j & -\mu_j/p_j \end{bmatrix} \begin{bmatrix} p_j(1-p_j)\mu_j^2 + \bar{v}_j^2 + p_j(\Sigma_0)_{jj} & p_j(1-p_j)\mu_j \\ p_j(1-p_j)\mu_j & p_j(1-p_j) \end{bmatrix} \begin{bmatrix} 1/p_j \\ -\mu_j/p_j \end{bmatrix}$$

$$= \frac{1}{p_j^2} \left(\bar{v}_j^2 + p_j(\Sigma_0)_{jj} \right).$$

The off-diagonal entries of $\boldsymbol{\Sigma}$ are

$$\Sigma_{jj'} = \begin{bmatrix} 1/p_j & -\mu_j/p_j \end{bmatrix} \begin{bmatrix} (p_{jj'} - p_j p_{j'})\mu_j\mu_{j'} + p_{jj'}(\Sigma_0)_{jj'} & (p_{jj'} - p_j p_{j'})\mu_j \\ (p_{jj'} - p_j p_{j'})\mu_{j'} & p_{jj'} - p_j p_{j'} \end{bmatrix} \begin{bmatrix} 1/p_{j'} \\ -\mu_{j'}/p_{j'} \end{bmatrix}$$

$$= \frac{p_{jj'}}{p_j p_{j'}}(\Sigma_0)_{jj'}.$$

To sum up, the asymptotic distribution of $\hat{\boldsymbol{\rho}}$ is

$$\sqrt{N}\left(\hat{\boldsymbol{\rho}} - \boldsymbol{\mu}\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \ \boldsymbol{\Sigma}\right)$$

where $\boldsymbol{\Sigma}$ is a matrix with

$$\Sigma_{jj} = \frac{1}{p_j^2}\bar{v}_j^2 + \frac{1}{p_j}(\Sigma_0)_{jj}, \quad \forall j,$$

$$\Sigma_{jj'} = \frac{p_{jj'}}{p_j p_{j'}}(\Sigma_0)_{jj'}, \quad \forall j \neq j'.$$

We can show that under Assumption 3 (d) and (e), the matrix $\boldsymbol{\Sigma}$ is positive-definite. Define matrix $\boldsymbol{M} \in \mathbb{R}^{m \times m}$ such that $M_{jj} = \frac{1}{p_j}$ and $M_{jj'} = \frac{p_{jj'}}{p_j p_{j'}}$ for $j \neq j'$. Consider random vector $\boldsymbol{U} = \left[\frac{1}{p_1}I_{i1}, \ldots, \frac{1}{p_m}I_{im}\right]^T$. We have $\boldsymbol{M} = \mathbb{E}[\boldsymbol{U}\boldsymbol{U}^T]$ as the Gram matrix of $\boldsymbol{U}$. From Assumption 3 (e), we know $\boldsymbol{U}$ has non-degenerate distribution, and thus $\boldsymbol{M}$ is positive definite. From Assumption 3 (d), we know $\boldsymbol{\Sigma}_0$ is positive definite. Hence by Schur product theorem, the Hadamard product $\boldsymbol{M} \circ \boldsymbol{\Sigma}_0$ is positive definite. Finally, because $\boldsymbol{\Sigma} = \boldsymbol{M} \circ \boldsymbol{\Sigma}_0 + \text{diag}([\frac{1}{p_1^2}\bar{v}_1^2, \ldots, \frac{1}{p_m^2}\bar{v}_m^2])$ and $\bar{v}_j^2 \geq 0$, it follows that $\boldsymbol{\Sigma}$ is positive definite.

A.11   Proof of Theorem 2

The proof largely follows Theorem 3.5 of Hong and Li (2018). In the following, we will verify that the assumptions of Theorem 3.5 holds.

We first verify that ϕ is convex. For any two points $\boldsymbol{x}_1, \boldsymbol{x}_2$, let $\boldsymbol{p}_1 = \mathrm{proj}_Q(\boldsymbol{x}_1)$ and $\boldsymbol{p}_2 = \mathrm{proj}_Q(\boldsymbol{x}_2)$. For any $\lambda \in (0,1)$, we have

$$
\begin{aligned}
\phi(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2) &= \|\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2 - \mathrm{proj}_Q\left(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2\right)\| \\
&\leq \|\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2 - (\lambda \boldsymbol{p}_1 + (1-\lambda)\boldsymbol{p}_2)\| \\
&= \|\lambda(\boldsymbol{x}_1 - \boldsymbol{p}_1) + (1-\lambda)(\boldsymbol{x}_2 - \boldsymbol{p}_2)\| \\
&\leq \lambda\|\boldsymbol{x}_1 - \boldsymbol{p}_1\| + (1-\lambda)\|\boldsymbol{x}_2 - \boldsymbol{p}_2\| \\
&= \lambda\phi(\boldsymbol{x}_1) + (1-\lambda)\phi(\boldsymbol{x}_2),
\end{aligned}
$$

which shows the convexity of ϕ. Here the first inequality follows from the optimality of the projection operator, and the second inequality follows from triangle inequality.

We next verify that ϕ is 1-Lipschitz. For any two points $\boldsymbol{x}_1, \boldsymbol{x}_2$, for any $\boldsymbol{q} \in Q$, triangle inequality gives $\|\boldsymbol{x}_1 - \boldsymbol{q}\| \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \|\boldsymbol{x}_2 - \boldsymbol{q}\|$. Take the infimum of both sides over $\boldsymbol{q} \in Q$, we have $\phi(\boldsymbol{x}_1) \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \phi(\boldsymbol{x}_2)$, using the fact that $\phi(\boldsymbol{x}) = \inf_q \|\boldsymbol{x} - \boldsymbol{q}\|$. Swapping $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ gives $\phi(\boldsymbol{x}_2) \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \phi(\boldsymbol{x}_1)$. Combining the two inequalities, we get the desired 1-Lipschitz property

$$
|\phi(\boldsymbol{x}_1) - \phi(\boldsymbol{x}_2)| \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|.
$$

It remains to verify that Assumptions 3.1 and 3.2 from Hong and Li (2018) hold in our case. Notice that we already showed in Lemma 4 that $\sqrt{N}(\hat{\boldsymbol{\rho}} - \boldsymbol{\mu})$ converges to the limiting distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\mu}))$ pointwise over all $\boldsymbol{\mu} \in Q$. Additionally, from Assumptions 1-3 we have that $\hat{\boldsymbol{\rho}}$ is restricted to the unit cube $[0,1]^m$, thus the covariance $\boldsymbol{\Sigma}$ is smooth and bounded. Hence the empirical process $X_N(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \sqrt{N}(\hat{\boldsymbol{\rho}} - \boldsymbol{\mu})$ is uniformly asymptotically tight. Moreover, the function $f_{\boldsymbol{\mu}}(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{\mu}$ is 1-Lipschitz in $\boldsymbol{\mu}$, and by Donsker Theorem, we have convergence in L^∞ of $X_N(\boldsymbol{\mu})$ (Theorem 19.3 and Example 19.7, (Van der Vaart, 2000)). This gives the first part of Assumption 3.1(i). The second part of Assumption 3.1(i) follows directly from the uniformly asymptotically tight property.

For Assumption 3.1(ii), in our setup the variable $\mathbb{Z}_n^*$ from Hong and Li (2018)'s notation is the bootstrap estimator residual $\sqrt{N}(\hat{\boldsymbol{\rho}}_b - \hat{\boldsymbol{\rho}})$. The validity of bootstrap process implies that the bootstrap estimator has the same (pointwise) asymptotic distribution. Thus the same argument as Assumption 3.1(i) above applies, giving convergence in L^∞ of the bootstrap estimator $\sqrt{N}(\hat{\boldsymbol{\rho}}_b - \boldsymbol{\mu})$ to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\mu}))$. Combining this with the convergence in L^∞ of $\sqrt{N}(\hat{\boldsymbol{\rho}} - \boldsymbol{\mu})$ and triangle inequality, it follows that $\sqrt{N}(\hat{\boldsymbol{\rho}}_b - \hat{\boldsymbol{\rho}})$ also converges in L^∞ to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\mu}))$.

For Assumption 3.2 from Hong and Li (2018), we temporarily adopt the notations from that paper. Notice that in our case, the limiting distribution $\mathbb{G}_0$ is $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}(\boldsymbol{\mu}))$, a multivariate Gaussian with full-rank covariance. Also notice that $\mathcal{C}_{a,d,x}$ is a polytope, hence $\partial\mathcal{C}_{a,d,x}$ is a set with Borel measure zero. Therefore we always have $\Pr[\mathbb{G}_0 \in \partial\mathcal{C}_{a,d,x}] = 0$, and Assumption 3.2 is satisfied.

Therefore, all assumptions of Theorem 3.5 of Hong and Li (2018) are satisfied, and we obtain the desired result.

A.12 Statistical Test for Heterogeneous Tastes

Model each participant i as having an "underlying" preference vector $\boldsymbol{\rho}_i \in [0, 1]^{17}$, and their choice for the j-th menu is a Bernoulli(ρ_{ij}) random variable independent across repetitions and menus. We want to test whether all participants share the same taste, in the sense that whether they prefer one type over another are consistent, and only the strength of the preference may vary. Concretely, the null hypothesis to test is that for any two participants i, i', we have $\text{sgn}(\rho_{ij} - \frac{1}{2}) = \text{sgn}(\rho_{i'j} - \frac{1}{2})$ for all j. [26]

Let $\hat{\rho}_{ij}^{(k)}$ be participant i's k-th repetition response to binary menu j, and denote $\bar{\rho}_{ij} = \frac{1}{10} \sum_{k=1}^{10} \hat{\rho}_{ij}^{(k)}$. To test this null hypothesis, first notice that it can be decomposed into a total of 2^{17} individual hypotheses, each corresponding to one fixed sign pattern of $\text{sgn}(\boldsymbol{\rho}_i - \frac{1}{2})$. For each individual hypothesis, we perform the likelihood ratio test using the maximum likelihood estimation of parameters $\boldsymbol{\rho}_i$. Let $\boldsymbol{s} \in \{\pm 1\}^{17}$ represent a given sign pattern of $\text{sgn}(\boldsymbol{\rho} - \frac{1}{2})$, then the maximum likelihood estimator of $\boldsymbol{\rho}_i$ is

$$\hat{\rho}_{ij}^{\text{MLE}}(\boldsymbol{s}) = \begin{cases} \max\{\bar{\rho}_{ij}, \frac{1}{2}\} & \text{if } s_j = 1 \\ \min\{\bar{\rho}_{ij}, \frac{1}{2}\} & \text{if } s_j = -1 \end{cases}. \tag{110}$$

From this, we can compute the maximum log-likelihood under the restricted model with sign pattern $\boldsymbol{s}$ as

$$\ell_0(\boldsymbol{s}) = \sum_i \sum_j \sum_k \hat{\rho}_{ij}^{(k)} \log \hat{\rho}_{ij}^{\text{MLE}}(\boldsymbol{s}) + \left(1 - \hat{\rho}_{ij}^{(k)}\right) \log \left(1 - \hat{\rho}_{ij}^{\text{MLE}}(\boldsymbol{s})\right). \tag{111}$$

On the other hand, under the unrestricted model, the maximum likelihood estimator of $\boldsymbol{\rho}_i$ will simply be the sample mean $\hat{\rho}_{ij}^{\text{MLE}} = \bar{\rho}_{ij}$, and the maximum log-likelihood under the unrestricted model is

$$\ell_1 = \sum_i \sum_j \sum_k \hat{\rho}_{ij}^{(k)} \log \bar{\rho}_{ij} + \left(1 - \hat{\rho}_{ij}^{(k)}\right) \log \left(1 - \bar{\rho}_{ij}\right). \tag{112}$$

For each individual null hypothesis, the likelihood ratio statistic is $\lambda_{\text{LR}}(\boldsymbol{s}) = -2(\ell_0(\boldsymbol{s}) - \ell_1)$. Let $\lambda_{\text{LR}} = \min_{\boldsymbol{s}} \lambda_{\text{LR}}(\boldsymbol{s})$ as the smallest statistic value among all 2^{17} values. This is the "most favorable" choice of $\boldsymbol{s}$ that gives the highest possible likelihood under the null. To reject the original null hypothesis, it is equivalent to reject all individual null hypotheses, and in turn equivalent to reject this one hypothesis with the smallest statistic value.

The asymptotic distribution of λ_{LR} is in general a mixture chi-squared distribution and difficult to compute,. However, since the total number of degrees of freedom in the parameters is $34 \times 17 = 578$, this mixture chi-squared can have at most 578 degrees of freedom, which serves as an inflated estimation of the distribution of the test statistic. In other words, using the $1 - \alpha$ quantile of χ_{578}^2 will provide a conservative proxy critical value, in the sense that rejection using this proxy critical value will be sufficient for rejection for the test.

---

[26]Here we omit the specific case of $\rho_{ij} = \frac{1}{2}$, which can be considered a limiting case both when $\text{sgn}(\rho_{ij} - \frac{1}{2}) = 1$ and when $\text{sgn}(\rho_{ij} - \frac{1}{2}) = 1$.

With this dataset, we get $\lambda_{\mathrm{LR}} = 3363.02$, and hence reject the null hypothesis with $p < 0.001$. Therefore, the test provides overwhelming evidence against the null hypothesis that all individuals share the same taste, further supporting the Collective SS model.