UNIVERSITY of York

Discussion Papers in Economics

No. 25/03

How different are we? Identifying the degree of revealed preference heterogeneity

Khushboo Surana

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

How different are we? Identifying the degree of revealed preference heterogeneity

Khushboo Surana*

July 9, 2025

Abstract

We propose a nonparametric method to identify the degree of heterogeneity in individual preferences. Using preference information revealed by observed behavior, the method estimates interpersonal preference heterogeneity as the Kemeny distance between individual preference rankings. Using data from the U.S. Panel Study of Income Dynamics (PSID), we derive bounds on the distance-based heterogeneity measure, which we then use to group individuals with similar preferences. We demonstrate that constructing such preference types can substantially strengthen empirical analysis by (i) producing more accurate demand predictions, (ii) improving welfare analysis, and (iii) detecting functional misspecifications in parametric methodologies.

JEL codes: C14, C60, D11, D12

Keywords: Preference heterogeneity, Revealed preference analysis, Kemeny distance

*University of York. E-mail: khushboo.surana@york.ac.uk.

I thank Abi Adams-Prassl, Miguel Ballester, Bart Capéau, Edoardo Ciscato, Laurens Cherchye, Sam Cosaert, Rahul Deb, Thomas Demuynck, Bram De Rock, Geert Dhaene, Mikhail Freer, Peter Neary, Krishna Pendakur, Frederic Vermeulen, Ewout Verriest and all seminar participants at KU Leuven, UC Louvain, University of Oxford, WU Vienna, IESEG School of Management, IIT Kanpur, University of Zurich, University of York, BI Norwegian Business School, University of Kent and RES Annual Conference for helpful comments.

1 Introduction

How differently would a policy or price change affect different individuals? Finding an accurate answer to this question has important implications for empirical economic analysis. If preferences are similar, econometricians can rely on pooled individual observations and make informative average predictions that could apply to everyone in the population. On the other hand, if preferences are heterogeneous, relying on average demand responses may give misleading policy recommendations. Importantly, a policy change targeted to maximize the welfare of a hypothetical "average" consumer may have varied effects, possibly harming the most vulnerable. This highlights the importance of understanding the extent and nature of preference heterogeneity across individuals.

Preference heterogeneity. The pervasive nature of unobserved preference heterogeneity and its importance in economics research is well recognized (Heckman, 2001). If we had access to long individual panel data, so that it was possible to analyze every individual separately, heterogeneity would not be such a troubling issue. Alternatively, if observable attributes like age, gender or education could reliably predict which individuals are likely to respond similarly to changing stimuli, their observations could be pooled. However, detailed individual-level data are scarce, and the literature has shown that observable attributes only partly explain the differences in observed choices (see, e.g., Banks, Blundell, and Lewbel, 1997; Lewbel and Pendakur, 2009). This suggests that even among the most similar-looking individuals, there exists rich forms of unobserved preference heterogeneity that influences their consumption behavior in complex, yet considerable, ways.

This paper proposes a nonparametric method to empirically quantify the degree of preference heterogeneity. Weitzman (1992) argues that "distance is such an absolutely fundamental concept in the measurement of dissimilarity that it must play an essential role in any meaningful theory of diversity or classification". Following this line of thought, the method quantifies interpersonal preference heterogeneity as the *Kemeny distance* between individual preference rankings over a finite set of choice alternatives. Intuitively, the Kemeny distance is defined as the number of pairwise disagreements between rankings. To empirically identify this distance, we exploit preference relations revealed through observed individual behavior. The method set identifies heterogeneity by constructing upper and lower bounds on the Kemeny distance.

A distinguishing feature of the method is that it follows the revealed preference tradition of Samuelson (1938); Afriat (1967); Diewert (1973), and Varian (1982). This makes it intrinsically nonparametric, meaning that it abstains from imposing any unverifiable functional structure on individual preferences. In the empirical revealed preference literature, preference heterogeneity has been examined by Gross (1995), who used revealed preference axioms on cross-sectional data to nonparametrically test assumptions of preference homogeneity across individuals. For cases where cross-sectional preference homogeneity is rejected, Crawford and Pendakur (2012) and Cosaert (2019) developed revealed preference algorithms to identify the smallest number of preference types required to rationalize the data.[1] In a longitudinal setting, Heufer (2014) used the revealed preference framework to compare the degree of risk aversion across individuals, while Castillo and Freer (2018) introduced a goodness-of-fit-based measure to determine whether individuals' preferences are revealed as heterogeneous. A key difference between these studies and the current paper is that, whereas existing measures yield binary outcomes - two individuals can either have the same or different preferences - our measure is cardinal and distance-based.

Our contributions. The paper makes three main contributions. First, it demonstrates that the Kemeny distance, a well-known and axiomatically justified heterogeneity measure, can be productively applied to standard choice datasets to identify revealed preference heterogeneity. Although the Kemeny distance is an established measure, it has primarily been used in contexts where preference rankings are assumed to be observed. For example, in the social choice literature, the Kemeny distance is employed to determine a consensus ranking that reflects the preferences of multiple individuals over a set of alternatives (see Kemeny, 1959; Kemeny and Snell, 1962; Baldiga and Green, 2013). In computer science, this distance function has been used to measure similarity between users' preferences and to develop recommender systems that predict user behavior based on the behavior of comparable users (see, e.g., Liu and Yang, 2008; Wang, Sun, Gao, and Ma, 2012, 2014).[2] Unlike these literatures, where individual preferences are typically assumed to be known, we consider a demand data setting in which only choices are observed. By integrating the Kemeny distance with the revealed preference literature, we demonstrate how it can be empirically identified from observed choices.

[1] Another interesting strand of literature focuses on studying preference heterogeneity within an individual. When individual choice observations cannot be rationalized with a single utility maximization hypothesis, an explanation could be that inconsistencies are due to multiple perfectly rationalizable preferences within an individual. May (1954) and Kalai, Rubinstein, and Spiegler (2002) study the problem of determining the minimum number of preferences required to perfectly rationalize the observed data. Such "multi-selves" models have been applied in various contexts to explain individual choice observations (e.g., Ambrus and Rozen, 2014; Cherchye, De Rock, Griffith, O'Connell, Smith, and Vermeulen, 2017).

[2] Other applications of rank-based distance functions include automatic evaluation of natural language processing systems (Lapata, 2003, 2006), computer vision applications, metasearch engine formation (Dwork, Kumar, Naor, and Sivakumar, 2001), and gene prioritization (Kim, Raisali, Farnoud, and Milenkovic, 2013), among others.

Second, existing revealed preference measures of heterogeneity focus on examining the minimal heterogeneity required to explain the observed choices of all individuals. In contrast, our method set identifies the Kemeny distance, thereby capturing the full range of heterogeneity within which the true value of the heterogeneity measure lies. Since there are usually many types of preference that can rationalize individual data, our method constructs lower and upper bounds for the Kemeny distance between preferences. The lower bound represents the closest two individuals' preferences could be, given their observed behavior. In contrast, the upper bound reflects the farthest their preferences could be under some set of preferences that explain their observed choices.

Third, we demonstrate the practical applicability of the method through an application to household demand data drawn from the Panel Study of Income Dynamics (PSID) survey. Our method allows us to set identify the interpersonal heterogeneity measure, which we use to construct groups of individuals with similar preferences. Specifically, we use the identified bounds to cluster the observed individuals into a small number of preference types such that preferences within each type are similar.[3] The principle underlying the construction of a small number of types is Occam's razor, which advocates for adopting a simpler model that is consistent with the data over a more complex one. Consequently, while a model where each individual is characterized by a unique type is plausible, a simpler model with a small number of preference types is preferable for generating more meaningful counterfactual predictions and conducting welfare analysis.

We illustrate the usefulness of constructing preference types based on our estimated bounds through three applications of empirical demand analysis. The first two applications address a common critique of revealed preference methods: their low empirical power. The third application focuses on concerns about functional form assumptions in parametric methods of demand analysis. Results from the first two applications show that working with our constructed preference types leads to substantial improvements in identifying power compared to using types based on observed attributes. These preference types produce significantly more accurate out-of-sample demand predictions and provide much more informative welfare analyses. Results from the third application demonstrate that the constructed preference types can serve as a valuable starting point for assessing the extent of misspecification in a particular functional form assumption.

[3] Introducing minimal heterogeneity using revealed preference techniques is becoming increasingly popular. For instance, Crawford and Pendakur (2012) and Cosaert (2019) use the partitioning idea to construct minimum number of types required to rationalize cross-sectional data. Liang (2019) considers cross-sectional demand inconsistencies due to preference heterogeneity and choice errors. She adopts a random utility model to account for choice errors while recovering the minimal number of preferences that can rationalize the data. Adams, Blundell, Browning, and Crawford (2015) apply a revealed preference framework to capture the minimal taste variation needed to rationalize patterns of tobacco consumption.

Outline. The rest of the paper is structured as follows. Section 2 presents the theoretical framework and details the identifying strategy for interpersonal preference heterogeneity. Section 3 discusses the data used in the empirical application and presents the recovered bounds for the heterogeneity measure. Section 4 demonstrates the usefulness of the recovered heterogeneity estimates through three empirical applications of consumer demand analysis. Section 5 concludes the paper.

2 Theoretical Framework

In what follows, we describe the distance function we use to characterize the degree of preference heterogeneity. This distance function cannot be directly computed if we have no information on individual preference rankings. To adapt for a typical observational setting, we next show how revealed preference conditions can be used to set-identify the heterogeneity measure.

2.1 Degree of Preference Heterogeneity

Consider two individuals a and b, and assume that we have complete information on their preference ranking over a set of bundles. Suppose that the preference rankings are strict. Now, consider the following reasonable conditions that any cardinal measure of preference heterogeneity $d(a, b)$ must satisfy:

(i) $d(a, b) \geq 0$, and equality holds if, and only if, preference rankings of a and b are the same.

(ii) $d(a, b) = d(b, a)$

(iii) $d(a, b) \leq d(a, c) + d(c, b)$ for all a, b, c

(iv) Relabeling of bundles should not change the distance between preferences rankings

(v) If the rankings of a and b agree at the beginning of the list and at the bottom of the list, and only disagree on ranking k bundles in the middle, for some number k, then the distance must be the same if only these k bundles in the middle were under consideration.

(vi) The minimum positive distance is 1.

Conditions $(i) - (iii)$ are the standard conditions for a distance function to qualify as a metric. Condition (iv) states that if names (or labels) of the bundles are exchanged, but the underlying preferences are the same, then the resulting distance measure should remain unchanged. Condition (v) requires that if two preference rankings are the same at the top and bottom of the ranked consumption bundles, then removing these bundles (at the top and bottom) should leave the distance unchanged. Finally, condition (vi) may be seen as a choice of unit of measurement.

One of the most popular, and the only, rank-based distance functions that satisfies these axiomatic properties is the well-known Kemeny distance (Kemeny, 1959; Kemeny and Snell, 1962).[4] The Kemeny distance is a simple and intuitive metric for measuring the "distance" between preferences and is commonly used in social choice literature to tackle preference aggregation issues and for deriving a consensus ranking that best represents a set of input rankings. Definition 1 uses the Kemeny distance to define a distance-based measure of preference heterogeneity.

Definition 1 *Given two individuals a and b, and their (strict) preference rankings $\succ_a$ and $\succ_b$ over a finite set of bundles $Q = \{q_t\}_{t \in T}$, define their degree of preference heterogeneity $d(a,b)$ as*

$$d(a,b) = \sum_{\{i,j \; | i>j, \; q_i, q_j \in Q\}} d_{ij}(a,b),$$

$$d_{ij}(a,b) = \begin{cases} 1 \; \textit{if} \; (q_i \succ_a q_j \; \textit{and} \; q_j \succ_b q_i) \; \textit{or} \; (q_j \succ_a q_i \; \textit{and} \; q_i \succ_b q_j) \\ 0 \; \textit{if} \; (q_i \succ_a q_j \; \textit{and} \; q_i \succ_b q_j) \; \textit{or} \; (q_j \succ_a q_i \; \textit{and} \; q_j \succ_b q_i) \end{cases}$$

The indicator function $d_{ij}(a,b)$ specifies if the two individuals are in conflict over ranking bundles q_i and q_j. If a prefers q_i over q_j while b prefers q_j over q_i (or vice versa), then $d_{ij}(a,b)$ equals one. Otherwise, if both agree in ranking the two bundles, then $d_{ij}(a,b)$ equals zero. The Kemeny distance is the sum of $d_{ij}(a,b)$ over all distinct pairs of q_i and q_j.[5] This distance measure allows us to compare the similarity of preferences between pairs

[4] The Kemeny distance is also popularly known as the Kendall tau rank distance. Some of the other popular rank-based distance functions are Spearman footrule distance, Spearman rank correlation, Hamming distance, Cayley distance, and Hausdorff distance (see Deza and Huang, 1998 for an overview). Identification of these alternate distance measures in a continuous demand setting can be operationalized using revealed preference methodology along similar lines as shown in the current paper. This can be a fruitful future research avenue.

[5] In Definition 1, all pairs of bundles are assigned equal weight, which means that preference conflicts between similar and dissimilar bundles receive the same level of penalty. However, in certain applications, it may be more appropriate to assign less penalty to conflicts between similar bundles or more penalty to conflicts that appear higher in the preference rankings. To accommodate these scenarios, weighted versions of the Kemeny distance can be used (see, e.g., Farnoud, Milenkovic, and Touri, 2012; Can, 2014). To

of individuals.[6] Specifically, we can say that individual a has preferences more similar to individual b than c if $d(a,b) \leq d(a,c)$.

What justifies the use of the Kemeny distance as a measure of interpersonal preference heterogeneity? In addition to the intuitive nature of the Kemeny distance and the axiomatic justifications provided above, we argue that there is a lack of a suitable nonparametric measure to quantify differences in preferences. Empirical revealed preference studies have used goodness-of-fit indices (such as the Critical Cost Efficiency Index; Afriat, 1973) to measure preference heterogeneity. However, these measures are inadequate for capturing heterogeneity in underlying preferences (Gross, 1991). Specifically, money metric-based heterogeneity indices suggest that consumers with similar incomes and prices are comparable, regardless of differences in their preferences. We illustrate this with an example in Appendix F.1. This limitation is addressed by our method as it uses a structural distance-based measure to quantify heterogeneity. In addition, goodness-of-fit measures only provide information about the minimum adjustment required for two individuals to avoid violations of rationality conditions. As a result, they tend to underestimate preference heterogeneity. By contrast, by set identifying the Kemeny distance, we capture both the minimum and maximum degree of preference heterogeneity revealed through the observed choices.

Many practical applications normalize the Kemeny distance by dividing the function with the total number of distinct pairs of consumption bundles ($|Q|(|Q|-1)/2$). The normalized Kemeny distance takes a value ranging from zero to one.[7] If preference rankings are identical, its value is zero; if they are completely opposite, its value is one. Intuitively, it indicates the probability that two preference rankings would rank two bundles in opposite orders. For the sake of exposition, in what follows, we use the normalized Kemeny distance as our measure of preference heterogeneity.

generalize the heterogeneity measure to more general choice functions, the distance measure proposed by Klamler (2008) can be used.

[6] Distance measures are typically defined for pairs of rankings. As an implication, the method proposed in this paper is limited to pairwise comparisons. As will be shown later, the estimated pairwise dissimilarity measures can be used to construct preference types. Nevertheless, extending the methodology introduced in this paper to identify a more general group dissimilarity measure is an interesting follow-up research question. For instance, in the social choice literature, the pairwise Kemeny distance is used by the Kemeny rule to define an aggregate ranking that best represents a collection of different rankings (Kemeny, 1959). Cosaert and Surana (2023) show that the distance between the aggregated ranking defined by the Kemeny rule and the input rankings is equivalent to the Swaps index (Apesteguia and Ballester, 2015) and can be used as a measure of group preference heterogeneity.

[7] Note that here it is implicitly assumed that none of the possible pairwise comparisons are strictly ordered in the sense that one bundle contains more of all goods than the other. If J of the total possible pairwise comparisons are strictly ordered, the normalized Kemeny distance can be defined by dividing the function with ($|Q|(|Q|-1)/2-J$). We account for this in my empirical application. The normalized Kemeny distance satisfies axiomatic properties $(i)-(iv)$, however, properties (v) and (vi) are not satisfied.

2.2 Revealed Preference

The Kemeny distance can be easily computed when individual preference rankings are known. However, preferences are typically unobserved. Instead, most datasets provide information about the bundles individuals choose in various budget situations. To identify the Kemeny distance in such empirical settings, it is necessary to infer the underlying preferences from the observed consumption behavior. The revealed preference framework is particularly useful in this context, as it allows preferences to be recovered nonparametrically. In what follows, we first describe the revealed preference characterization of utility-maximizing behavior and then focus on identifying the Kemeny distance.

Rationalizability. Consider a setting with N goods and a finite dataset $\mathcal{S} = \{(p_t, q_t)\}_{t \in T}$, where $T = \{1, \cdots, |T|\}$ represents the set of observations. For each observation $t \in T$, $p_t \in \mathbb{R}^N_{++}$ and $q_t \in \mathbb{R}^N_{+}$ denote the vectors of prices and chosen quantities, respectively.[8] The observed data $\mathcal{S}$ is considered "rationalizable" if there exists a continuous, strictly concave, and strictly monotone utility function such that, for each observation $t \in T$, q_t maximizes the utility function over all affordable bundles. Definition 2 describes this formally.

Definition 2 *A dataset $\mathcal{S} = \{(p_t, q_t)\}_{t \in T}$ is said to be rationalizable if there exists a continuous, strictly concave, and strictly monotone utility function $u : \mathbb{R}^N_+ \longrightarrow \mathbb{R}$ such that for all $t \in T$, $u(q_t) > u(q)$ where q $(\neq q_t)$ satisfies $p_t q \leq p_t q_t$.*

Next, we define the revealed preference relations. Given a dataset $\mathcal{S} = \{(p_t, q_t)\}_{t \in T}$, we say that q_t is directly revealed preferred to q_s (denoted as $q_t R^D q_s$) if $p_t q_s \leq p_t q_t$. We say that q_t is revealed preferred to q_s (denoted as $q_t R q_s$) if there exists a sequence $(u, v, \cdots, r \in T)$ such that $q_t R^D q_u, q_u R^D q_v, \cdots q_r R^D q_s$. We use these concepts to define the Strong Axiom of Revealed Preference (SARP).

Definition 3 *A dataset $\mathcal{S} = \{(p_t, q_t)\}_{t \in T}$ satisfies SARP if for all $s, t \in T$, $q_t R q_s$ and $q_t \neq q_s$ implies not $q_s R^D q_t$.*

In words, a dataset $\mathcal{S}$ satisfies SARP if for any two observed bundles q_t and q_s, if $q_t \neq q_s$ and q_t is revealed preferred to q_s then it must not be the case that q_s is directly revealed preferred to q_t. Matzkin and Richter (1991) have shown that SARP is a necessary and sufficient condition for the set $\mathcal{S}$ to be rationalizable.[9]

[8] We assume that individuals exhaust their budget when choosing the most optimal consumption bundle (i.e., $y_t = p_t q_t \quad \forall t \in T$), so all relevant information is contained in the prices and quantities data.

[9] For the sake of exposition, we focus on the definition of the Kemeny distance defined over strict preference rankings. This implies strict concavity of the utility function and validates the use of SARP conditions for

Recoverability. We can use SARP to recover preferences. Suppose we want to identify preferences over a finite set of bundles Q and we have an observed dataset $\mathcal{S}$. The set Q may contain several bundles that are not part of the observed data. Let us denote these bundles by $\{q_k\}_{k\in K}$. The first step in recovering preferences is to define mutually feasible support prices for each of those bundles in Q that are not in $\mathcal{S}$. We say that a set of prices $\{p_k\}_{k\in K}$ is mutually feasible if the prices and corresponding bundles are all together SARP consistent with the demand observations in $\mathcal{S}$.

Definition 4 *Given a dataset $\mathcal{S} = \{(p_t, q_t)\}_{t\in T}$ and a set of bundles $\{q_k\}_{k\in K}$, the set of mutually feasible support prices is defined as:*

$$S_K = \left\{ \{p_k\}_{k\in K} : \begin{array}{l} p_k > 0 \textit{ for all } k \in K \\ p_k q_k = 1 \textit{ for all } k \in K \\ \mathcal{S} \cup \{(p_k, q_k)\}_{k\in K} \textit{ satisfies SARP} \end{array} \right.$$

The mutual feasibility of support prices ensures that any candidate preference ordering over the set of bundles Q, which also rationalizes the observed data $\mathcal{S}$, can be represented by a continuous, strictly concave, and strictly monotone utility function. Appendix B provides further details on support prices and the requirement of their mutual feasibility. To bring these concepts into practice, we next present a mixed-integer linear programming (MILP) method to operationalize the solution set of mutually feasible support prices. Proposition 1 provides an MILP characterization of the conditions described in Definition 4.

Proposition 1 *Given a dataset $\mathcal{S} = \{(p_t, q_t)\}_{t\in T}$ and a set of bundles $\{q_k\}_{k\in K}$, a set of price vectors $\{p_k\}_{k\in K}$ is mutually feasible if and only if there exist $U_j \in [0, 1]$ and $R_{ij} \in \{0, 1\}$ for all $i, j \in K \cup T$ such that the following conditions are satisfied*

$$p_i > 0 \quad \textit{for all } i \in K \tag{1}$$

$$p_i q_i = 1 \quad \textit{for all } i \in K \tag{2}$$

$$U_i - U_j < R_{ij} \quad \textit{for all } i, j \in K \cup T \tag{3}$$

$$R_{ij} - 1 \leq U_i - U_j \quad \textit{for all } i, j \in K \cup T \tag{4}$$

$$p_i q_i - p_i q_j < M_i R_{ij} \quad \textit{for all } i, j \in K \cup T \tag{5}$$

$$q_{in} - q_{jn} \leq q_{in}(2 - R_{ij} - R_{ji}) \quad \textit{for all } i, j \in K \cup T \textit{ and } n \in \{1, \cdots, N\} \tag{6}$$

identifying the heterogeneity measure. It is worth noting that the definition of the Kemeny distance can be extended to allow for ties (see Kemeny, 1959), in which case the identification can be carried out by replacing SARP with the Generalized Axiom of Revealed Preference (GARP) conditions.

where M_i is any constant strictly greater than $p_i q_i$ and N is the number of consumption categories.

Appendix A contains the proof. Proposition 1 states that a set of vectors $\{p_k\}_{k \in K}$ is mutually feasible if and only if equations (1)-(6) are satisfied. In this formulation, the numbers U_j can be considered as the utility levels corresponding to the consumption bundles q_j. These utility numbers are constrained to be consistent with individual preference relations. The binary variables R_{ij} capture the order imposed by the utility numbers. Thus, for a pair of bundles q_i and q_j, $R_{ij} = 1$ if and only if $U_i \geq U_j$. The interpretation of the six conditions is as follows. Constraints (1) and (2) follow directly from the definition of mutual feasibility. Constraints (3)-(6) impose SARP on the combined set of observed ($t \in T$) and unobserved data ($k \in K$). Specifically, constraints (3) and (4) impose transitivity on the binary variables R_{ij}. Constraint (5) ensures that $R_{ij} = 1$ if bundle q_i is directly revealed preferred over q_j. Finally, by restricting $q_i = q_j$ whenever $R_{ij} = R_{ji} = 1$, constraint (6) ensures that SARP is satisfied.

2.3 Set Identification

So far, we have focused on recovering individual preference rankings. Now, we discuss how the recovered rankings can be compared between individuals to set-identify the Kemeny distance. Given a set of bundles Q, the recoverability method described above generally identifies incomplete individual preference rankings. Consequently, multiple complete extensions of the recovered individual preference relations may exist. The identification method described below uses all complete extensions of the recovered preferences to determine the feasible values of the Kemeny distance between individuals.

Assume we have a finite number of demand observations from each individual and that we want to compare preferences over a set of bundles Q. For any individual a with observed data $\mathcal{S}^a$, define a set of bundles $Q^a = \{q_{a_k}\}_{a_k \in a_K}$ containing all those bundles in Q that are not present in $\mathcal{S}^a$. For bundles $q_{a_k} \in Q^a$, we do not observe the price vectors at which the individual would have chosen these bundles, so we define mutually feasible support price vectors $\{p_{a_k}\}_{a_k \in a_K}$. This set of mutually feasible price vectors is characterized by the conditions in Proposition 1, where $\mathcal{S} = \mathcal{S}^a$ and $\{q_k\}_{k \in K} = Q^a$. The solutions to the MILP characterization in Proposition 1 produce rationalizable preference rankings of individual a over the set of alternatives Q. These preference rankings can be compared across individuals to construct lower and upper bounds on the heterogeneity measure. By requiring the mutual feasibility of price vectors, we ensure that the identification of the Kemeny distance is based on preference rankings that are rationalizable and consistent with the observed individual data.

Upper bound on $d(a, b)$. To find an upper bound on the Kemeny distance between the preferences of individuals a and b over the set Q, we make use of the theoretical restrictions of mutual feasibility of support prices. Specifically, given $\mathcal{S}^a$, $\mathcal{S}^b$ and Q, we solve the following optimization problem:

$$\text{maximize} \quad \sum_{\{i,j\ |i>j,\ q_i,q_j \in Q\}} D_{ij} \quad \text{such that}$$

$$R^a_{ji} + R^b_{ij} \leq 1 + D_{ij} \text{ for all } i > j \text{ and } q_i, q_j \in Q \tag{7}$$

$$R^b_{ji} + R^a_{ij} \leq 1 + D_{ij} \text{ for all } i > j \text{ and } q_i, q_j \in Q \tag{8}$$

$$R^a_{ji} + R^b_{ji} \leq 2 - D_{ij} \text{ for all } i > j \text{ and } q_i, q_j \in Q \tag{9}$$

$$R^a_{ij} + R^b_{ij} \leq 2 - D_{st} \text{ for all } i > j \text{ and } q_i, q_j \in Q \tag{10}$$

$$D_{ij} \in \{0, 1\} \tag{11}$$

where R^a_{ij} (R^b_{ij}) satisfy conditions (1)-(6) in Proposition 1, with $\mathcal{S} = \mathcal{S}^a$ ($\mathcal{S}^b$) and $\{q_k\}_{k \in K} = Q^a$ ($\{q_k\}_{k \in K} = Q^b$), as defined above.

In the above formulation, the binary variables D_{ij} can be interpreted as revealed preference conflict indicators, carrying a similar interpretation to the conflict indicators $d_{ij}(a, b)$ in the definition of the degree of preference heterogeneity (Definition 1). If individual a reveals a preference for q_j over q_i and b reveals a preference for q_i over q_j (or vice-versa), then D_{ij} equals 1 (constraints (7) and (8)). Otherwise, if the revealed preferences over these bundles agree, then D_{ij} equals 0 (constraints (9) and (10)). Aggregating the conflict indicators over all possible pairs of bundles yields the total pairs of bundles where preferences of a and b assign opposite rank orders to the bundles. The objective function maximizes the sum of all conflict indicators. Dividing the solution of this optimization problem with the total number of pairs ($\sum_{\{i,j\ |i>j,\ q_i,q_j \in Q\}} 1$) provides an upper bound for the Kemeny distance over all rationalizable preference rankings of a and b over the set of bundles Q.

Lower bound on $d(a, b)$. In an analogous way, given $\mathcal{S}^a$, $\mathcal{S}^b$ and Q, we can obtain a lower bound by solving the following optimization problem:

$$\text{minimize} \quad \sum_{\{i,j\ |i>j,\ q_i,q_j \in Q\}} D_{ij}.$$

subject to equations (7)-(11) in the above optimization program and equations (1)-(6) in Proposition 1.

Solving these two optimization problems set identifies the Kemeny distance between individual preferences.[10] The lower bound represents the smallest possible Kemeny distance between two individuals' (unobserved) preference rankings, given the preference relations revealed by their observed choices. The true preference rankings of the individuals would be at least as far apart as this lower bound. Conversely, the upper bound is the largest attainable Kemeny distance over all complete preference rankings compatible with the observed choices. The true, unobserved preference rankings would be closer than this upper bound. Together, the lower and upper bounds demonstrate what can be nonparametrically learned about preference heterogeneity from the observed choices.

We note that our heterogeneity measure has a limitation in that it depends on comparisons over the set Q. In the empirical application, we use the observed bundles to define this set, implying that our estimates are local sample estimates and subject to sampling uncertainty. To analyze the sensitivity of these estimates, the asymptotic behavior of the measure could be examined by drawing on the copula literature (see Nelsen, 2007 for an introduction). Such a framework would allow for a nonparametric statistical test to determine whether the preferences of two individuals differ. Further development of the method to allow for inference is an interesting avenue for future research. The linear programming nature of our method may also prove useful in this regard. For instance, Kaido, Molinari, and Stoye (2019) propose a bootstrap-based procedure for inference on parameters that are (partially) identified through linear moment inequalities (see also Bugni, Canay, and Shi, 2017; Fang, Santos, Shaikh, and Torgovitsky, 2023).

Cost efficiency indices. Sometimes, empirical applications of revealed preference methods require small deviations from exact rationality conditions. This may be necessary, for example, when individual behavior does not satisfy the strict SARP conditions but is close to satisfying them.[11] The most common approach to allowing for small deviations from SARP

[10] Based on the Kemeny distance, Ha and Haddawy (2003) proposed a probabilistic distance for partial rankings. The probabilistic distance is defined as the average Kemeny distance of all possible weak-order extensions of the given partial rankings (see also Kidwell, Lebanon, and Cleveland, 2008). They note the impracticality of computing this measure directly and instead propose approximation algorithms based on Monte Carlo simulation as an alternative. The methodology proposed in this paper complements the probabilistic distance measure. In principle, the probabilistic distance would lie between the lower and the upper bound obtained from the proposed methodology.

[11] This is a common finding in the literature. Using a variety of goodness-of-fit indices, the empirical literature often finds that even when the proportion of individuals who are "perfectly rational" is small, most individuals are close to satisfying strict rationality conditions. For instance, using the Critical Cost Efficient Index, Choi, Fisman, Gale, and Kariv (2007) report that, on average, only 4.6% of expenditure loss is needed to rationalize the observed behavior. Similarly, Echenique, Lee, and Shum (2011) find that while

is to define the Afriat Critical Cost Efficiency Index (CCEI), $0 < e \leq 1$ (Afriat, 1973). Intuitively, the CCEI measures the minimum adjustment to total expenditures needed to remove all SARP violations. In particular, the optimization problems to obtain lower and upper bounds on the Kemeny distance can be easily modified to allow for some inefficiency in observed behavior. Specifically, to permit a level of efficiency e, we replace constraints (5) and (6) in Proposition 1 with equations (5a) and (6a), respectively,

$$ep_iq_i - p_iq_j < M_iR_{ij} \text{ for all } i, j \in K \cup T \tag{5a}$$

$$eq_{in} - q_{jn} \leq q_{in}(2 - R_{ij} - R_{ji}) \text{ for all } i, j \in K \cup T \text{ and } n \in \{1, \cdots, N\} \tag{6a}$$

where e is a given pre-specified efficiency level.

Illustrative example. We now demonstrate how set-identification works in practice with a simple example. Consider individuals a and b, whose demand observations are visualized in Figure 1. For individual a (solid budgets), we have two observations corresponding to the bundles q_1 and q_2. Individual b's data (dashed budgets) correspond to the bundles q_3 and q_4. Let $Q = \{q_1, q_2, q_3, q_4\}$ and assume that the preference rankings of a and b are $q_4 \succ_a q_2 \succ_a q_3 \succ_a q_1$ and $q_1 \succ_b q_4 \succ_b q_3 \succ_b q_2$, respectively. By Definition 1, the degree of preference heterogeneity between a and b is 0.67 (see columns 2-4 in Table 1).

Both individuals' observed demands are SARP consistent and align with their true preferences. Using revealed preference arguments, we can infer some preference rankings. For individual a, we can conclude that $q_2 \succ q_3$ and $q_4 \succ q_3$ (by monotonicity). For individual b, we can conclude that $q_3 \succ q_2$, $q_4 \succ q_2$ and $q_4 \succ q_3$. It is evident that rationalizability conditions typically recover only partial rank orders. By considering all possible complete extensions of the recovered partial orders, we can compute lower and upper bounds on the degree of preference heterogeneity. As illustrated in Table 1, all complete extensions imply a lower bound of 0.17, while at least one pair of complete extensions implies an upper bound of 0.83.

3 Empirical Application

We apply the method to a sample of U.S.-based household panel data from the 1999–2019 waves of the Panel Study of Income Dynamics (PSID). The PSID, collected biennially since

only 20% of households satisfy rationality conditions, the severity of these violations, as measured by the Money Pump Index, is small - about 6% of total expenditure. Dean and Martin (2016) find that although only 29% households are perfectly rational, the cost of rationalizing irrational households, as measured by the Minimum Cost Index, is just 0.08% of total expenditure.

Figure 1: Illustrative example

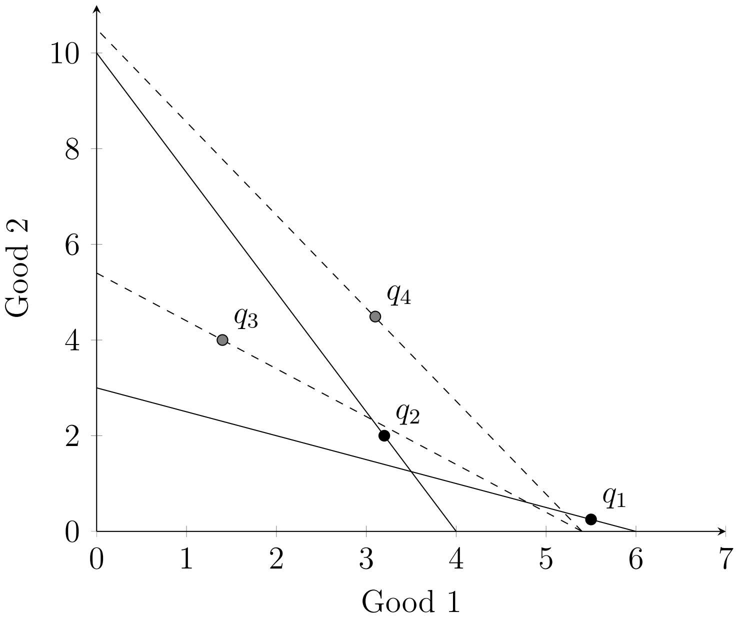


Table 1: Identifying preference heterogeneity

	true preferences and distance			revealed preferences and estimated bounds			
	a	b	$d(a,b)$	a	b	lower bound	upper bound
(q_1,q_2)	$q_2 \succ_a q_1$	$q_1 \succ_b q_2$	1	-	-	0	1
(q_1,q_3)	$q_3 \succ_a q_1$	$q_1 \succ_b q_3$	1	-	-	0	1
(q_1,q_4)	$q_4 \succ_a q_1$	$q_1 \succ_b q_4$	1	-	-	0	1
(q_2,q_3)	$q_2 \succ_a q_3$	$q_3 \succ_b q_2$	1	$q_2 \succ_a q_3$	$q_3 \succ_b q_2$	1	1
(q_2,q_4)	$q_4 \succ_a q_2$	$q_4 \succ_b q_2$	0	-	$q_4 \succ_b q_2$	0	1
(q_3,q_4)	$q_4 \succ_a q_3$	$q_4 \succ_b q_3$	0	$q_4 \succ_a q_3$	$q_4 \succ_b q_3$	0	0
			0.67			0.17	0.83

1968, is a nationally representative survey of over 5,000 households in the United States. It includes a rich set of economic and socio-demographic information. Since 1999, the dataset has also included data on consumption expenditures. In what follows, we first discuss the sample selection criteria. Next, we assess whether, and to what extent, individual observed behavior is rationalizable. Finally, we identify the distance-based heterogeneity measure between individual preferences.

3.1 Data

The sample is subject to the following selection criteria. First, the analysis focuses on singles. Any household where the head is married or cohabiting at any point of observation is excluded. Couples are omitted to avoid preference aggregation issues within households with multiple decision-makers.[12] Next, we focus on individuals on the intensive margin of labor supply—that is, those actively participating in the labor market during each period of study. We also exclude individuals whose demand observations imply an obvious complete preference ranking due to monotonicity (i.e., when either $q_t \leq q_s$ or $q_t \geq q_s$ for all t, s). Finally, after removing individuals with missing basic information (e.g., wage, time use, education), we obtain 2,343 observations (213 individuals observed 11 times).

We assume a simple labor supply setting in which individuals allocate their entire potential income to two consumption categories: leisure and a Hicksian aggregate.[13] Leisure quantities are calculated under the assumption that individuals require 8 hours per day for personal care and sleep. Daily leisure hours are defined as the time available for labor market participation that was not spent working (i.e., 24 - 8 - hours spent on market work). Hicksian consumption represents daily expenditures on market goods and is calculated as the sum of household spending on food, housing, transportation, education, childcare, and healthcare. The price of leisure is assumed to be the individual's hourly wage, while the price of the Hicksian aggregate is normalized to 1. Appendix D.1 provides additional details about the selected sample. In the following analysis, we use the first eight observations (1999–2013) to

[12] Imposing a unitary assumption on a multi-person household models the household as a single decision-maker. Although convenient for intrahousehold welfare analysis, this approach can lead to misleading conclusions. Moreover, a growing body of empirical research has shown that this assumption is often rejected by observed multi-member household data (see, for example, Browning and Chiappori, 1998; Dauphin, El Lahga, Fortin, and Lacroix, 2011). Extending the proposed methodology to collective household models is a promising avenue for follow-up research. A starting point for such an extension could be integrating the proposed methodology with the revealed preference characterization of rational consumption for collective household models developed by Cherchye, De Rock, and Vermeulen (2007, 2011).

[13] Similar settings have been considered by Manski (2014) and Cherchye, Demuynck, and De Rock (2014). Although simple, this setup effectively demonstrates the main message of the proposed methodology. Moreover, Manski (2014) shows that ignoring unobserved preference heterogeneity could lead to fallacious conclusions even in this simple setting.

identify the distance-based preference heterogeneity measure. The remaining three observations (2015–2019) are used in Section 4 to evaluate the out-of-sample predictive performance of the method.

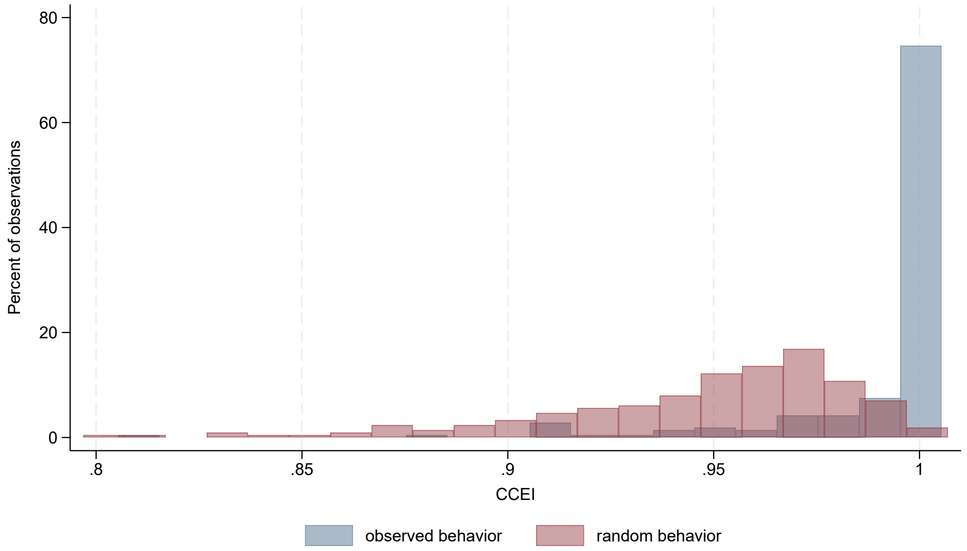
3.2 Individual Rationalizability

We begin by empirically investigating the assumption of individual rationality. Rationality can be tested by checking whether the observed individual data satisfy SARP. However, the SARP test is strict in nature; the data will either pass or fail. When an individual's behavior is irrational, it is useful to empirically assess the degree of violation. To do so, we use the CCEI, which measures how closely the data align with exact rationality conditions. The CCEI ranges from zero to one, where a value of one indicates that the data fully satisfy the strict conditions, and lower values reflect more severe violations of SARP.

We find that approximately 66% of individuals are consistent with the strict conditions (i.e., their CCEI is equal to one), and around 80% are very close to satisfying SARP (with CCEI values greater than 0.99). The high average CCEI value of 0.99 indicates that even if the observed individuals are not perfectly consistent, they are very close to meeting the rationalizability conditions. Table 8 in Appendix D.2 provides further details. Overall, these results strongly support the rationality assumption in the selected sample. In most cases, only small adjustments to the data are needed to achieve consistency with rationality. However, there is no strong consensus in the literature regarding a specific CCEI threshold for defining rationality. Varian (1991) suggests using a threshold of 0.95, but this remains a subjective choice. As an alternative, we follow Bronars (1987) by comparing the CCEI values of observed behavior with those derived from simulated irrational behavior. Becker (1962) proposes simulating irrational behavior by randomly selecting quantity bundles from a uniform distribution over all bundles on the budget hyperplane. For each individual, we simulate 1,000 random responses based on their budgets and compute the CCEI value for each simulated data point.

Figure 2 shows the distribution of CCEI values for the observed individuals in the sample alongside the distribution of CCEI values for simulated random behavior. The horizontal axis represents various CCEI values, and the vertical axis indicates the percentage of observations corresponding to each value. The histograms clearly demonstrate that observed behavior is much closer to rationality than randomly generated behavior. While a small proportion of individuals are not perfectly rational, they are still close to satisfying the strict conditions. In the analysis that follows, we focus on the subsample of 171 individuals who are almost rational (i.e., those with a CCEI greater than 0.99).

Figure 2: Distribution of CCEI values in the observed and simulated random data



3.3 Revealed Preference Heterogeneity

The MILP method outlined in Section 2.3 provides a simple and practical approach for set identifying the degree of preference heterogeneity between individuals. To apply the method to the data, we must define the set of alternatives Q over which preferences will be compared. In principle, rankings can be compared over any set of alternatives deemed suitable for preference comparison; however, specifying this set in practice can be challenging. For this application, we consider pair-specific sets of alternatives. Specifically, when comparing the preferences of individuals a and b, we define the set of alternatives as the union of their observed demands (i.e., $Q = \{q_t^a\}_{t \in T} \cup \{q_t^b\}_{t \in T}$).[14] Additionally, we exclude preference comparisons for pairs of bundles that would be ranked identically by any preference due to monotonicity. This prevents two individuals with nested budget sets from being falsely identified as having similar preferences.

[14] An alternative approach could define Q as the set of all bundles observed in the sample ($Q = \cup_i \cup_t \{q_t^i\}$). However, in a typical survey setting where the number of consumers far exceeds the number of observations per consumer (e.g., 171 versus 8 in the current setup), this choice would require a large number of unobserved price vectors (see Proposition 1). While theoretically feasible, this approach would be computationally expensive to implement in practice. At this point, we note that, while we have focused on the classical data setting involving choices made under linear budget conditions, it is possible to extend the method to incorporate choices made under more complex settings, such as nonlinear budget sets, discrete consumption spaces, or dynamic contexts. To take a specific example, for choices made under risk and uncertainty, the Generalized Restriction of Infinite Domains (GRID) method proposed by Polisson, Quah, and Renou (2020) could be used to construct the preference comparison set and identify the Kemeny distance. Although we acknowledge that these options represent valuable extensions of the method used in the current paper, we leave their exploration for future research.

Table 2 summarizes the identified bounds. Columns 2 and 3 present the lower and upper bounds, respectively, while column 4 ("difference") shows the difference between the upper and the lower bounds. The results illustrate what can be learned about preference heterogeneity under the minimalistic assumption of rationality. The average lower bound is 3.73%, and the average upper bound is 64.21%.[15] This corresponds to an average tightness of 60.48 percentage points.[16]

Table 2: Degree of preference heterogeneity

	lower bound	upper bound	difference
mean	0.0373	0.6421	0.6048
sd	0.0627	0.1590	0.1635
min	0.0000	0.1333	0.0600
p25	0.0000	0.5270	0.4848
p50	0.0000	0.6441	0.6042
p75	0.0500	0.7595	0.7222
max	0.6500	1.0000	1.0000

The lower and upper bounds indicate the extent to which the assumption of preference homogeneity can be empirically supported. Smaller values for both the lower and upper bounds provide stronger evidence of preference similarity between individuals. The lower bounds represent the best-case scenario for the preference homogeneity assumption between individuals. For example, a lower bound of zero indicates that there exists at least one common preference ranking that is rationalizable with both individuals' observed choices. The lower bound estimates in column 2 strongly reject the complete homogeneity assumption, as the minimal degree of heterogeneity for some pairs is as high as 65%. The upper bounds, on the other hand, indicate the maximum preference heterogeneity that can be present given the observed individual choices. The estimates for the upper bounds in column 3 show that even among individuals who could feasibly be modeled as having the same preferences (if the lower bound is zero), it is possible that they may have very dissimilar preferences, as

[15] The recovered bounds reveal considerable variation in patterns of preference heterogeneity. We explore this further in Appendix D.3 by relating the estimates to observed characteristics. This exploratory analysis provides insights into the factors driving differences in individual preferences. The regression estimates suggest that age and gender are significant factors explaining preference heterogeneity.

[16] Admittedly, there is room for further improvement in the bounds to obtain reasonably tight heterogeneity estimates. One possibility is to impose additional assumptions on the nature of demand. For instance, recent studies have shown that imposing mild normality assumptions can substantially improve recovery compared to using rationality restrictions alone (see, e.g., Blundell, Browning, and Crawford, 2003, 2007, 2008 and Cherchye, Demuynck, De Rock, and Surana, 2020). While we do not explore this approach in the current paper, it could be an interesting topic for future research.

reflected by a high upper bound estimate.

The results in Table 2 sharply contrast with existing revealed preference measures, which are binary in nature: individuals are either classified as having different preferences or not. In contrast, our method identifies the extent to which preferences are revealed to be heterogeneous. To further illustrate this point, in Appendix F.2, we provide a theoretical and empirical comparison between our method and a goodness-of-fit-based heterogeneity measure proposed by Castillo and Freer (2018). We show that, although the two measures are correlated (as they should be, given that they aim to quantify the same phenomenon), there are individuals who are not revealed as different by the latter method but exhibit large distance-based heterogeneity according to our method.

4 Demonstrations

We demonstrate the practical usefulness of our heterogeneity measure through three applications of consumer demand analysis. The first two applications address the frequently cited criticism that nonparametric revealed preference methods have low empirical power, while the third application focuses on concerns about functional form assumptions in parametric methods. For all three applications, we use the identified bounds on the heterogeneity measure to cluster the sample into similar preference types. The results from the first two applications show that employing these preference types as units of analysis leads to a substantial improvement in the identifying power of nonparametric demand predictions and welfare evaluations. The results from the third application illustrate that using the constructed preference types can serve as a helpful starting point for assessing whether a particular functional form assumption is appropriate. In the following, we begin by discussing the use of estimated bounds to form similar preference types. After that, we discuss the three applications.

4.1 Similar Preference Types

As individual panel datasets are usually short, empirical applications often address the issue of data scarcity by assuming preference homogeneity among groups of individuals. Homogeneity allows observations to be pooled across individuals as if they were maximizing the same utility function. Consequently, the choice behavior of one person within a group can be used to predict the behavior of another group member. Naturally, a key question is which individuals should be grouped together. One approach is to group observationally similar individuals, assuming that those with similar observable characteristics also have similar preferences. However, this assumption has been shown to lack empirical support and often results in low R^2 values when modeling the observed demand distribution (see Banks, Blundell, and Lewbel, 1997; Lewbel and Pendakur, 2009). This highlights the appeal of directly using the revealed degree of preference heterogeneity when identifying "comparable" individuals.

To form similar preference types, we generalize the algorithm proposed by Cosaert (2019), which uses revealed preference conditions to assign individuals to preference types. Their algorithm follows a two-step procedure. In the first step, the minimum number of types required to partition the sample is identified. In the second step, observed individual characteristics are used to assign individuals to types so that observably similar individuals are more likely to be grouped together. We generalize their algorithm in two ways. First, instead of using strict revealed preference conditions to determine which individuals cannot belong to the same type, we use a cardinal measure (the estimated lower bounds) to determine whether two individuals can be grouped into the same type. Second, rather than relying on observed characteristics, we utilize the estimated upper bounds on preference heterogeneity to assign individuals to types. We elaborate on both steps below.

Step 1: finding the number of preference types. The problem of determining the number of preference types can be reformulated as a graph coloring problem, a well-known problem in computer science. A graph $G = (V, E)$ is defined as a combination of a set of nodes (V) and a set of edges (E) between pairs of nodes. The graph coloring problem seeks to find the minimum number of colors needed to color all the nodes of a graph such that no two adjacent nodes (i.e., nodes connected by an edge) are assigned the same color. By definition, any solution to this graph coloring problem ensures that no two nodes assigned the same color are connected to each other. The smallest number of colors required to solve this graph coloring problem is known as the chromatic number of the graph.

To draw an analogy between graph coloring and the problem of determining the number of preference types, consider each individual as a node and each preference type as a color. To construct the graph, draw an edge between two individuals if they must belong to different preference types. In our application, two individuals are required to belong to different types if the lower bound of their degree of heterogeneity is strictly positive. This ensures that individuals must belong to different types when no common utility function can rationalize their observed behavior. The chromatic number of this graph represents the smallest number of preference types needed to partition the observed individuals. Applying this procedure to our sample of 171 individuals results in 29 preference types.

Three remarks are in order. First, the algorithm proposed by Cosaert (2019) focuses on a two-goods setting and requires two individuals to be assigned to different types if their combined data exhibit WARP violations. In a two-goods setting, this strict requirement of no WARP violations ensures that all observations within a type are rationalizable.[17] By contrast, we use our estimated lower bounds on preference heterogeneity to determine who cannot belong to the same type. While our application also considers a two-goods setting, the algorithm is applicable to more general settings. Second, because Cosaert (2019) focuses on crosssectional data, all individuals are rational by construction. In contrast, we consider a longitudinal setting and allow individuals to exhibit small deviations from rationality. As such, our algorithm cannot guarantee within-type rationalizability. Instead, its aim is to construct preference types that are "almost" rationalizable. Finally, because we use a cardinal measure (the lower bound), our algorithm provides researchers with a natural criterion for adjusting the threshold used to determine who cannot belong to the same type. This may be necessary, for instance, if subsequent analysis requires a smaller number of types for computational reasons. Naturally, as larger lower bound values are permitted, fewer types are required to partition the sample. Figure 7 in Appendix D.4 shows how the number of required preference types changes with alternative definitions of incompatible individuals. In Appendix E.2, we conduct a robustness check using an alternative approach to determine the number of preference types. Specifically, we require two individuals to belong to different types if the lower bound on their preference heterogeneity measure is greater than 1%. The results from this robustness check are similar to those in the subsequent applications.

Step 2: allocation of individuals to preference types. Although the chromatic number of a graph is unique, there may be multiple ways to color the graph. Consequently, there can be several ways of allocating individuals to different types. A natural question that follows is: Given the number of preference types, how should individuals be allocated to different types?

In the absence of a cardinal measure of preference heterogeneity, Cosaert (2019) proposed minimizing the variation in observed characteristics to determine type assignments. The underlying idea is that observed characteristics may partially capture unobserved preference heterogeneity. In contrast, since our measure of heterogeneity is cardinal, we can use it directly, rather than relying on observed characteristics, to assign individuals to types. In this application, we use the estimated upper bounds on preference heterogeneity. Specifically, we adopt a minimax strategy, which minimizes the sum of within-cluster upper bounds under the restriction that the lower bounds on preference heterogeneity among individuals within

[17] This follows from the fact that, in a two-goods setting, WARP = SARP and transitivity has no testable implication (see Rose, 1958; Cherchye, Demuynck, and De Rock, 2018).

each type are zero.

To formalize the problem of constructing preference types, let s_{at} be a binary variable that indicates whether individual a belongs to type t. Constructing preference types based on the estimated bounds boils down to solving the following optimization problem.

$$\{s^*_{at}\} = \arg\min_{s_{at}, s_{bt}} \sum_{t \leq \tau} \sum_{a,b} \overline{d}(a,b) s_{at} s_{bt} \ \text{ such that}$$

$$\forall a : \sum_{t \leq \tau} s_{at} = 1 \tag{12}$$

$$\forall a, b \text{ and } \forall t \leq \tau : \ \underline{d}(a,b) > 0 \Rightarrow s_{at} + s_{bt} \leq 1 \tag{13}$$

$$\forall a, \text{ and } \forall t \leq \tau : \ s_{at} \in \{0,1\} \tag{14}$$

where τ is the minimum number of types required (as estimated in step 1), and $\overline{d}(a,b)$ and $\underline{d}(a,b)$ are the upper and lower bounds on the degree of preference heterogeneity between individuals a and b. The objective function minimizes the sum of within-type upper bound of the Kemeny distance.

Constraint (12) ensures that each individual belongs to exactly one type, while constraint (13) enforces that a and b must belong to different types if their lower bound degree of heterogeneity is strictly positive. Solving this optimization problem yields a partition of individuals such that the sum of within-type upper bounds on the degree of heterogeneity is minimized, and the lower bounds between any two individuals are zero.[18] In what follows, we refer to these types as the "distance-based" types.

To evaluate the effectiveness of constructing distance-based types, we compare them with a more conventional method of forming types based on observable characteristics. For simplicity, we keep the number of preference types the same as above (i.e., 29) and focus on a unidimensional criterion—age—to construct preference types.[19] Specifically, we use a standard clustering algorithm with age differences between individuals as the distance criterion to be minimized within each type. In the following, we refer to these as the "age-

[18] In practice, forming the distance-based preference types boils down to solving a k-medoid constrained clustering problem. In this problem, incompatible individuals are constrained to belong to different clusters, and a distance function (here, upper bounds) is used as the criterion for selecting the optimal cluster. Finding an exact solution to a constrained clustering problem is feasible only for small sample sizes. For larger samples, we use greedy algorithms to approximate a solution. Appendix C.2 provides further discussion and presents the greedy algorithm used in the following analysis.

[19] A number of studies have identified age as a crucial determinant in explaining preference heterogeneity (see, e.g., Andersen, Harrison, Lau, and Rutström, 2010; Von Gaudecker, Van Soest, and Wengström, 2011; Falk, Becker, Dohmen, Enke, Huffman, and Sunde, 2018), making it a useful benchmark for comparison with distance-based types. Of course, preference types can be constructed using any set of observable characteristics. In Appendix E.1, we present a robustness check where we use age, gender and/or education to construct observable-characteristics-based types. We obtain similar qualitative conclusions.

based" types. Appendix D.4 shows the distribution of the sizes of distance-based and age-based preference types.

4.2 Out-of-Sample Demand Prediction

The first application focuses on forecasting demands for counterfactual budgets. Theory-consistent demands can be nonparametrically (set) identified using rationality restrictions (see Varian, 1982 for further details). We conduct out-of-sample demand predictions for counterfactual budgets defined using the last three observations of each individual (years 2015 to 2019). Specifically, given a counterfactual budget for an individual, a predicted demand interval is constructed by identifying feasible values of the budget share on leisure that are consistent with the data from all individuals within their preference type. In a two-goods setting, this automatically identifies the feasible budget shares for the Hicksian good.

We compare the predicted distance-based and age-based demand intervals with the observed demands to evaluate the predictive performance of the two preference-type constructions. Since the predictions are intervals of feasible budget shares, multiple aspects must be considered when evaluating predictive performance. First, the interval may or may not include the observed demand. If the observed demand is included in the predicted interval, the method successfully forecasts the demand; otherwise, it fails. Therefore, the number of correctly predicted intervals is an important criterion for evaluating an approach. However, it is equally important to consider how informative (or tight) the predicted intervals are. A predicted interval as wide as [0,1] will certainly contain the observed budget share, but it is not informative. On the other hand, a tightly identified predicted set that just misses the observed demand may be preferable to a successful prediction that is as wide as the naive bound (i.e., [0, 1]).

To assess both the correctness and informativeness of the predictions, we define δ^{upper} (δ^{lower}) as the difference between the estimated upper (lower) bound of the predicted demand interval and the observed demand. Specifically, $\delta^{upper} = s^{upper} - s^*$ and $\delta^{lower} = s^{lower} - s^*$, where s^{upper} and s^{lower} represent the upper and lower bounds of the predicted interval, and s^* is the observed demand (in terms of the budget share on leisure). Clearly, if zero lies within the interval $[\delta^{lower}, \delta^{upper}]$, the predicted set contains the observed demand. Otherwise, the farther the interval is from zero, the greater the inaccuracy of the prediction. Furthermore, the difference $\delta^{upper} - \delta^{lower}$ reflects the tightness of the predicted set. Ideally, we want the predictions such that the intervals $[\delta^{lower}, \delta^{upper}]$ are both small and include zero (or are close to zero).

Figure 3 shows the δ^{upper} and δ^{lower} measures based on predictions from distance-based types (top panel) and age-based types (bottom panel). The filled markers correspond to δ^{upper}, while the hollow markers correspond to δ^{lower}. To help visualize the results, the predictions are sorted by the tightness of the predicted intervals along the x-axis. In each plot, gray markers represent predictions that failed to capture the observed demand, while the colored (blue or red) markers indicate predictions that include the observed demand.

As mentioned above, two important factors must be considered: informativeness and correctness. For informativeness, note that predictions situated towards the left-hand side represent tighter predictions while those on the right-hand side, with a tightness value of one, lack predictive power. The informative nature of distance-based predictions is clearly evident. While a large majority of distance-based predictions exhibit less than a 20 percentage point difference between δ^{upper} and δ^{lower}, very few age-based predictions are similarly tight. Most age-based predictions show a difference of more than 80 percentage points, with a significant number as wide as the naive bounds.

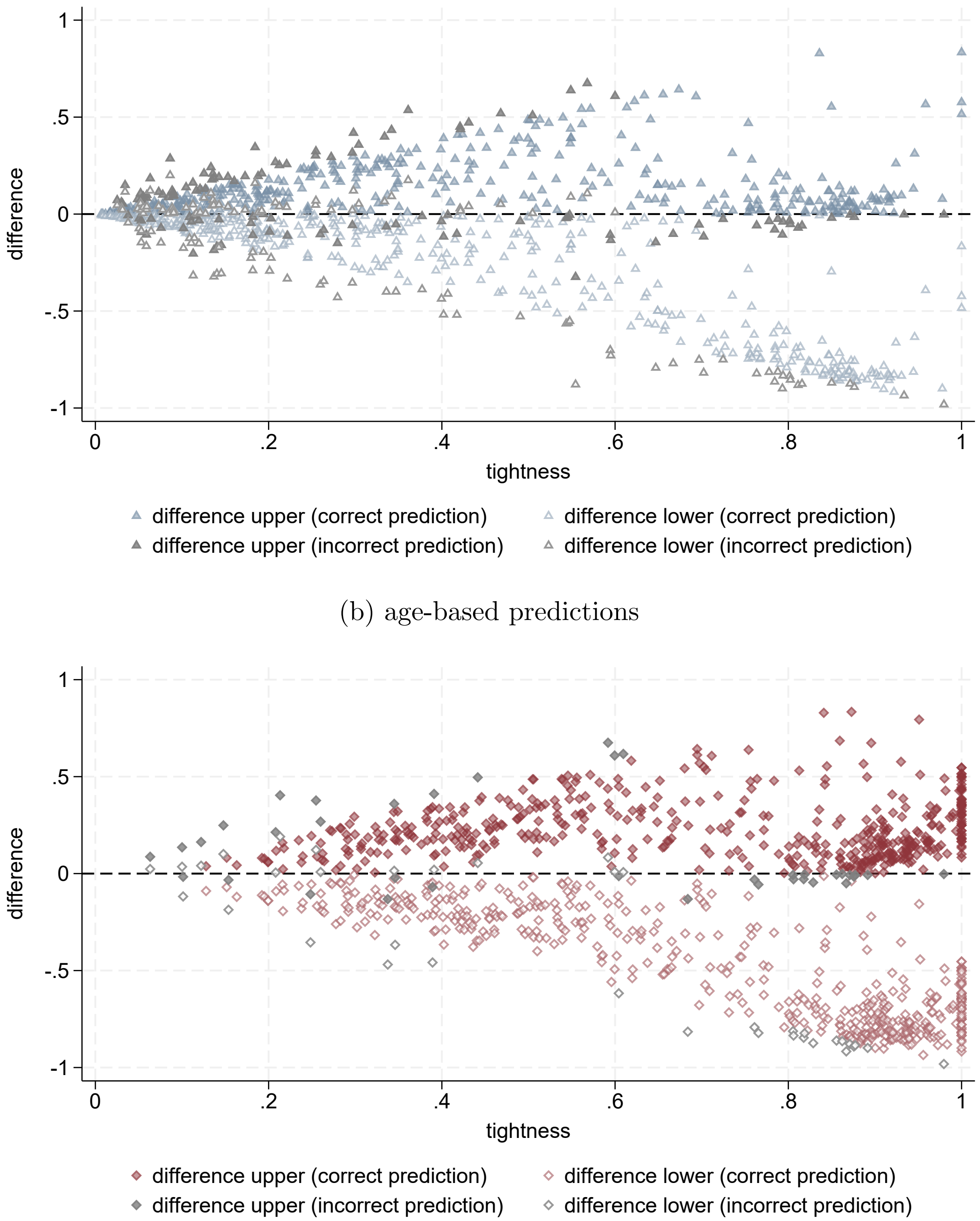
Next, to assess the correctness of demand predictions, note that if the dashed horizontal line through zero passes between δ^{upper} and δ^{lower}, the predicted set is correct. Otherwise, the father these values are from the dashed line, the greater the inaccuracy. In the plots, incorrect predictions are colored gray. While there are more incorrect distance-based predictions compared to age-based predictions, these predictions are still close to the target, indicating they narrowly missed the observed demand. Table 10 in Appendix D.5 summarizes the information presented in Figure 3 by computing the tightness of the predicted demand intervals and the distance between the observed demands and the predicted intervals. The results confirm the patterns observed in the graphical analysis.

4.3 Welfare Analysis

Another important application of demand analysis is determining how changes in budget situations affect an individual's well-being. Researchers and policymakers are interested in knowing not only which price-income combination is preferred but also "by how much" one situation is preferred over the other. Revealed preference axioms can be used to bound the welfare effect of a budget change. Although this nonparametric approach is conceptually appealing, its practical usefulness is often questioned due to the low empirical bite it offers. In this second application, we show that the distance-based types can help obtain more informative estimates when evaluating welfare effects.

To measure "by how much" one budget is preferred over the other, we use the money metric utility concept introduced by Samuelson (1974). Consider the budget in observation

Figure 3: Out-of-sample demand predictions



(a) distance-based predictions

(b) age-based predictions

t, (p_t, y_t), and a reference price-income regime (p_r, y_r). The money metric utility function gives the minimum expenditure required in observation t (with prices p_t) to be as well off as in the reference regime (with prices p_r and income y_r). Formally, it is defined as

$$\mu(p_t; p_r, y_r) \equiv e(p_t, v(p_r, y_r)),$$

where $e(p, u)$ is the expenditure function, which quantifies the minimum income needed to obtain utility u at prices p and $v(p, y)$ is the indirect utility function, quantifying the maximum utility achievable with income y at prices p. We can use the money metric utility function to define a cost-of-living index,

$$c_{t,r} = \frac{\mu(p_t; p_r, y_r)}{y_t}.$$

The cost-of-living index, $c_{t,r}$, represents the proportion of an individual's budget in observation t that is needed to as well off as in the reference regime. If $c_{t,r}$ is greater than one, the individual needs more money than available in observation t to be as well off as in the reference regime. Thus, they are worse off in observation t compared to the reference regime. Otherwise, if $c_{t,r}$ is less than one, they are better off in observation t than in the reference regime.

We use the cost-of-living index to quantify the welfare effects of the 2008 financial crisis. For each individual, we identify $c_{2011,2007}$ to estimate the difference in living costs between 2007 and 2011. In words, it measures the fraction of an individual's 2011 income needed in 2011 to be as well off as they were in 2007. The revealed preference characterization of rational demand behavior can be used to nonparametrically identify the cost-of-living index. The method provides set identification by obtaining upper and lower bounds on $c_{2011,2007}$, effectively capturing all feasible values of the cost-of-living index.[20]

To identify $c_{2011,2007}$ for a given individual, we use the demand observations of all individuals within their type. As before, we consider two types: distance-based and age-based. Table 3 summarizes the estimated lower and upper bounds for the sample of individuals under consideration. Columns 2-4 show the results for the distance-based types, and columns 5-7 show the results for the age-based types. Columns Δ_d and Δ_a show the difference between the upper and lower distance-based and age-based bounds. These quantify the identifying power that follows from each method of constructing types. Finally, the last column reports the relative difference between Δ_d and Δ_a, quantifying the extent to which distance-based

[20] We adapt the nonparametric method developed by Varian (1982) to compute the lower and upper bounds on the cost-of-living indices. The identification procedure involves solving optimization problems with linear objectives and linear inequality constraints (Appendix C.3 provides further details).

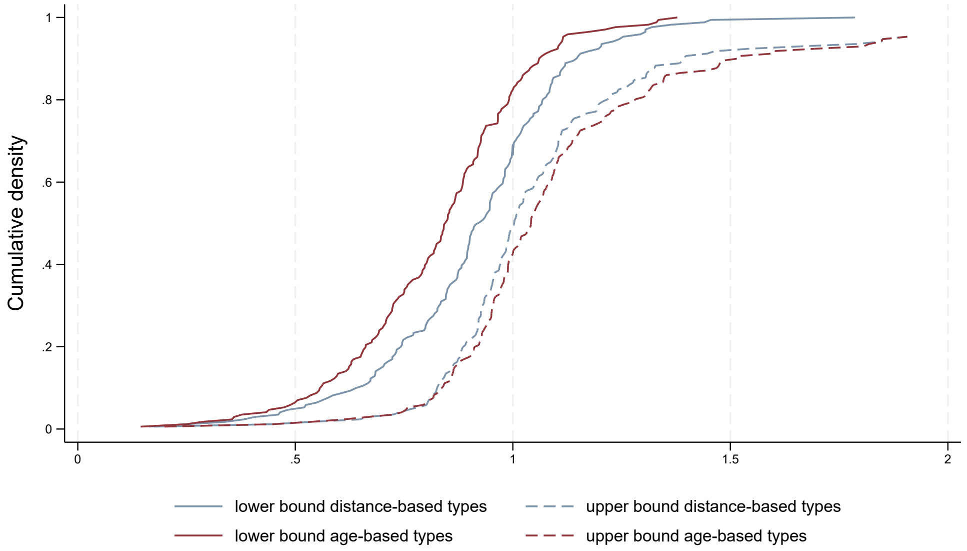
bounds are tighter than age-based bounds.

Table 3: Bounds on $c_{2011,2007}$

	distance-based types			age-based types			
	lower	upper	Δ_d	lower	upper	Δ_a	$\frac{\Delta_a - \Delta_d}{\Delta_a}$
mean	0.912	1.172	0.260	0.828	1.202	0.374	0.450
sd	0.224	0.978	0.935	0.208	0.970	0.962	0.452
min	0.164	0.175	0.001	0.145	0.199	0.025	-1.345
p25	0.800	0.919	0.026	0.704	0.939	0.120	0.172
p50	0.924	1.003	0.077	0.846	1.042	0.185	0.542
p75	1.040	1.140	0.202	0.966	1.213	0.356	0.813
max	1.786	12.530	11.277	1.378	12.530	11.808	0.996

The results show that the distance-based bounds are significantly tighter than the age-based bounds. The mean (median) difference between the upper and lower bounds estimated using the distance-based types is 0.26 (0.08) while for the age-based types, it is 0.37 (0.18). On average, the distance-based bounds are 45% tighter compared to the age-based bounds and the relative difference between Δ_d and Δ_a is more than 54% for about half of the sample. Figure 4 shows the empirical cumulative distribution functions of distance-based and age-based lower and upper bonds for $c_{2011,2007}$. In line with the results in Table 3, we find that the distance-based CDFs are much closer to each other than the age-based CDFs.

Figure 4: CDF of distance-based and age-based bounds



As a follow-up exercise, we use the estimated cost-of-living index to determine whether an individual is better-off or worse-off in 2011 compared to 2007. An individual is defined as better off in 2011 if the identified upper bound is less than one, while they are worse-off if the identified lower bound is greater than one. If the lower bound is less than one and the upper bound is greater than one, we cannot robustly conclude whether they are better or worse off in 2011.

Table 4 shows the fraction of individuals in the sample that are classified as better-off, worse-off or unclassified based on the identified upper and lower bounds on $c_{2011,2007}$. Columns 2 and 3 show the classification based on the bounds identified from the distance-based and age-based types, respectively. Using the distance-based bounds, 48.54% are classified as better-off, 30.99% are worse-off, while a residual of 20.47% cannot be classified in either category. By contrast, the age-based bounds classify 42.11% as better-off and 17.54% as worse-off. The fraction of individuals that cannot be classified in either category is much larger 40.35%. These results support the findings in Table 3, showing that the bounds obtained through the distance-based types lead to a significantly more informative classification of individuals after the crisis period.

Table 4: Worse-off and better-off individuals

	distance-based types	age-based types
better-off in 2011	48.54	42.11
worse-off in 2011	30.99	17.54
cannot say	20.47	40.35

4.4 Parametric Estimation

While the two applications above focused on the nonparametric approach of demand analysis, the third application focuses on the parametric approach, wherein one assumes a specific functional form for the utility function. Observed data are used to estimate the parameters of the utility function, which can then be used to point-identify parameters of interest, such as income compensation, the cost-of-living index, or demand in a counterfactual situation. A crucial concern with this parametric approach is that it relies heavily on prior assumptions about the functional form of individual preferences. The following exercise demonstrates that constructing distance-based types can be a helpful starting step for examining the parametric specification.

Varian (1990) suggests separating any parametric estimation into two parts. The first part should be independent of the parametric specification; it should test for rationalizability and quantify how close the observed behavior is to maximizing a well-behaved utility function.

The second part should focus on the prior functional form assumption and quantify how close the observed behavior is to maximizing the assumed utility specification. Following this line of thought, Halevy, Persitz, and Zrill (2018) suggest a method to recover the parameters of an assumed utility function such that a revealed preference-based loss function is minimized. They show that the loss function can be decomposed into two additive components. The first component measures the deviation of the observed behavior from rationality, and the second component measures the extent of misspecification in the functional form.

In this exercise, we use a money metric CCEI-based loss function. If the observed behavior is consistent (satisfies SARP), there exists a continuous, strictly concave, and strictly monotone utility function that rationalizes the observed demands. Consider a choice observation (p_t, q_t) and a utility function U. The observed demand implies that q_t is revealed preferred to all other feasible bundles q such that $p_t q \leq p_t q_t$. If U is the correct specification, it must be the case that $U(q_t) > U(q)$ for all $p_t q \leq p_t q_t$. However, if the utility function is misspecified, there exists a feasible bundle q that U ranks above q_t (that is, $U(q) > U(q_t)$). The CCEI-based loss function measures the minimal budget adjustment needed such that all those bundles that U ranks above q_t are no longer feasible. In this scenario, the loss quantifies the extent of misspecification caused by considering a specific utility function rather than all utility functions. By contrast, if the observed behavior violates SARP, the loss function can be decomposed into an internal inconsistency loss and a misspecification loss. The internal inconsistency loss, which is constant, represents the minimal loss required across all continuous, strictly concave, and strictly monotone utility functions, while the misspecification loss measures the additional loss that arises from restricting to the assumed utility function.

Following Halevy et al. (2018), we estimate a Constant Elasticity of Substitution (CES) functional form for each distance-based and age-based preference type. Table 5 shows the total loss (columns labeled "total") and its decomposition into internal inconsistency loss (columns labeled "internal") and misspecification loss (columns labeled "misspecification"). The total loss values suggest that, on average, the distance-based types require 16.05% of income to be wasted, while the age-based types require 18.23% of income to be wasted to be explained by a CES utility function. Although the distribution of total loss is similar across distance-based and age-based types, the decomposition into internal inconsistency and misspecification loss differs significantly. On average, misspecification loss is lower, while internal inconsistency loss is higher for age-based types compared to distance-based types. Intuitively, this is because age-based types are more likely to exhibit SARP violations than distance-based types. As explained above, when the data are not rationalizable, the method adjusts preference rankings in a minimal way to ensure that the observed choices becomes rationalizable. This adjustment removes some preference information, requiring any parametric utility function to be fitted to the remaining, smaller set of preference rankings. Consequently, when internal inconsistencies are larger, misspecification loss for any parametric class (not just the CES class) is smaller by construction. We interpret this as an inability to determine whether the CES class can describe the preferences of age-based types. By contrast, because distance-based types are constructed to have similar preferences, they exhibit fewer SARP violations. Thus, higher misspecification loss for these types can be interpreted as an indication that the CES class is not well-suited to represent group preferences.

Table 5: CCEI-based total, internal inconsistency, and misspecification loss from CES parametric estimation

	distance-based types			age-based types		
	total	internal	misspecification	total	internal	misspecification
mean	0.1605	0.0029	0.1576	0.1823	0.0636	0.1188
sd	0.0897	0.0033	0.0896	0.0940	0.0709	0.0695
min	0.0418	0.0000	0.0418	0.0207	0.0000	0.0207
p25	0.0909	0.0000	0.0900	0.1397	0.0185	0.0617
p50	0.1373	0.0010	0.1314	0.1668	0.0559	0.1066
p75	0.2145	0.0059	0.2115	0.2311	0.0820	0.1443
max	0.4613	0.0091	0.4612	0.4726	0.3312	0.2870

In the next step, we use the estimated CES utility functions to make out-of-sample demand predictions for each preference type. This exercise is the parametric counterpart to the illustration in Section 4.2. For types with misspecification loss below certain thresholds (5%, 10% and 15%), we conduct out-of-sample demand predictions by constructing counterfactual budgets using observations from 2015, 2017 and 2019. Comparing the point-identified predictions with the actual demands reflects the accuracy of the estimated utility functions.

Table 6 shows the Euclidean distance between the predicted and observed demands. The results clearly indicate that the predicted demands for the distance-based types are significantly more accurate than those for the age-based types. For types with misspecification loss below 5%, the average Euclidean distance between the predicted and observed demands is 0.1415. This is much smaller than the corresponding distance for the age-based types (0.4138). Considering slightly larger misspecification losses (10% and 15%) leads to similar conclusions. These out-of-sample demand predictions demonstrate that constructing types based on the revealed heterogeneity measure can be a valuable starting point for the parametric approach.

It should be noted that the purpose of this exercise is not to determine whether the assumed functional form (CES) provides a better or worse fit to the observed choices across

Table 6: Euclidean distance between predicted and observed demands

	distance-based types			age-based types		
	$\leq 5\%$	$\leq 10\%$	$\leq 15\%$	$\leq 5\%$	$\leq 10\%$	$\leq 15\%$
N	21	153	273	63	249	426
mean	0.1415	0.3939	0.3476	0.4138	0.4027	0.4612
std. dev.	0.1261	0.7667	0.6020	0.4039	0.4711	0.6648
min	0.0030	0.0003	0.0003	0.0036	0.0014	0.0014
p25	0.0525	0.0827	0.0901	0.1275	0.0939	0.1099
p50	0.1218	0.1683	0.1837	0.2540	0.2497	0.2676
p75	0.1912	0.3981	0.3996	0.5571	0.5380	0.5380
max	0.5272	6.4144	6.4144	1.6092	2.4960	7.0087

various types. Instead, the primary goal of this section is to illustrate how using distance-based types, which ensure greater internal consistency of choices, can enhance our ability to identify misspecifications in any functional form. Furthermore, this exercise demonstrates that, for a given level of misspecification, the utility functions estimated for the distance-based types result in more accurate out-of-sample demand predictions that those for the age-based types with similar levels of misspecification. Therefore, creating types based on the revealed heterogeneity measure can serve as a useful starting point for the parametric approach.

5 Conclusion

We presented a structural method to quantify the degree of preference heterogeneity between individuals. Using individual demand observations at finite combinations of prices and income, the method identifies preference heterogeneity as the Kemeny distance between individual preference rankings. The method is easily operationalized in practice using linear programming techniques and provides empirical results that offer insights into the degree of dissimilarity in individual preferences.

We demonstrated the potential of our method through an empirical application on the labor supply behavior of a sample of singles drawn from the PSID. The method obtains bounds on the heterogeneity measure, which can be valuable for applications in demand analysis. We further illustrated alternative uses of the recovered estimates through three applications of consumer demand analysis. Using the identified heterogeneity estimates, we constructed groups of similar preference types and showed that using these types as basic units of analysis can be highly beneficial in various demand applications. Specifically, we demonstrated that this approach can substantially strengthen the identifying power of nonparametric demand analysis and aid in diagnosing functional misspecifications in parametric estimations.

This paper made some simplifying choices which helped to keep the discussion focused. Weakening these assumptions could enrich the scope of the analysis. For example, the method was limited to comparing the preferences of two individuals. An interesting avenue for future research could involve extending the pairwise preference comparison approach to group comparison. Generalizing the method to allow for the structural identification of heterogeneity at the group level would significantly broaden the range of empirical questions that can be addressed. For instance, it could be used to examine differences in consumer behavior and their welfare effects both within and across countries and cultural groups (see, for example, the contributions of Dubois et al., 2014; Atkin, 2016 and Bertrand and Kamenica, 2023). Similarly, the method could be extended to alternative comparison criteria. For instance, Deb, Kitamura, Quah, and Stoye (2018) introduced the concept of revealed preferences over prices. Extending the current method to their framework would allow for the investigation of heterogeneity in price preferences.

References

A. Adams. Mutually consistent revealed preference demand predictions. *American Economic Journal: Microeconomics*, 2019.

A. Adams, R. Blundell, M. Browning, and I. Crawford. Prices versus preferences: taste change and revealed preference. Technical report, IFS Working Papers, 2015.

S. N. Afriat. The construction of utility functions from expenditure data. *International economic review*, 8(1):67–77, 1967.

S. N. Afriat. On a system of inequalities in demand analysis: an extension of the classical method. *International economic review*, pages 460–472, 1973.

A. Ambrus and K. Rozen. Rationalising choice with multi-self models. *The Economic Journal*, 125(585):1136–1156, 2014.

S. Andersen, G. W. Harrison, M. I. Lau, and E. E. Rutström. Preference heterogeneity in experiments: Comparing the field and laboratory. *Journal of Economic Behavior & Organization*, 73(2):209–224, 2010.

J. Apesteguia and M. A. Ballester. A measure of rationality and welfare. *Journal of Political Economy*, 123(6):1278–1310, 2015.

D. Atkin. The caloric costs of culture: Evidence from indian migrants. *American Economic Review*, 106(4):1144–81, 2016.

K. A. Baldiga and J. R. Green. Assent-maximizing social choice. *Social Choice and Welfare*, 40(2):439–460, 2013.

J. Banks, R. Blundell, and A. Lewbel. Quadratic engel curves and consumer demand. *Review of Economics and statistics*, 79(4):527–539, 1997.

G. S. Becker. Irrational behavior and economic theory. *Journal of political economy*, 70(1): 1–13, 1962.

M. Bertrand and E. Kamenica. Coming apart? cultural distances in the united states over time. *American Economic Journal: Applied Economics*, 15(4):100–141, 2023.

R. Blundell, M. Browning, and I. Crawford. Improving revealed preference bounds on demand responses. *International Economic Review*, 48(4):1227–1244, 2007.

R. Blundell, M. Browning, and I. Crawford. Best nonparametric bounds on demand responses. *Econometrica*, 76(6):1227–1262, 2008.

R. W. Blundell, M. Browning, and I. A. Crawford. Nonparametric engel curves and revealed preference. *Econometrica*, 71(1):205–240, 2003.

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

M. Browning and P.-A. Chiappori. Efficient intra-household allocations: A general characterization and empirical tests. *Econometrica*, pages 1241–1278, 1998.

F. A. Bugni, I. A. Canay, and X. Shi. Inference for subvectors and other functions of partially identified parameters in moment inequality models. *Quantitative Economics*, 8(1):1–38, 2017.

B. Can. Weighted distances between preferences. *Journal of Mathematical Economics*, 51: 109–115, 2014.

M. Castillo and M. Freer. Revealed differences. *Journal of Economic Behavior & Organization*, 145:202–217, 2018.

L. Cherchye and F. Vermeulen. Nonparametric analysis of household labor supply: goodness of fit and power of the unitary and the collective model. *The Review of Economics and Statistics*, 90(2):267–274, 2008.

L. Cherchye, B. De Rock, and F. Vermeulen. The collective model of household consumption: a nonparametric characterization. *Econometrica*, 75(2):553–574, 2007.

L. Cherchye, B. De Rock, and F. Vermeulen. The revealed preference approach to collective consumption behaviour: Testing and sharing rule recovery. *The Review of Economic Studies*, 78(1):176–198, 2011.

L. Cherchye, T. Demuynck, and B. De Rock. Revealed preference analysis for convex rationalizations on nonlinear budget sets. *Journal of economic theory*, 152:224–236, 2014.

L. Cherchye, B. De Rock, R. Griffith, M. O'Connell, K. Smith, and F. Vermeulen. A new year, a new you? heterogeneity and self-control in food purchases. 2017.

L. Cherchye, T. Demuynck, and B. De Rock. Transitivity of preferences: when does it matter? *Theoretical Economics*, 13(3):1043–1076, 2018.

L. Cherchye, T. Demuynck, B. De Rock, and K. Surana. Revealed preference analysis with normal goods: Application to cost-of-living indices. *American Economic Journal: Microeconomics*, 12(3):165–188, 2020.

S. Choi, R. Fisman, D. Gale, and S. Kariv. Consistency and heterogeneity of individual behavior under uncertainty. *American economic review*, 97(5):1921–1938, 2007.

S. Cosaert. What types are there? *Computational Economics*, 53(2):533–554, 2019.

S. Cosaert and K. Surana. A new interpretation and derivation of the swaps index. *Economics Letters*, 226:111109, 2023.

I. Crawford and K. Pendakur. How many types are there? *The Economic Journal*, 123 (567):77–95, 2012.

A. Dauphin, A.-R. El Lahga, B. Fortin, and G. Lacroix. Are children decision-makers within the household? *The Economic Journal*, 121(553):871–903, 2011.

M. Dean and D. Martin. Measuring rationality with the minimum cost of revealed preference violations. *Review of Economics and Statistics*, 98(3):524–534, 2016.

R. Deb, Y. Kitamura, J. K.-H. Quah, and J. Stoye. Revealed price preference: theory and empirical analysis. *arXiv preprint arXiv:1801.02702*, 2018.

M. Deza and T. Huang. Metrics on permutations, a survey. In *Journal of Combinatorics, Information and System Sciences*. Citeseer, 1998.

W. E. Diewert. Afriat and revealed preference theory. *The Review of Economic Studies*, 40 (3):419–425, 1973.

P. Dubois, R. Griffith, and A. Nevo. Do prices and attributes explain international differences in food purchases? *American Economic Review*, 104(3):832–67, 2014.

C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622. ACM, 2001.

F. Echenique, S. Lee, and M. Shum. The money pump as a measure of revealed preference violations. *Journal of Political Economy*, 119(6):1201–1223, 2011.

A. Falk, A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde. Global evidence on economic preferences. *The quarterly journal of economics*, 133(4):1645–1692, 2018.

Z. Fang, A. Santos, A. M. Shaikh, and A. Torgovitsky. Inference for large-scale linear systems with known coefficients. *Econometrica*, 91(1):299–327, 2023.

F. Farnoud, O. Milenkovic, and B. Touri. A novel Distance-Based approach to constrained rank aggregation. Dec. 2012.

J. Gross. On expenditure indices in revealed preference tests. *Journal of Political Economy*, 99(2):416–419, 1991.

J. Gross. Testing data for consistency with revealed preference. *The Review of Economics and Statistics*, pages 701–710, 1995.

V. Ha and P. Haddawy. Similarity of personal preferences: Theoretical foundations and empirical analysis. *Artificial Intelligence*, 146(2):149–173, 2003.

Y. Halevy, D. Persitz, and L. Zrill. Parametric recoverability of preferences. *Journal of Political Economy*, 126(4):1558–1593, 2018.

J. J. Heckman. Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of political Economy*, 109(4):673–748, 2001.

J. Heufer. Nonparametric comparative revealed risk aversion. *Journal of Economic Theory*, 153:569–616, 2014.

H. Kaido, F. Molinari, and J. Stoye. Confidence intervals for projections of partially identified parameters. *Econometrica*, 87(4):1397–1432, 2019.

G. Kalai, A. Rubinstein, and R. Spiegler. Rationalizing choice functions by multiple rationales. *Econometrica*, 70(6):2481–2488, 2002.

J. Kemeny and J. Snell. Mathematical models in the social sciences, chapter preference rankings: An axiomatic approach, 1962.

J. G. Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959.

P. Kidwell, G. Lebanon, and W. Cleveland. Visualizing incomplete and partially ranked data. *IEEE Transactions on visualization and computer graphics*, 14(6):1356–1363, 2008.

M. Kim, F. Raisali, F. Farnoud, and O. Milenkovic. Gene prioritization via weighted kendall rank aggregation. In *2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 184–187. IEEE, 2013.

C. Klamler. A distance measure for choice functions. *Social Choice and Welfare*, 30(3): 419–425, 2008.

M. Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 545–552. Association for Computational Linguistics, 2003.

M. Lapata. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484, 2006.

A. Lewbel and K. Pendakur. Tricks with hicks: The easi demand system. *American Economic Review*, 99(3):827–63, 2009.

A. Liang. Inference of preference heterogeneity from choice data. *Journal of Economic Theory*, 179:275–311, 2019.

N. N. Liu and Q. Yang. Eigenrank: a ranking-oriented approach to collaborative filtering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 83–90. ACM, 2008.

C. F. Manski. Identification of income–leisure preferences and evaluation of income tax policy. *Quantitative Economics*, 5(1):145–174, 2014.

R. L. Matzkin and M. K. Richter. Testing strictly concave rationality. *Journal of Economic Theory*, 53(2):287–303, 1991.

K. O. May. Intransitivity, utility, and the aggregation of preference patterns. *Econometrica: Journal of the Econometric Society*, pages 1–13, 1954.

R. B. Nelsen. *An introduction to copulas.* Springer Science & Business Media, 2007.

F. T. Nobibon, L. Cherchye, Y. Crama, T. Demuynck, B. De Rock, and F. C. Spieksma. Revealed preference tests of collectively rational consumption behavior: formulations and algorithms. *Operations Research*, 64(6):1197–1216, 2016.

H.-S. Park and C.-H. Jun. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.

M. Polisson, J. K.-H. Quah, and L. Renou. Revealed preferences over risk and uncertainty. *American Economic Review*, 110(6):1782–1820, 2020.

H. Rose. Consistency of preference: the two-commodity case. *The Review of Economic Studies*, 25(2):124–125, 1958.

P. J. Rousseeuw and L. Kaufman. Finding groups in data. *Hoboken: Wiley Online Library*, 1990.

P. A. Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17): 61–71, 1938.

P. A. Samuelson. Complementarity: An essay on the 40th anniversary of the hicks-allen revolution in demand theory. *Journal of Economic literature*, 12(4):1255–1289, 1974.

E. Schubert and P. J. Rousseeuw. Faster k-medoids clustering: Improving the pam, clara, and clarans algorithms. *arXiv preprint arXiv:1810.05691*, 2018.

H. R. Varian. The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, pages 945–973, 1982.

H. R. Varian. Goodness-of-fit in optimizing models. *Journal of Econometrics*, 46(1-2): 125–140, 1990.

H. R. Varian. *Goodness-of-fit for revealed preference tests.* Department of Economics, University of Michigan, 1991.

H.-M. Von Gaudecker, A. Van Soest, and E. Wengström. Heterogeneity in risky choice behavior in a broad population. *American Economic Review*, 101(2):664–694, 2011.

K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. *Icml*, 1:577–584, 2001.

S. Wang, J. Sun, B. J. Gao, and J. Ma. Adapting vector space model to ranking-based collaborative filtering. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1487–1491. ACM, 2012.

S. Wang, J. Sun, B. J. Gao, and J. Ma. Vsrank: A novel framework for ranking-based collaborative filtering. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):51, 2014.

M. L. Weitzman. On diversity. *The Quarterly Journal of Economics*, 107(2):363–405, 1992.

For online publication

Appendix A Proof of Proposition 1

Graph theoretical interpretation of SARP. Consider a finite, directed graph $G = (V, E)$, where V represents the set of vertices and E represents the set of edges. A directed graph G is called *strongly connected* if for every pair of distinct vertices i and $j \in V$, there exists a directed path from i to j and a directed path from j to i. A *subgraph* of graph G induced by $W \subseteq V$ is a graph $G(W) = (W, F)$ such that $F \subseteq E$ consists of all edges from E whose both end points are in W. A *strongly connected component* of graph G is a subgraph that is strongly connected and maximal with this property. That is, no additional edges or vertices can be included in this subgraph such that it still remains strongly connected. The set of strongly connected components of graph G forms a partition of its set of vertices.

Consider a dataset $\mathcal{S} = \{(p_t, q_t)\}_{t \in T}$. The preference graph $G = (T, E)$ associated with $\mathcal{S}$ is defined as follows. Each vertex corresponds to an observation in $\mathcal{S}$. Draw an edge between two distinct observations s, t if and only if $p_s q_s \geq p_s q_t$. That is, the edge $s \to t$ belongs to the preference graph if and only if $q_s R^D q_t$. Let E_{scc} be the set of edges that belong to the strongly connected components of this preference graph. Define $G_{scc} = (T, E_{scc})$ as the union of all strongly connected components of G.

We can now restate SARP in graph theoretic terminology. A dataset $\mathcal{S} = \{(p_t, q_t)\}_{t \in T}$ satisfies SARP if for all edges $s \to t \in E_{scc}$, it must be that $q_s = q_t$. The following proposition gives an alternate characterization of SARP in terms of a graph-theoretical interpretation.

Proposition 2 *A dataset $\mathcal{S} = \{(p_t, q_t)\}_{t \in T}$ satisfies SARP if and only if there exist $U_t \in \mathbb{R}$ for all $t \in T$ such that (i) $q_s = q_t$ if and only if $U_s = U_t$ and (ii) $q_s \neq q_t$ and $p_s q_s \geq p_s q_t$ implies $U_s > U_t$.*

Proof. The proof of Proposition 2 follows from the proof of Proposition 5 in Nobibon, Cherchye, Crama, Demuynck, De Rock, and Spieksma (2016). Suppose there exist numbers $U_t \in \mathbb{R}$ such that conditions (i) and (ii) in Proposition 2 are satisfied. Suppose $q_s \neq q_t$ and q_s is preferred over q_t. This means that there exist a sequence of observations $s, u, v, \cdots, z, t \in T$ such that $p_s q_s \geq p_s q_u, p_u q_u \geq p_u q_v, \cdots, p_z q_z \geq p_z q_t$. By condition (ii), this implies $U_s \geq U_u \geq \cdots \geq U_t$. If $q_t R^D q_s$ then $p_t q_t \geq p_t q_s$. If so, condition (ii) implies $U_t > U_s$ (a contradiction). Thus, it cannot be the case that $q_t R^D q_s$.

Conversely, suppose that $\mathcal{S}$ satisfies SARP. We need to show that there exist numbers $U_t \in \mathbb{R}$ such that conditions (i) and (ii) are satisfied. Consider the preference graph $G = (T, E)$ associated with $\mathcal{S}$. Let $G_1, G_2, \cdots, G_\alpha$ be the strongly connected components of G

indexed from 1 to α in a reverse topological order. That is, every arc $s \to t \in G$ goes from G^i to G^j such that $i \geq j$. For every observation $t \in T$, define $U_t = e_t/|T|$ where e_t is the index of the strongly connected component that contains q_t. These U_t's satisfy conditions $(i) - (ii)$. If $p_s q_s \geq p_s q_t$, then $s \to t \in E$. By construction, $U_s \geq U_t$. If $q_s = q_t$, then $s \to t \in E_{scc}$. Thus, by construction $U_s = U_t$. On the other hand, if $U_s = U_t$, then s and t are in a strongly connected component of G. By the alternate characterization of SARP given above, $q_s = q_t$. Otherwise, if $q_s \neq q_t$, then $s, t \notin E_{scc}$. Thus, by construction, $U_s > U_t$. ■

Proof of Proposition 1. Let $\{p_k\}_{k \in K}$ be a mutually feasible set of price vectors corresponding to a set of bundles $\{q_k\}_{k \in K}$ and a dataset $\mathcal{S} = \{(p_t, q_t)\}_{t \in T}$. We need to show that there exist $U_j \in [0, 1]$ and $R_{ij} \in \{0, 1\}$ for all $i, j \in K \cup T$ such that equations (1)-(6) are satisfied. By definition 4, $\{p_k\}_{k \in K}$ being a set of mutually feasible support prices implies that $p_k > 0$ and $p_k q_k = 1$ for all $k \in K$. Thus, constraints (1) and (2) are satisfied by definition. Further, $\{(p_j, q_j)\}_{j \in K \cup T}$ satisfies SARP. The following arguments will show that this translates to a non-empty domain defined by constraints (3)-(6). Let $\bar{\mathcal{S}} = \{(p_j, q_j)\}_{j \in K \cup T}$. Consider the preference graph $G = (V, E)$ associated with $\bar{\mathcal{S}}$ and its corresponding strongly connected components $G_1, G_2 \cdots, G_\alpha$ indexed in reverse topological order. For all $j \in K \cup T$, let e_j denote the index of the strongly connected component which consists q_j. For each observation q_j, define $U_j = e_j/|\bar{\mathcal{S}}|$ and define $R_{ij} = 1$ if $e_i \geq e_j$ and $R_{ij} = 0$ otherwise. It remains to be shown that U_j and R_{ij} defined this way satisfy constraints (3)-(6). First, note that because $1/|\bar{\mathcal{S}}| \leq U_t \leq 1$, we have $0 \leq |U_i - U_j| < 1$. For $i \neq j$, if $e_i \geq e_j$, by construction $R_{ij} = 1$ and $U_i \geq U_j$. If $e_i < e_j$, by construction, we have $R_{ij} = 0$ and $U_i < U_j$. Thus, constraints (3)-(4) are satisfied. If $p_i q_i < p_i q_j$, constraint (5) is trivially satisfied. If $p_i q_i \geq p_i q_j$, we have $i \to j \in E$. By construction, $e_i \geq e_j$ and $R_{ij} = 1$. Thus constraint (5) is also satisfied. Finally, for constraint (6), if $q_i = q_j$, the constraint is trivially satisfied. If $q_i \neq q_j$, then by proposition 2, $U_i \neq U_j$ and $e_i \neq e_j$. This implies that either $R_{ij} = 0$ or $R_{ji} = 0$. This ensures that condition (6) is also satisfied.

Conversely, suppose that $\{p_k\}_{k \in K}$ is a set of price vectors such that conditions (1)-(6) are satisfied. We need to show that $\{p_k\}_{k \in K}$ is a mutually feasible set of price vectors corresponding to the set of bundles $\{q_k\}_{k \in K}$ and the dataset $\mathcal{S} = \{(p_t, q_t)\}_{t \in T}$. The first two constraints ensure that $p_k > 0$ and $p_k q_k = 1$ for all $k \in K$. Next, showing $\bar{\mathcal{S}} = \{(p_j, q_j)\}_{j \in K \cup T}$ satisfies SARP is equivalent to showing that the U_j's corresponding to conditions (3)-(6) are such that conditions (i) and (ii) of Proposition 2 are also satisfied. If $q_i = q_j$, constraint (5) implies that $R_{ij} = R_{ji} = 1$. By constraint (4), we have $U_i = U_j$. Conversely, if $U_i = U_j$, constraint (3) implies that $R_{ji} = R_{ij} = 1$. By constraint (6), we have $q_i = q_j$. These ensure that condition (i) of Proposition 2 is satisfied. If $q_i \neq q_j$ and $p_i q_i \geq p_i q_j$, then constraint (5) implies that $R_{ij} = 1$. By condition (4), we have $U_j \leq U_i$. As $q_i \neq q_j$, we know that $U_j \neq U_i$. Thus, $U_j < U_i$. This ensures that condition (ii) of Proposition 2 is also satisfied. By Proposition 2, we can conclude that $\{(p_j, q_j)\}_{j \in K \cup T}$ satisfies SARP, and, thus, $\{p_k\}_{k \in K}$ is a set of mutually feasible price vectors. ■

Appendix B Support Prices and Mutual Feasibility

Support prices. Let $\mathcal{S}$ be a given dataset and consider two bundles, q' and q''. If $(p', q') \in \mathcal{S}$ and $p'q'' \leq p'q'$, then q' is revealed preferred to q''. Using the principle of revealed preference, we can conclude that q' is ranked higher in the preference ranking than q''. However, if $q' \notin \mathcal{S}$, can we use the observed demand behavior is $\mathcal{S}$ to infer anything about the preference relation between q' and q''? Varian (1982) argues that if every possible price vector p' at which q' could be chosen implies that q' is revealed preferred to q'', then we can conclude that the bundle q' is preferred over q''.[21] To formalize this reasoning, we start by defining the set of support prices for the quantity bundle q'.

Definition 5 *Given a dataset $\mathcal{S} = \{(p_t, q_t)\}_{t \in T}$ and a bundle q', the set of prices that support q' is defined as:*

$$S(q') = \left\{ p' : \begin{array}{l} p' > 0 \\ p'q' = 1 \\ \mathcal{S} \cup \{(p', q')\} \text{ satisfies SARP} \end{array} \right.$$

Intuitively, the set of support prices $S(q')$ contains all price vectors at which q' could be chosen, and the observation (p', q') would be rationalizable with the observed data $\mathcal{S}$. (Note that requiring $p'q' = 1$ is simply a convenient normalization.)

Revealed worse and revealed preferred bundles. Using support prices, we can define the set of bundles revealed worse than and revealed preferred over q'. We say that a bundle q is revealed worse than q' if, for *all* price vectors that support q', q' is revealed preferred over q. Specifically, we define the set of revealed worse-than bundles as,

$$RW(q') = \{q : \text{for all } p' \in S(q'),\ q'Rq\}.$$

For any bundle q in the set $RW(q')$, it can be safely concluded that $q' \succ q$.

[21] In terms of a utility function, this is equivalent to the condition that if every well-defined utility function u that rationalizes $\mathcal{S}$ also implies that $u(q') > u(q'')$, then it is safe to conclude that $q' \succ q''$.

Similarly, a bundle q is revealed preferred over q' if *all* prices that support q imply that qRq'. The set of bundles revealed preferred over q' can be formalized as,

$$RP(q') = \{q : \text{ for all } p \in S(q), qRq'\}.$$

For any bundle q in the set $RP(q')$, we can conclude that $q \succ q'$.

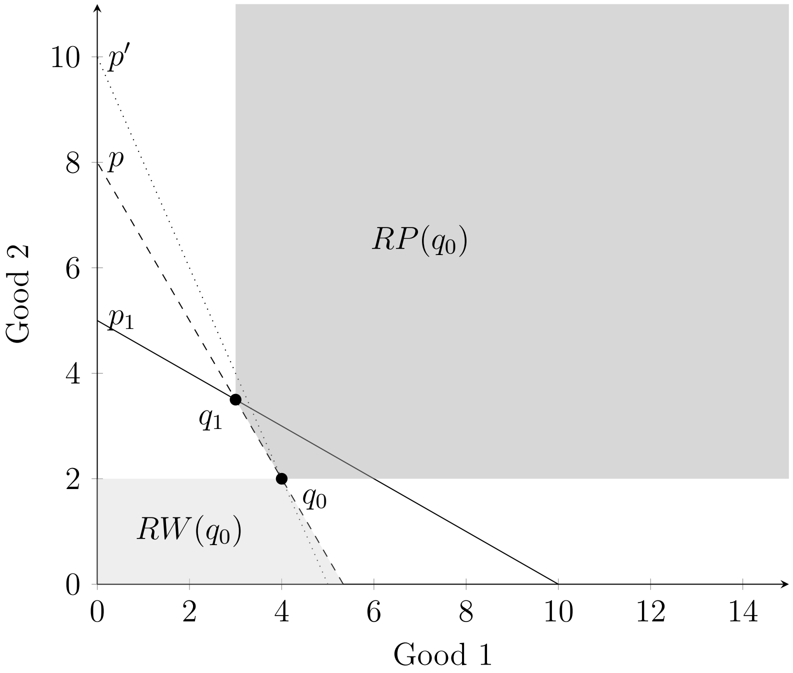
Example 2 The following example explains, in a two-goods setting, the concepts of support prices, revealed worse and revealed preferred sets. Consider a dataset with a single choice observation $\mathcal{S} = \{(p_1, q_1)\}$ and a bundle q_0, as shown in Figure 5. The bundle q_0 is not in the observed dataset, but we want to reliably recover the bundles which would be preferred over q_0 and those that would be preferred less than q_0. Since q_0 is not in the observed data, we do not know the prices (say, p_0) at which q_0 could have been demanded. However, the observed data impose some restrictions on the feasibility of p_0. Specifically, p_0 cannot be strictly steeper than the dashed budget line p (passing through q_0 and q_1) because any price vector stepper than p would imply a SARP violation with the individual's observed data. For example, (p', q_0) and (p_1, q_1) are SARP-inconsistent choices. Conversely, any non-negative price vector that passes through q_0 and is less steep than p could be a supporting price vector for q_0.

It is easy to verify that all supporting prices for q_0 imply that q_0 is preferred over all the bundles in the light-shaded region $RW(q_0)$. For any $q \in RW(q_0)$, we can infer that $q_0 \succ q$. Using similar arguments, we can conclude that all bundles in the dark-shaded region $RP(q_0)$ are preferred over q_0. For any $q \in RP(q_0)$, we can infer that $q \succ q_0$.

Mutual feasibility of support prices. Individually recovered support prices may not be mutually feasible. This can be illustrated with a simple example. Consider the situation depicted in Figure 6, where the observed data $\mathcal{S} = \{(p_1, q_1)\}$ and two bundles $\{q_0, q_0'\}$ are not in the observed data. To recover sets of bundles that are revealed preferred over and revealed worse than q_0 and q_0', one can use the observed data $\mathcal{S}$ to first identify sets of support prices $S(q_0)$ and $S(q_0')$. Then, using these identified support prices, one can recover the sets of bundles revealed worse than and revealed preferred over q_0 and q_0'.

However, this approach is problematic because such separate recoveries of support prices do not guarantee that all combinations of support prices for q_0 and q_0' would be mutually rationalizable. For instance, applying the definition of support prices separately to each bundle could mean that p_0 is a feasible price vector for the bundle q_0, and p_0' is a feasible price vector for the bundle q_0'. However, as shown in Figure 6, these price vectors are not mutually feasible. Specifically, these particular choices of price vectors— $((p_0', q_0')$ and

Figure 5: Revealed worse and revealed preferred set



(p_0, q_0))— imply a SARP violation. To ensure mutual feasibility, we define mutually feasible support price vectors (see Definition 4).[22]

Appendix C Practical Implementation

C.1 Set Identification

We use the rationalizability conditions to set identify the distance between preference rankings. This boils down to defining upper and lower bounds on the number of preference conflicts over a given set Q, given individual demand data $\mathcal{S}^a$ and $\mathcal{S}^b$. To obtain an upper bound, we solve the following optimization problem.

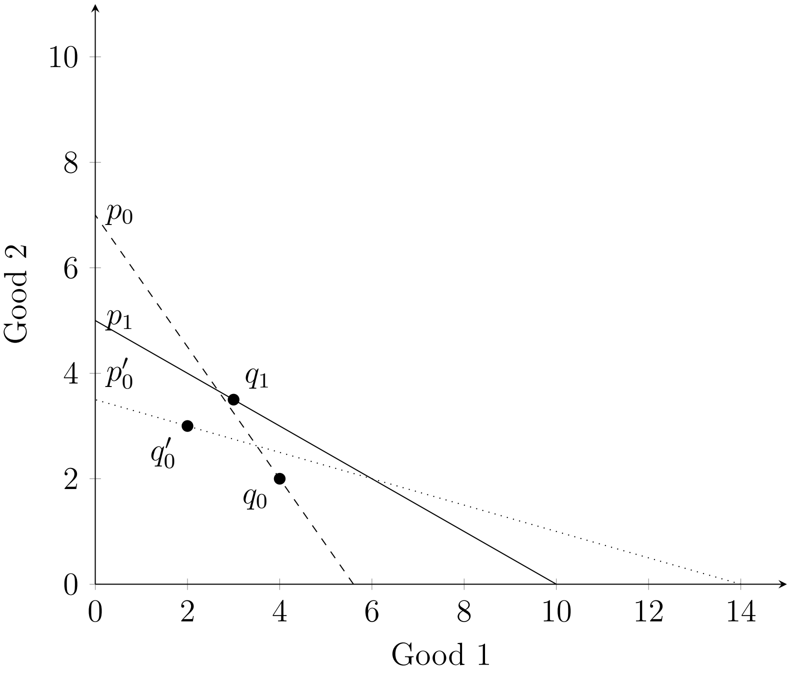
$$\text{maximize} \sum_{\{i,j \ |i>j,\ q_i,q_j \in Q\}} D_{ij} \ \text{ such that}$$

$$p_i^m > 0 \text{ for all } q_i \in Q^m \text{ and } m \in \{a,b\}$$

$$p_i^m q_i = 1 \text{ for all } q_i \in Q^m \text{ and } m \in \{a,b\}$$

[22] Adams (2019) shows a similar result in the context of demand predictions. Specifically, the author demonstrates that separate predictions of rational demand responses over different counterfactual budgets do not guarantee mutual feasibility (or joint rationality) of the predicted demands.

Figure 6: Mutually feasibility of support prices



$$U_i^m - U_j^m < R_{ij}^m \text{ for all } q_i, q_j \in Q^m \cup \mathcal{S}^m \text{ and } m \in \{a, b\}$$

$$R_{ij}^m - 1 \le U_i^m - U_j^m \text{ for all } q_i, q_j \in Q^m \cup \mathcal{S}^m \text{ and } m \in \{a, b\}$$

$$p_i^m q_i - p_i^m q_j < M_i R_{ij}^m \text{ for all } q_i, q_j \in Q^m \cup \mathcal{S}^m \text{ and } m \in \{a, b\}$$

$$q_{in} - q_{jn} \le q_{in}(2 - R_{ij}^m - R_{ji}^m) \text{ for all } q_i, q_j \in Q^m \cup \mathcal{S}^m, n \in \{1, \cdots, N\} \text{ and } m \in \{a, b\}$$

$$R_{ji}^a + R_{ij}^b \le 1 + D_{ij} \text{ for all } i > j \text{ and } q_i, q_j \in Q$$

$$R_{ji}^b + R_{ij}^a \le 1 + D_{ij} \text{ for all } i > j \text{ and } q_i, q_j \in Q$$

$$R_{ji}^a + R_{ji}^b \le 2 - D_{ij} \text{ for all } i > j \text{ and } q_i, q_j \in Q$$

$$R_{ij}^a + R_{ij}^b \le 2 - D_{st} \text{ for all } i > j \text{ and } q_i, q_j \in Q$$

$$D_{ij} \in \{0, 1\},\ R_{ij}^m \in \{0, 1\},\ U_i^m \in [0, 1] \text{ for all } q_i, q_j \in Q^m \cup \mathcal{S}^m \text{ and } m \in \{a, b\}$$

Dividing the optimal value of the objective function in the above optimization problem by the number of pairwise bundles, we obtain an upper bound on the distance-based preference heterogeneity measure. Similarly, to obtain a lower bound, we minimize the objective function subject to the same linear constraints.

C.2 K-medoids Constrained Clustering

As the estimated Kemeny distances are non-Euclidean distances with unknown underlying variables generating them, we cannot use a k-means clustering algorithm to assign individuals to types. To adapt to this empirical setting, we use a modified k-means algorithm known as the k-medoids algorithm, which is suitable for any arbitrary distance measure (see Rousseeuw and Kaufman, 1990; Park and Jun, 2009). The difference between the two algorithms is that in a k-means algorithm, the centroids (mean coordinates of the cluster objects) are used for assigning objects to clusters, while in a k-medoids algorithm, the data points (medoids) are used for the assignment. A medoid of a cluster is a data point with the smallest sum of distances to all other data points in the cluster. Formally, a medoid of a cluster C is defined as follows.

$$\text{medoid}(C) := \arg\min_{x_i \in C} \sum_{x_j \in C} d(x_i, x_j).$$

We adapt the algorithms of Wagstaff, Cardie, Rogers, and Schrödl (2001) and Park and Jun (2009) to create a k-medoids clustering algorithm which allows for pairwise incompatibility (or cannot-link) constraints. This is described in Algorithm 1. There are three remarks on the algorithm. First, the algorithm is considered "converged" if there is no further change in the assignment of data points to clusters. Second, the use of incompatibility (cannot-link) constraints implies that the assignment of data points to clusters may depend on the order in which data points appear in the algorithm. If the initial medoids are chosen randomly, it may reach a point where no valid clustering is possible. To avoid such a situation, we take initial assignments from the solutions of the corresponding graph coloring algorithm (see MatGraph package). The medoids of these valid clusters are then used as initial medoids in the algorithm. Third, it has been shown that the effectiveness of such algorithms in finding the optimal clustering strongly relies on a good initialization (see Schubert and Rousseeuw, 2018). To account for this, we run Algorithm 1 1000 times and choose the cluster with the lowest total deviation (TD) defined as,

$$TD := \sum_{k} \sum_{x_i, x_j \in c_k} d(x_i, x_j).$$

C.3 Identification of Cost of Living Indices

Bounds on the cost of living index $c_{t,r}$ can be obtained by identifying the money metric utility function $\mu(p_t; p_r, y_r)$. In this section, we revisit the procedure outlined by Varian (1982) to find approximations to the lower and upper bounds of the money metric utility function and show that these problems can be represented as MILP optimization problems.

Algorithm 1: constrained k-medoids

```
Input: Distance $D$, Incompatibility $I$, number of clusters $k$
choose $k$ initial medoids $c_1, c_2, \cdots, c_k$
while not converged do
    foreach data point $d_i$ do
        assign it to the closest medoid $C_j$ such that violate-constraints($d_i, C_j, I$) = FALSE
        if no such cluster exists then
            return {}
        foreach cluster $C_k$ do
            update its medoid
Output: $\{C_1, C_2, \cdots, C_k\}$
Function violate-constraints($d_i, C_j, I$):
    violate = FALSE
    foreach data point $d_j$ in cluster $C_j$ do
        if $I(d_i, d_j) = 1$ then
            violate = TRUE
    return violate
```

Upper bound. The money metric utility function is defined as,

$$\mu(p_t; p_r, y_r) \equiv e(p_t, v(p_r, y_r)),$$
$$\mu(p_t; p_r, y_r) \equiv \inf p_t x \text{ such that } x \in \{x : u(x) \geq v(p_r, y_r)\}.$$

Consider a dataset $\mathcal{S} = \{(p_k, q_k)\}_{k \in K}$. An upper bound to $\mu(p_t; p_r, y_r)$ can be obtained by computing the indirect overcompensation function:

$$\mu^+(p_t; p_r, y_r) = \inf p_t x \text{ such that } x \in XRP(p_r, y_r)$$

where $XRP(p_r, y_r)$ is the set of all bundles revealed preferred to the budget (p_r, y_r). That is, the set of all bundles x such that, at any prices p at which x could be bought, x is revealed preferred to all bundles $x_r \in S(p_r, y_r)$. Here, $S(p_r, y_r)$ represents the set of bundles that support (p_r, y_r). Formally,

$$XRP(p_r, y_r) = \{x : \text{ for all } p \in S(x) \text{ and all } x_r \in S(p_r, y_r), xRx_r\},$$
$$S(p_r, y_r) = \{x_r : \{(p_r, x_r)\} \cup \mathcal{S} \text{ satisfies SARP and } p_r x_r = y_r\}.$$

If the demands in observation t and reference regime r are unknown, we can approximate

μ^+ as the maximum expenditure over all bundles that could be demanded in the reference regime (p_r, y_r). That is:

$$a\mu^+(p_t; p_r, y_r) = \max p_t x \text{ such that } x \in S(p_r, y_r).$$

The following MILP formulation defines an optimization problem that we can solve to obtain an approximation to the overcompensation function $a\mu^+(p_t; p_r, y_r)$.

$$\begin{gathered}
\text{maximize } p_t q_r \text{ such that} \\
q_r \geq 0,\ p_r q_r = y_r \\
U_i - U_j < R_{ij} \text{ for all } i, j \in \{r\} \cup K \\
R_{ij} - 1 \leq U_i - U_j \text{ for all } i, j \in \{r\} \cup K \\
q_{in} - q_{jn} \leq q_{i,n}(2 - R_{ij} - R_{ji}) \text{ for all } t, s \in \{r\} \cup K \text{ and } n \in 1, \cdots, N \\
p_i q_i - p_i q_j < M_i R_{ij} \text{ for all } i, j \in \{r\} \cup K \\
R_{ij} \in \{0, 1\}, U_i \in [0, 1] \text{ for all } i, j \in \{r\} \cup K
\end{gathered}$$

Lower bound. A lower bound to $\mu(p_t; p_r, y_r)$ can be obtained by computing the indirect undercompensation function:

$$\mu^-(p_t; p_r, y_r) = \inf p_t x \text{ such that } x \in NXRW(p_r, y_r).$$

where $NXRW(p_r, y_r)$ is the set of bundles not revealed worse than budget (p_r, y_r). A bundle is revealed worse than a budget (p_r, y_r) if, for all bundles $x_r \in S(p_r, y_r)$, x_r is preferred over x. That is,

$$\begin{aligned}
XRW(p_r, y_r) &= \{x : \text{ for all } x_r \in S(p_r, y_r), x_r R x\}, \\
&= \{x : p_r x_k \leq y_r \text{ for some } x_k R x\}.
\end{aligned}$$

The following MILP formulation gives an optimization problem that can be solved to obtain an approximation to the undercompensation function $\mu^-(p_t; p_r, y_r)$.

$$\begin{gathered}
\text{minimize } p_t q_v \text{ such that} \\
q_v, q_r \geq 0,\ p_r q_r = y_r \\
U_i - U_j < R_{ij} \text{ for all } i, j \in \{v, r\} \cup K \\
R_{ij} - 1 \leq U_i - U_j \text{ for all } i, j \in \{v, r\} \cup K \\
q_{in} - q_{jn} \leq q_{in}(2 - R_{ij} - R_{ji}) \text{ for all } i, j \in \{v, r\} \cup K \text{ and } n \in 1, \cdots, N
\end{gathered}$$

$$p_i q_i - p_i q_j < M_i R_{ij} \text{ for all } i, j \in \{v, r\} \cup K$$
$$p_i q_i - p_i q_v < M_i R_{iv} \text{ for all } i \in \{r\} \cup K$$
$$R_{rv} = 0,\ R_{ij} \in \{0, 1\} \text{ for all } i, j \in \{r\} \cup K$$

In the above optimization problem, q_v is constrained to be a bundle which is not revealed worse than the reference regime bundle by constraining $R_{rv} = 0$. Since the binary variables R_{iv} for $i \in \{r\} \cup K$ capture whether (p_i, y_i) is revealed preferred to the unknown bundle q_v, transitivity ensures that q_v is chosen such that it is not revealed worse than bundle q_r.

Cost efficiency indices. In some instances, the dataset may not be exactly consistent with SARP. When the data do not satisfy the exact conditions, small deviations from SARP can be permitted by extending Afriat's notion of Critical Cost Efficiency Index (CCEI; Afriat, 1973) to our setting. In particular, the MILP formulations can be easily modified by introducing a parameter $e \in [0, 1]$. In the identification of the overcompensation function, we would replace constraints 5 and 6 with

$$eq_{in} - q_{jn} \leq q_{in}(2 - R_{ij} - R_{ji}) \text{ for all } i, j \in \{v, r\} \cup K \text{ and } n \in 1, \cdots, N,$$
$$ep_i q_i - p_i q_j < M_i R_{ij} \text{ for all } i, j \in \{v, r\} \cup K.$$

where $0 \leq e \leq 1$ is a pre-specified efficiency level. Similarly, when identifying the undercompensation function, we would replace constraints 5, 6 and 7 with

$$eq_{in} - q_{jn} \leq q_{in}(2 - R_{ij} - R_{ji}) \text{ for all } i, j \in \{v, r\} \cup K \text{ and } n \in 1, \cdots, N,$$
$$ep_i q_i - p_i q_j < M_i R_{ij} \text{ for all } i, j \in \{v, r\} \cup K,$$
$$ep_i q_i - p_i q_v < M_i R_{iv} \text{ for all } i \in \{r\} \cup K.$$

Appendix D Additional Empirical Results

D.1 Data

The empirical setting considers two consumption categories: leisure and aggregate Hicksian consumption. The Hicksian consumption consists of six broad categories of expenditures: food, housing, transportation, education, childcare, and healthcare.[23] Food expenditures

[23] Since 2005, PSID reports household expenditures on two additional consumption category: clothing and recreation. To remain consistent with the composition of Hicksian consumption across the years, we do not

record spending on food at home, delivered food, and food eaten away from home. Housing expenditures record spending on mortgage and loan payments, rent, property taxes, insurance, utilities, cable TV, telephone, internet charges, home repairs, and home furnishings. Transportation expenditures record spending on vehicle loans, leases, and down payments, insurance, other vehicle expenses (e.g., repairs and maintenance), gasoline, parking, carpooling, bus fares, train fares, taxis, and other transportation costs. Education expenditures record total school-related expenses. Childcare expenditures represent total spending on childcare, and healthcare expenditures include spending on hospitals, nursing homes, doctors, prescription drugs, and insurance.

Table 7 provides a summary of the selected sample. Years of education represent the actual grade of school completed, Hicksian consumption is the average daily dollar expenditures on the six consumption categories described above and leisure hours is the average total daily hours spent on leisure.

D.2 Individual Rationalizability

We use the CCEI to measure the severity of SARP violations. Column 2 of Table 8 summarizes the CCEI indices for the individuals in the selected sample. The first row shows that a large majority of individuals (65.73%) are consistent with the strict conditions (i.e., their CCEI $= 1$). The second row shows that 80.28% are very close to satisfying SARP (with CCEI ≥ 0.99). The average CCEI (0.9902) is also very high, which indicates that even if the observed individuals are not exactly consistent, they are quite close to satisfying the rationalizability conditions. Some individuals, however, are quite far from satisfying the exact conditions. For example, the minimum CCEI of 0.8054 indicates that this individual needs about 20% wasted income to be consistent with rationality. Overall, these results provide strong support for the rationality assumption applied to the selected sample. In most cases, we need only a small adjustment in the data to obtain consistency with rationality.

While the CCEI indices suggest that the data are mostly rationalizable, it can be argued that a favorable goodness-of-fit is due to the low restrictiveness of the test. Thus, it is equally important to examine the extent of empirical "bite" present in the rationalizability conditions. This can be quantified by computing the power of the test. Bronars (1987) suggests measuring power as the probability of detecting irrational behavior. Bronars' power uses irrational behavior as a benchmark to compare the observed behavior. A high power measure, complemented with a favorable goodness-of-fit, provides a convincing support for the rationality hypothesis.

use these additional expenditure categories.

Table 7: Summary statistics

	mean	sd	min	max
age in year 1999	37.36	8.52	19.00	63.00
years of education	13.51	2.23	0.00	17.00
gender (0 = female/1 = male)	0.19	0.39	0.00	1.00
height (in inches)	5.73	2.99	1.00	11.00
race = White	0.42	0.49	0.00	1.00
race = Black	0.57	0.50	0.00	1.00
race = Asian	0.01	0.10	0.00	1.00
race = Other	0.00	0.07	0.00	1.00
Hicksian consumption in 1999	50.60	29.28	6.23	225.94
Hicksian consumption in 2001	54.71	34.18	6.27	317.48
Hicksian consumption in 2003	61.20	39.54	7.50	335.26
Hicksian consumption in 2005	74.72	44.49	0.49	449.10
Hicksian consumption in 2007	84.36	83.49	13.60	1158.11
Hicksian consumption in 2009	82.00	46.48	3.45	393.05
Hicksian consumption in 2011	90.56	117.46	11.67	1669.46
Hicksian consumption in 2013	86.16	41.86	19.73	256.81
Hicksian consumption in 2015	85.66	59.83	6.36	724.37
Hicksian consumption in 2017	89.27	45.16	13.53	319.40
Hicksian consumption in 2019	95.93	71.56	21.13	912.33
hourly wage in 1999	13.75	9.95	0.71	56.07
hourly wage in 2001	16.12	14.85	2.45	176.40
hourly wage in 2003	16.60	12.48	2.38	133.33
hourly wage in 2005	18.42	18.51	1.59	192.31
hourly wage in 2007	19.01	16.09	2.04	183.33
hourly wage in 2009	20.76	12.73	0.09	72.83
hourly wage in 2011	20.94	16.51	3.17	156.25
hourly wage in 2013	21.20	14.81	0.94	138.89
hourly wage in 2015	21.10	12.97	1.44	70.76
hourly wage in 2017	22.31	15.25	1.66	90.09
hourly wage in 2019	25.58	28.82	3.21	326.54
leisure hours in 1999	10.21	1.34	4.57	15.71
leisure hours in 2001	10.20	1.24	6.00	14.57
leisure hours in 2003	9.73	1.54	3.00	13.57
leisure hours in 2005	9.95	1.46	1.00	13.86
leisure hours in 2007	9.66	1.98	0.86	14.86
leisure hours in 2009	10.15	1.54	3.14	14.71
leisure hours in 2011	10.02	1.63	1.71	15.29
leisure hours in 2013	10.05	1.85	0.29	15.43
leisure hours in 2015	10.20	1.48	1.71	14.86
leisure hours in 2017	10.32	1.62	1.71	14.57
leisure hours in 2019	10.32	1.72	4.57	15.71

Table 8: Critical cost efficiency index and Bronars power

	CCEI	Bronars power
CCEI = 1	140 (65.73%)	
CCEI $\geq$ 0.99	171 (80.28%)	
mean	0.9902	0.6891
sd	0.0245	0.1533
min	0.8054	0.0220
p25	0.9946	0.6160
p50	1.0000	0.7170
p75	1.0000	0.7950
max	1.0000	0.9380

Following Cherchye and Vermeulen (2008), we quantify the power measure at the individual level. Becker (1962) suggests simulating irrational behavior by choosing the quantity bundles randomly from a uniform distribution across all bundles in the budget hyperplane. We simulate 1000 random responses corresponding to each individual's budgets and check the consistency of each of these simulated random data. Bronars' power is defined as the number of times rationality is rejected in the random samples divided by the number of simulated samples. This approximates the probability of detecting an irrational behavior. Column 3 in Table 8 shows the proportion of random samples that reject rationality. The estimates show that the power measures are reasonably high, with an average of about 69%. Overall, these results provide empirical validation for the use of rationality as the identifying assumption for preference heterogeneity.

D.3 Revealed and Observed Preference Heterogeneity

The recovered estimates in Table 2 reveal quite some heterogeneity in the patterns of preference dissimilarities. We investigate this further by relating the distance in preferences to differences in observed characteristics. This lends additional insight into what is driving the heterogeneity in individual preferences. Specifically, we estimate two regression models using ordinary least squares (OLS), where the lower and upper bounds are used as the dependent variables. Table 9 shows the model estimates. Columns 2 and 3 report the estimates when the dependent variables are the estimated lower and upper bounds, respectively. The results show that a few observable characteristics correlate significantly with the identified lower and upper bound of the preference heterogeneity measure. Particularly important ones are age and gender. A larger age difference is associated with higher preference dissimilarity, while individuals of the same gender tend to have lower dissimilarities.

Table 9: Preference dissimilarity and observed chacteristics

	lower bound	upper bound
Δ birth year	0.0002*	-0.0003
	(0.0001)	(0.0002)
Δ completed level of education	0.0002	0.0004
	(0.0003)	(0.0007)
same gender	-0.0036*	-0.0078*
	(0.0016)	(0.0031)
same race	-0.0001	-0.0010
	(0.0009)	(0.0020)
born in same state	-0.0035	-0.0079
	(0.0030)	(0.0060)
grew up in same state	0.0015	-0.0030
	(0.0028)	(0.0059)
born in same region	-0.0002	0.0044
	(0.0011)	(0.0026)
N	14535	14535
R-squared	0.3464	0.5241

Notes: *p< 0.1, **p<0.05, ***p< 0.01. Standard errors are given in parentheses. Controls include Bronars power and number of distinct pairs over which distance was computed.

D.4 Preference Types

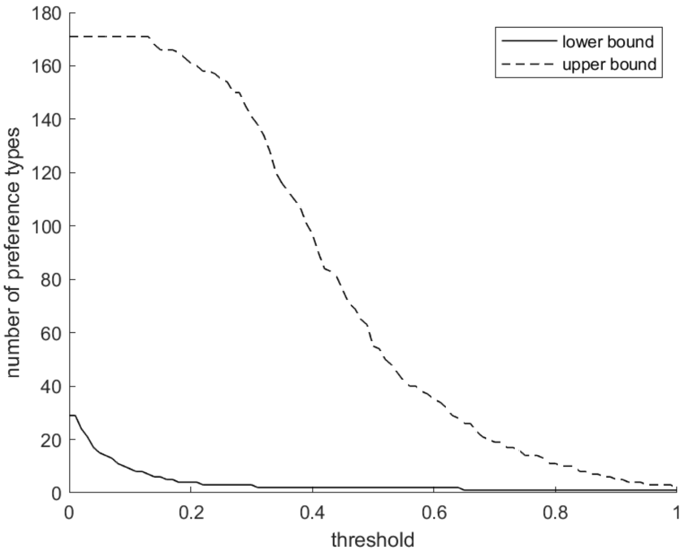
Number of preference types. In the main text, we required two individuals to belong to different preference types if the lower bound of their distance was strictly positive. However, there may be other ways to define such incompatibilities among individuals. For instance, in the spirit of Varian (1990), we may allow for "almost" similar preference rankings to belong to the same preference type. Then, two individuals would be incompatible only when their distance is above some threshold value (see, e.g., Castillo and Freer, 2018; Liang, 2019). Unfortunately, there is no theoretical foundation for what constitutes a reasonable threshold value and the answer is likely to depend on the problem at hand.[24] In this context, it is

[24] In a closely analogous context of goodness-of-fit measures, Varian (1990) suggests "the magic number of significance tests, 5%, as a reasonable choice".

interesting to explore how the required number of preference types will change with different threshold values and the use of lower or upper bound estimates of preference heterogeneity. This would reflect the sensitivity of the number of required preference types to alternative definitions of incompatibility constraints.

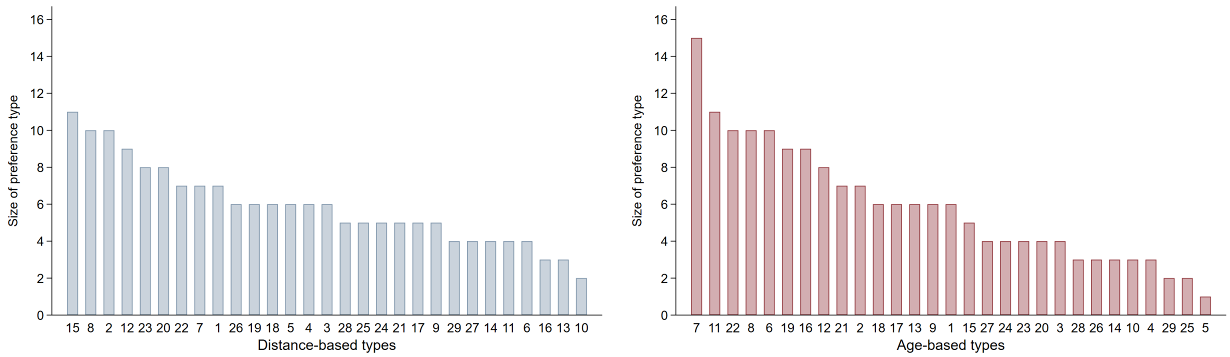
Figure 7 plots the required minimum number of types (approximated by a greedy algorithm) versus the threshold value. The dashed line plots the minimum number of types required when we use the upper bounds to define incompatible pairs of individuals, while the solid line shows the number of types required when using the lower bounds. Unsurprisingly, for a given threshold value, the number of required preference types is higher when incompatibilities are based on the upper bounds than when they are based on the lower bounds. Further, as expected, the number of required preference types decreases with the threshold value. That is, the sample can be clustered in a smaller number of preference types if we allow larger heterogeneity within types. For example, if we use the lower bounds and choose a high enough threshold value (above 0.65), all individuals would belong to the same preference type. This threshold corresponds to the maximum value of the lower bound estimates summarized in Table 2. In the other extreme case with a threshold value of zero, the number of preference types required by the upper bound estimates is 171 and that by the lower bound estimates is 29.

Figure 7: Number of preference types by incompatibility threshold



Sizes of preference types. Figure 8 shows the sizes of the distance-based types (graph (a)) and age-based types (graph(b)) with types ordered from largest to smallest. For the distance-based types, the largest type consists of 11 individuals (6.43% of the sample), and the smallest type consists of two individuals (1.17% of the sample). As for the age-based preference types, the largest type consists of 15 individuals (8.77% of the sample), and the smallest type consists of a single individual (0.58% of the sample).

Figure 8: Size of preference types



(a) distance-based types

(b) age-based types

D.5 Out-of-Sample Demand Prediction

To assess the correctness of the predicted demand intervals, we compute the distance between the observed demands and the bound of the predicted intervals that is closest to the observed demand. If the observed demand lies within the predicted interval, the distance is defined as zero. Columns 2 and 3 of Table 10 summarize the results. The first row shows the number (percent) of predictions that contain the observed demands. Considering both the number of correct predictions and the distance to the closest interval, it appears that age-based predictions perform better than distance-based predictions. The age-based predictions include the observed demands for about 93% of the counterfactual budgets, while the distance-based predictions are correct for about 77% of the counterfactual budgets.

However, looking only at the number of correct predictions can be misleading. It is essential to consider the informativeness of the predicted demand intervals. We quantify this by computing the difference between the upper and lower bounds of the predicted demand intervals. Columns 4 and 5 show the tightness of the predicted intervals for distance and age-based predictions, respectively. The results clearly indicate that the favorable predictive performance of age-based types is mainly driven by the large width of the predicted intervals. While the average width of predicted intervals for the age-based types is 0.6892, it is significantly smaller for distance-based types, at 0.3979.

Table 10: Distance to and width of predicted intervals

	distance to predicted interval		width of predicted interval	
	distance-based	age-based	distance-based	age-based
correct predictions	394 (76.80%)	476 (92.79%)		
mean	0.0135	0.0030	0.3979	0.6892
sd	0.0357	0.0162	0.2931	0.2571
min	0.0000	0.0000	0.0069	0.0633
p25	0.0000	0.0000	0.1357	0.4591
p50	0.0000	0.0000	0.3240	0.7563
p75	0.0000	0.0000	0.6554	0.9157
max	0.3243	0.1906	1.0000	1.0000

Appendix E Robustness Checks

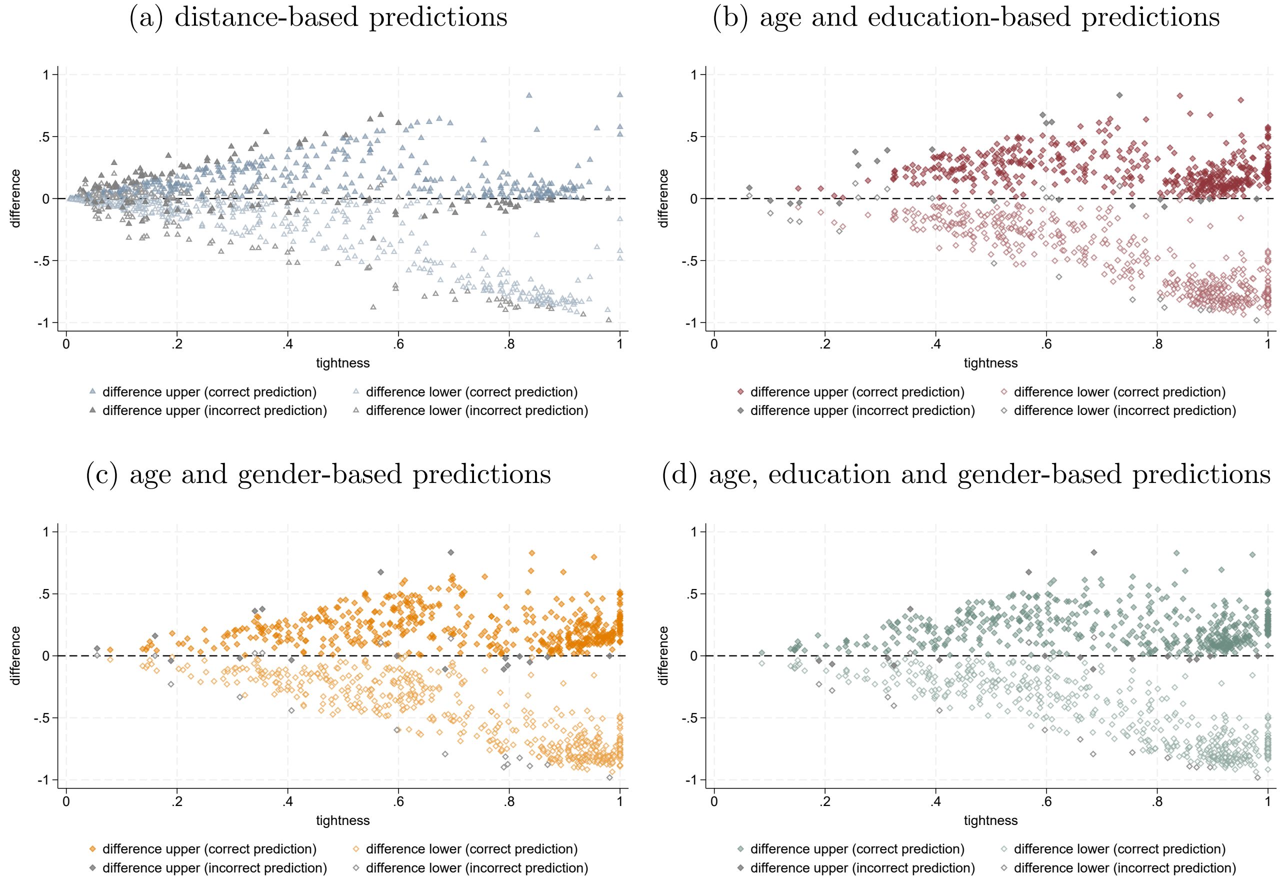
E.1 Age, Education and Gender-based Clustering

In our main empirical analysis, we used a single dimension (age) to define types based on observable characteristics. As a robustness check, we consider additional dimensions to define types based on observable heterogeneity. Specifically, we use age, gender and/or years of education to group "similar" individuals. The predictive performances of the out-of-sample demand predictions for these types are shown in Figure 9, which can be compared with Figure 3 in the main text. Additionally, we compare the bounds on cost-of-living indices identified from these types with those obtained from distance-based types. The results are presented in Table 11 and Figure 10. Comfortingly, our qualitative conclusions remain very similar to those in the main text.

E.2 Distance-based Types with Relaxed Incompatibility Constraints

In the main analysis, we constructed distance-based types by requiring that two individuals belong to different types if the lower bound of their preference heterogeneity measure is strictly positive. As a robustness check, we consider an alternative approach that allows for small positive values of lower-bound estimates when defining incompatible individuals. Specifically, we require two individuals to belong to different types only if their lower-bound distance exceeds 1%. Figure 11 shows the predictive performance of out-of-sample predictions based on these relaxed distance-based types. We find that even with this less restrictive construction, distance-based types perform significantly better than age-based types. Next, we compute bounds on cost-of-living indices and compare those obtained from distance-based

Figure 9: Out-of-sample demand predictions; alternative clustering



(a) distance-based predictions

(b) age and education-based predictions

(c) age and gender-based predictions

(d) age, education and gender-based predictions

Figure 10: CDF of distance-based and alternative criteria-based bounds

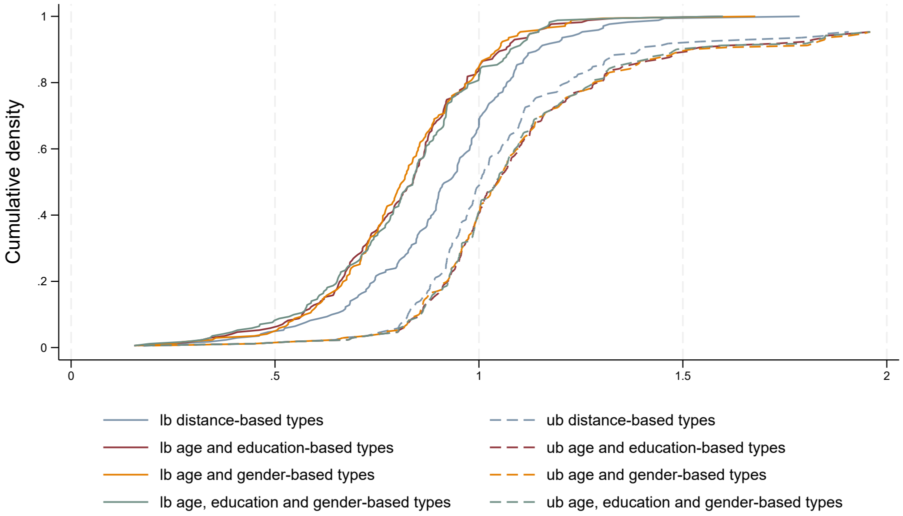


Table 11: Bounds on $c_{2011,2007}$; alternative clustering

Panel A: distance-based types			
	lower	upper	Δ_d
mean	0.912	1.172	0.260
sd	0.224	0.978	0.935
min	0.164	0.175	0.001
p25	0.800	0.919	0.026
p50	0.924	1.003	0.077
p75	1.040	1.140	0.202
max	1.786	12.530	11.277

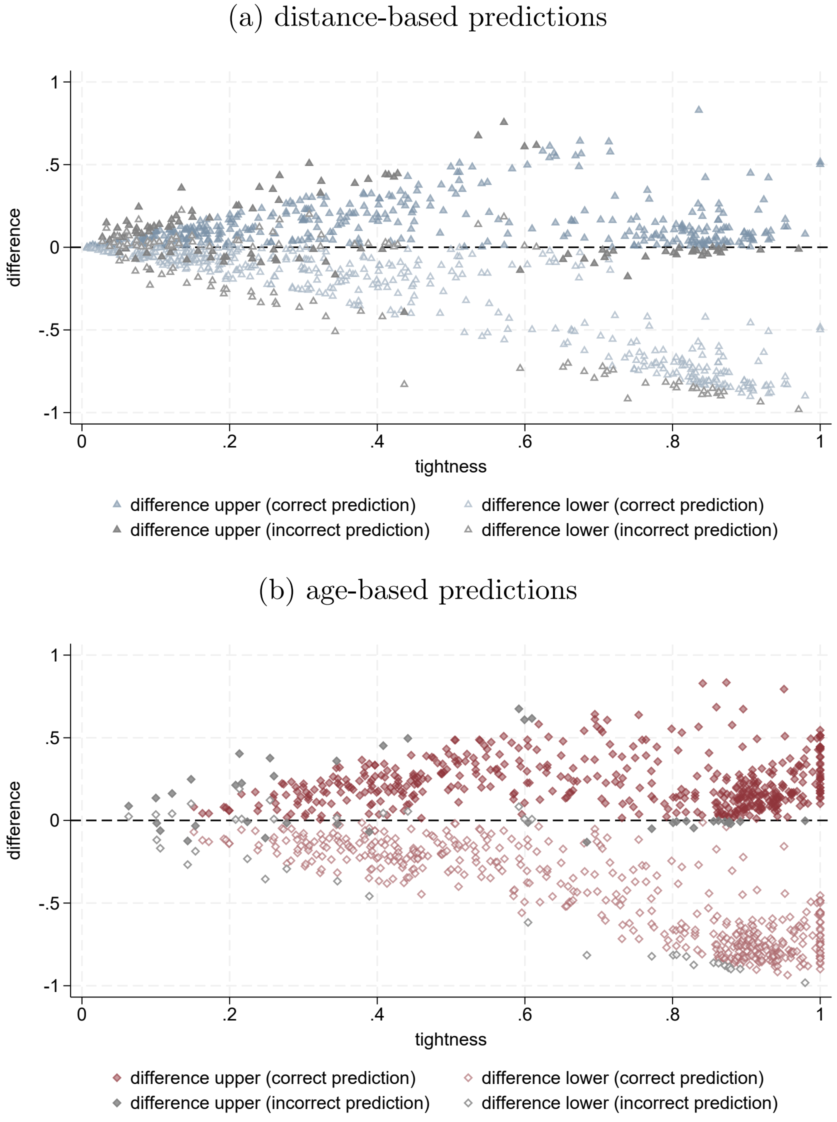
Panel B: age and education-based types				
	lower	upper	Δ_{ae}	$\frac{\Delta_{ae}-\Delta_d}{\Delta_{ae}}$
mean	0.815	1.220	0.405	0.518
sd	0.208	0.980	0.975	0.438
min	0.155	0.178	0.018	-2.747
p25	0.683	0.949	0.126	0.301
p50	0.838	1.046	0.217	0.619
p75	0.934	1.219	0.389	0.817
max	1.598	12.530	11.814	0.994

Panel C: age and gender-based types				
	lower	upper	Δ_{ag}	$\frac{\Delta_{ag}-\Delta_d}{\Delta_{ag}}$
mean	0.814	1.216	0.402	0.535
sd	0.202	0.977	0.964	0.346
min	0.157	0.199	0.012	-0.357
p25	0.708	0.941	0.140	0.284
p50	0.815	1.049	0.238	0.594
p75	0.934	1.213	0.357	0.837
max	1.678	12.530	11.808	0.996

Panel D: age, education and gender-based types				
	lower	upper	Δ_{aeg}	$\frac{\Delta_{aeg}-\Delta_d}{\Delta_{aeg}}$
mean	0.817	1.216	0.399	0.514
sd	0.217	0.981	0.974	0.414
min	0.155	0.178	0.013	-2.383
p25	0.690	0.945	0.129	0.284
p50	0.838	1.040	0.216	0.600
p75	0.934	1.213	0.383	0.814
max	1.598	12.530	11.869	0.994

types with those derived from age-based types. The results, shown in Table 12 and Figure 12, confirm that distance-based types yield tighter bounds than age-based types. Overall, our main qualitative conclusions remain robust.

Figure 11: Out-of-sample demand predictions; distance-based types with relaxed incompatibility constraints

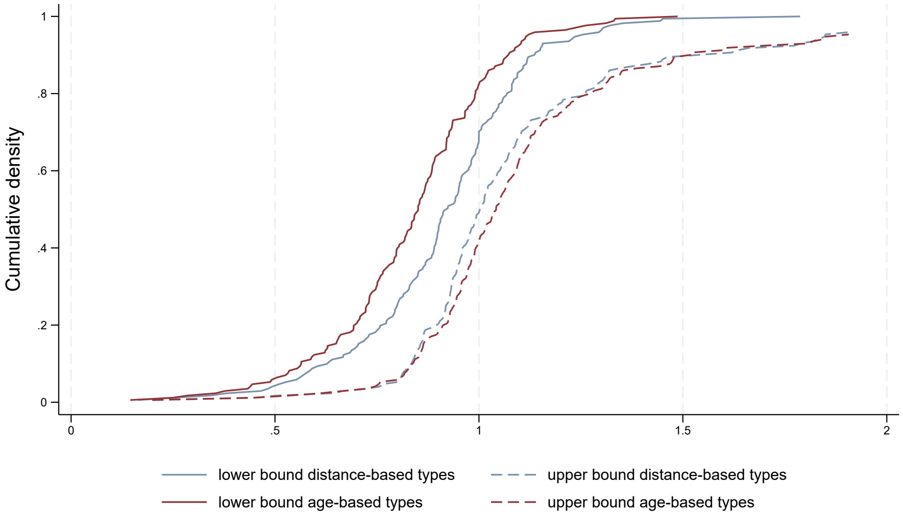

(a) distance-based predictions

(b) age-based predictions

Table 12: Bounds on $c_{2011,2007}$; distance-based types with relaxed incompatibility constraints

	distance-based types			age-based types			
	lower	upper	Δ_d	lower	upper	Δ_a	$\frac{\Delta_a - \Delta_d}{\Delta_a}$
mean	0.913	1.180	0.267	0.837	1.202	0.365	0.425
sd	0.220	0.976	0.948	0.206	0.970	0.960	0.502
min	0.164	0.199	0.001	0.145	0.199	0.025	-1.459
p25	0.800	0.919	0.026	0.725	0.939	0.119	0.122
p50	0.924	1.003	0.071	0.851	1.042	0.183	0.496
p75	1.042	1.171	0.207	0.966	1.207	0.321	0.837
max	1.788	12.530	11.450	1.487	12.530	11.808	0.997

Figure 12: CDF of distance-based and age-based bounds; distance-based types with relaxed incompatibility constraints



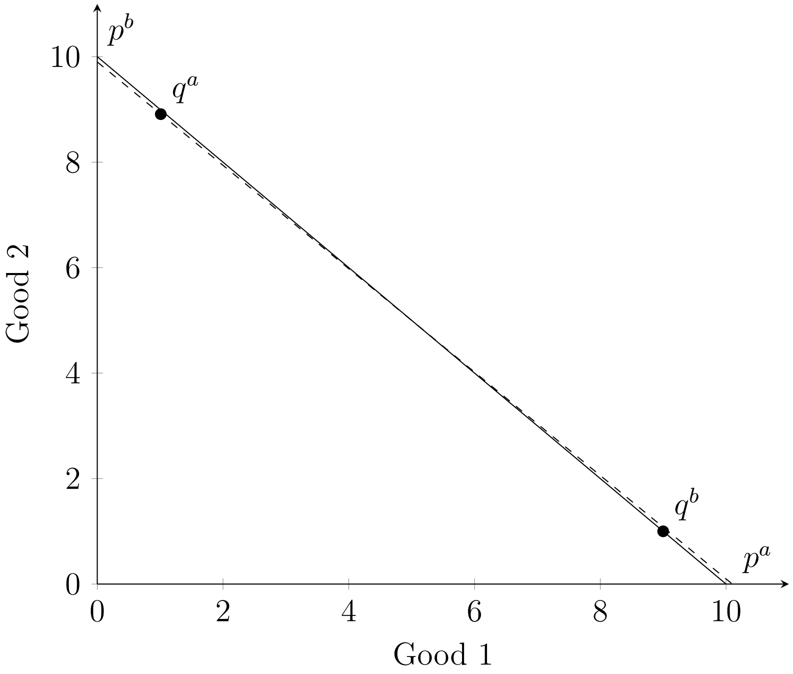
Appendix F Goodness-of-fit Based Heterogeneity Measure

F.1 Expenditure-based Index to Measure Preference Heterogeneity

Expenditure-based goodness-of-fit indices are appealing due to their simplicity and ease of computation. However, they are unsuitable for capturing differences in underlying preferences. Specifically, if individuals face similar prices and have similar levels of income, expenditure-based measures will indicate a high degree of preference similarity, regardless of actual differences in preferences. To illustrate this point, we revisit the example suggested by Gross (1991).

Consider two individuals, a and b, with Cobb-Douglas preferences $u^a(q_1, q_2) = q_1^{0.1} q_2^{0.9}$ and $u^b(q_1, q_2) = q_1^{0.9} q_2^{0.1}$. Clearly, their preferences are very different, so the individuals will make very different choices. Suppose they have the same income of 1000 and face prices $(p_1^a, p_2^a) = (99, 101)$ and $(p_1^b, p_2^b) = (100, 100)$, respectively. It is easy to compute that their optimal choices will be $(q_1^a, q_2^a) = (1.01, 8.91)$ and $(q_1^b, q_2^b) = (9, 1)$. These choices are shown in Figure 13.

Figure 13: Expenditure-based heterogeneity measure



It is evident from Figure 13 that the observed demands of the two individuals are not rationalizable by a common preference, as they violate SARP. Suppose we use the CCEI index to measure how different their preferences are. For the observations described above, we find that the CCEI index is 0.992, which leads us to incorrectly conclude that the individuals may have similar preferences. That is, a high CCEI index indicates that it is possible to construct a utility function that almost rationalizes both individuals' observed behavior. However, it does not reveal how much their underlying preferences may actually differ, as the true preferences could still be substantially different. This limitation is addressed by our distance-based approach, which (a) does not rely on an expenditure-based method and (b) derives lower and upper bounds. These bounds not only show whether a common utility function can rationalize both individuals' behavior (lower bound), but also quantify the potential degree of misspecification if their true utility functions differ significantly.

F.2 Comparison with Castillo and Freer (2018)

Below, we compare our proposed distance-based heterogeneity measure with a goodness-of-fit-based heterogeneity measure proposed by Castillo and Freer (2018) (hereafter referred to as CF).

Given an individual's dataset $\mathcal{S}$, CF first define distance from rationality $\rho(\mathcal{S})$, which takes a value between zero and one, such that $\rho(\mathcal{S}) = 0$ if and only if the dataset is rationalizable. Next, they define two individuals to be *revealed different* if their joint data is further from rationality than each of their data taken separately. More specifically,

$$RD_\epsilon(\mathcal{S}^a, \mathcal{S}^b) = \begin{cases} 1, & \text{if } \rho(\mathcal{S}^a \cup \mathcal{S}^b) - \max\{\rho(\mathcal{S}^a), \rho(\mathcal{S}^b)\} > \epsilon \\ 0, & \text{otherwise} \end{cases}$$

where ϵ is a precision parameter to account for measurement issues in estimating the true distance from rationality.

There are several important differences between our proposed distance-based heterogeneity measure and CF's revealed different measure. First, the RD measure uses an auxiliary "distance from rationality" function to measure heterogeneity, whereas our proposed measure is based on a more fundamental concept: comparing conflicts in preference rankings to measure heterogeneity. Second, the RD measure produces binary outcomes; two individuals are either revealed different or they are not. In contrast, our proposed measure is cardinal, capturing the degree of difference between two preference rankings.[25] Third, the

[25] Because the revealed different measure produces binary outcomes, it does not satisfy all the properties required of a distance function. Specifically, it satisfies two of the three properties of a distance function: (1) $RD_\epsilon(\mathcal{S}^a, \mathcal{S}^b) \geq 0$ and (2) $RD_\epsilon(\mathcal{S}^a, \mathcal{S}^b) = RD_\epsilon(\mathcal{S}^b, \mathcal{S}^a)$. However, it does not satisfy the triangle inequality and is therefore not a metric. In contrast, the Kemeny distance satisfies all three properties and qualifies as

RD measure implicitly favors preference similarity by defining RD as one only when the change in the distance from rationality is sufficiently large. Conversely, our set identification method determines both the lowest and highest possible values of distance between preferences. Thus, our lower and upper bounds reflect all the preference ranking information that can be extracted from the observed data. Finally, the RD measure is crucially dependent on two factors: ρ, which measures distance from rationality, and ϵ, the precision parameter. In their application, the authors use the Critical Cost Efficiency Index (CCEI), the Money Pump Index (MPI), and the Houtman-Maks Index (HMI) to measure distance from rationality. They demonstrate that results may vary significantly depending on the chosen index. Similarly, the precision parameter ϵ, plays a crucial role in shaping results. Higher values of ϵ reveal substantially less heterogeneity.[26] On the other hand, our proposed measure is dependent on the set of bundles Q over which preference rankings are compared. In the paper, we used the observed bundles to define the set over which two individuals' preference rankings are compared. In this context, we note that existing measures are also typically dependent on the observed choices. For example, the distance from rationality measure used in the definition of the revealed different measure also depends on the set of observed choices.

In the following, we empirically compare our estimated distance-based heterogeneity measure with the RD measure. For ρ, we use the CCEI index. Specifically, given a dataset $\mathcal{S}$, define $\rho(\mathcal{S}) = 1 - CCEI(\mathcal{S})$. For ϵ, we consider two values: $\epsilon = 0$ (labeled RD_0) and $\epsilon =$ the 95th percentile of the empirical distribution of the change in the distance from rationality (labeled RD_{95CI}). Table 13 shows the estimated lower and upper bounds for individuals that are and are not revealed different from each other. Columns 2-5 show the results for the RD_0 measure and columns 6-9 show the results for the RD_{95CI} measure.

Several notable patterns emerge. First, we find that the estimated lower and upper bounds are higher among pairs that are revealed different from each other than among pairs not revealed to be different. This is unavoidably so, as both measures aim to capture the same underlying concept. Second, the difference is more pronounced for the lower bound than for the upper bound. This is expected, as the RD measure implicitly favors preference similarity. Third, there is substantial heterogeneity in the estimated bounds among pairs revealed different to each other. For instance, among pairs with $RD_0 = 1$, the estimated lower bounds range from 0.01 to 0.65, and the upper bound range from 0.18 to 1.00. Finally, the two measures are not perfectly correlated. There are individuals who are not revealed

a metric.

[26] For example, in their sample of 1182 individuals, setting ϵ equal to the 95th percentile of the empirical distribution of the change in the distance from rationality, the number of preference types is 35 with the CCEI, 17 with the MPI and 5 with the HMI. Setting ϵ equal to zero, increases the number of preference types to 131 with CCEI, 116 with the MPI, and 248 with the HMI.

to be different from each other according to either the RD_0 or RD_{95CI} measures but exhibit large distance-based differences. For example, one pair of individuals has an estimated lower bound as high as 0.14, yet the pair is not revealed to be different from each other.

Table 13: Distance-based preference heterogeneity by revealed different measure

	RD_0				RD_{95CI}			
	$RD_0 = 0$		$RD_0 = 1$		$RD_{95CI} = 0$		$RD_{95CI} = 1$	
	lower	upper	lower	upper	lower	upper	lower	upper
mean	0.0006	0.6420	0.0802	0.6422	0.0309	0.6389	0.1579	0.7022
sd	0.0051	0.1629	0.0713	0.1544	0.0538	0.1581	0.0912	0.1641
min	0.0000	0.1333	0.0110	0.1800	0.0000	0.1333	0.0137	0.1800
p25	0.0000	0.5263	0.0299	0.5278	0.0000	0.5246	0.0889	0.5890
p50	0.0000	0.6458	0.0556	0.6413	0.0000	0.6406	0.1429	0.7188
p75	0.0000	0.7632	0.1064	0.7561	0.0426	0.7547	0.2083	0.8333
max	0.1429	1.0000	0.6500	1.0000	0.6500	1.0000	0.5696	1.0000