Predictive Power in Behavioral Welfare Economics

Elias Bouacida, Daniel Martin

Predictive Power in Behavioral Welfare Economics

Elias Bouacida* and Daniel Martin†

April 3, 2017

Abstract

When choices appear inconsistent due to behavioral biases, there is a theoretical debate about whether or not it is necessary to impose the structure of a model in order to provide meaningful welfare guidance based on such choices. To address this question empirically, we evaluate the predictive power of the "model-free" approach to welfare analysis proposed by Bernheim and Rangel (2009) using two standard data sets. For most of the demand functions that could *possibly* be observed in this data, their approach does not offer clear welfare guidance. However, we find that for most of the demand functions that are *actually* observed, their approach can be used to make tight predictions, even when these demands exhibit inconsistencies. For the experimental choices of Manzini and Mariotti (2010), we show that the welfare guidance provided by such predictions are largely consistent with delay aversion, while the guidance provided by revealed preferences is ambiguous.

JEL Codes: I30, C91, D12

Keywords: Welfare economics, behavioral economics, revealed preferences, limited attention, experimental economics, scanner data

*Université Paris 1 Panthéon Sorbonne – Paris School of Economics, 48 Boulevard Jourdan, Paris 75014, elias.bouacida@psemail.eu.

†Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, d-martin@kellogg.northwestern.edu.

1 Introduction

The welfare benefits of an economic policy are difficult to ascertain if individuals do not make consistent choices about the goods impacted by that policy. For instance, should healthy foods be subsidized even though consumers sometimes choose unhealthy foods over healthy ones? Put more formally, it is difficult to determine whether a policy will maximize utility if choices do not appear to correspond to a well-behaved utility function.

This is a real issue in practice. In addition to the large number of preference reversals identified in the behavioral economics literature, several recent papers have demonstrated choice inconsistencies in standard data sets – both experimental (Choi, Fisman, Gale, and Kariv 2007; Choi, Kariv, Müller, and Silverman 2014) and observational (Blundell, Browning, and Crawford 2003; Dean and Martin 2016). These findings suggest that individuals cannot be modeled *as if* they maximize a standard, stable utility function in a wide range of choice settings.

However, it is still normatively appealing to retain choice as the basis for welfare assessments. One choice-based solution is to find a model of choice procedures, decision-making errors, or behavioral biases that explains observed choices and to use that model to conduct welfare analysis.[1] An alternative choice-based solution is to generate a (possibly incomplete) relation from choices without imposing much *ad hoc* model structure and to use that relation to conduct welfare analysis (Bernheim and Rangel 2009; Chambers and Hayashi 2012; Apesteguia and Ballester 2015; Nishimura 2015). For the latter, a leading proposal is a welfare relation P^* introduced by Bernheim and Rangel (2009) that only allows comparisons when observed choices are unambiguous about which alternative is preferred, which is called the Strict Unambiguous Choice Relation (Strict UCR).[2]

Given the nature of this divide, a debate has emerged as to how much model structure is necessary to provide meaningful welfare guidance from inconsistent choices (Bernheim 2009; Rubinstein and Salant 2012; Manzini and Mariotti 2014; Bernheim 2016). We provide empirical evidence for this theoretical debate by determining the predictive power of the "model-free" approach of Bernheim and Rangel (2009) for two standard data sets.[3] We calculate predictive power for both the demands that are actually observed in the data and the demands that could have been observed given the choice sets that individuals faced.[4]

[1] There are many such examples from the decision theory and behavioral economics literatures, including Caplin, Dean, and Martin (2011), Manzini and Mariotti (2012), and Rubinstein and Salant (2012).

[2] Masatlioglu, Nakajima, and Ozbay (2012) provide an example of where Strict UCR and their model provide different welfare guidance, so Strict UCR is not completely free of model structure.

[3] This is technically related to the work of Manzini and Mariotti (2010), Beatty and Crawford (2011), Andreoni, Gillen, and Harbaugh (2013), Dean and Martin (2016), and Boccardi (2017), who use predictive power for a different purpose: to evaluate how demanding revealed preference tests are to pass.

[4] To generate a wide range of these "hypothetical" demands, we draw random sequences of choices from

Predictive power is a useful measure for the clarity of welfare guidance for choice-based approaches because what a relation predicts will be chosen corresponds to what is welfare optimal for that relation. For instance, if Strict UCR indicates that just one option should be selected from a choice set, then it has both maximal predictive power and offers the clearest possible welfare guidance. However, if Strict UCR says that any option could be selected from a choice set, then it has minimal predictive power and offers no welfare guidance.

As an example, imagine the choices of $\{x\}$ from $\{x, y, z\}$, $\{x\}$ from $\{x, a\}$, and $\{a\}$ from $\{x, y, a\}$. From these choices, Strict UCR says that xP^*y, xP^*z, and aP^*y. For the choice set $\{x, y\}$, Strict UCR predicts a single option will be chosen and offers clear welfare guidance – it says that the *individual welfare optimum* for that choice set is x. On the other hand, for other choice sets, such as $\{x, a\}$, Strict UCR predicts that anything would be chosen and offers no welfare guidance.

To calculate the predictive power of Strict UCR for a given data set, we generate prospective "virtual" choice sets by randomly selecting subsets of alternatives. For a virtual choice set, we determine the predictions of Strict UCR by determining which options in that choice set are consistent with the relation generated from the observed choices in the data set.[5]

We test Strict UCR's predictive power for two types of data: from the lab, a set of choices from an incentivized experiment; and from the field, a set of scanned grocery purchases. The former is composed of choices from menus of payment plans for 102 students, which comes from an experiment carried out by Manzini and Mariotti (2010).[6] The latter is composed of choices from budget sets for 1,183 single-person households over 10 years, which comes from Nielsen's National Consumer Panel (NCP) – formerly known as the Homescan Consumer Panel.[7]

We selected these data sets for four reasons. First, both are representative of widely used types of data in the economic literature. Second, in both data sets individuals make inconsistent choices: for the experimental data, 52.94% of individuals make choices that generate revealed preference cycles, and for the consumption data, 100% of individuals exhibit revealed preference cycles. Third, both have unique features that make them rich enough to effectively test predictive power: the experimental data contains choices from all subsets of alternatives, and the consumption data contains a large number of observations per individual. Fourth, they are quite different from each other in terms of subject pools, choice settings, and choice alternatives.

---

the choice sets that individuals faced.

[5] Specifically, we eliminate those alternatives that are dominated by other alternatives in the choice set according to the relation.

[6] We are very grateful to the authors for providing this data to us.

[7] Data from The Nielsen Company (US), LLC and marketing databases provided by the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business.

Consistent with what has been hypothesized in the literature, in both settings we find that Strict UCR does not offer clear welfare guidance for most hypothetical demand functions. However, for most observed demands, we find that Strict UCR is acyclic and far from coarse. Instead, it is a complete or nearly complete welfare relation for most individuals. For the experimental data, Strict UCR is complete for 47.06% of individuals and is within one comparison of being complete for 74.51% of individuals. For the consumption data, Strict UCR is never complete, but on average it contains 99.12% of the comparisons in a complete and acyclic preference relation over chosen bundles. For those individuals with an acyclic relation, Strict UCR contains on average 99.23% of the comparisons in a complete and acyclic preference relation.

In addition, we find that Strict UCR has a high level of predictive power for observed demands. In both choice settings, a complete and acyclic relation predicts that a single alternative would be chosen from any virtual choice set, so the predictive power benchmark is 1 predicted alternative. For individuals in the experimental data, Strict UCR predicts an average of 1.17 alternatives would be chosen from a virtual choice set. For 84.36% of these choice sets, Strict UCR predicts just a single alternative would be chosen. For individuals in the consumption data with an acyclic Strict UCR, the relation predicts an average of 1.03 alternatives would be chosen from a virtual choice set, with 96.74% of choice sets having just a single predicted alternative.

Looking at these precise predictions for the experimental data, the unambiguous welfare optimal payment plans – according to Strict UCR – are those with earlier payments. This pattern is obscured when looking at revealed preferences alone. In fact, ambiguities exist when comparing all types of plans, which suggests that behavioral biases are leading to non-systematic choice inconsistencies.

In addition, we compare Strict UCR to a proposal by Nishimura (2015) called the "transitive core", which also takes a conservative model-free approach to constructing welfare relations. For both data sets, we find that the transitive core of revealed preferences performs similarly to Strict UCR in terms of predictive power. It also makes similar predictions for which option is welfare optimal for subjects in the experiment.

Finally, we also compare Strict UCR to the relation produced by the revealed attention approach of Masatlioglu, Nakajima, and Ozbay (2012). For the experimental data, we find that when Strict UCR is close to fully incomplete, the preference relation implied by the revealed attention approach is close to fully complete. However, the revealed attention relation is frequently cyclic, making it difficult to extract clear welfare guidance from the relation. Also, we find no cases where these approaches provide different welfare guidance, though Masatlioglu, Nakajima, and Ozbay (2012) provide a hypothetical example where Strict UCR and their preference relation offer different welfare guidance.

To the best of our knowledge, this paper presents the first non-parametric empirical applications of Strict UCR, as well as the first empirical applications of both the transitive core and the revealed attention approach.[8] In addition, we introduce predictive power as a tool for assessing behavioral welfare relations. From this, we help to provide an answer to the question of how much model structure is necessary to provide meaningful welfare guidance. For the standard choice data sets we consider, it appears that one can give clear welfare guidance without imposing many assumptions – on the form of utility, on the nature of the behavioral biases, or on which choice sets to consider.

In section 2, we briefly introduce Strict UCR and describe alternative welfare relations. In section 3, we describe the two data sets. In section 4, we provide results on the completeness and predictions of Strict UCR for both data sets. In that section, we also provide robustness checks. We conclude with a brief discussion in section 5.

2 Strict UCR and Other Welfare Relations

It is well-known that a set of choices can be rationalized by a utility function if and only if the preferences revealed by those choices contain no cycles (excluding cycles of indifference). To express this technically, we need to introduce some notation.[9] First, x is directly revealed preferred to y (denoted xR^0y) if x is chosen when y is available, and this relationship is strict if xR^0y and not yR^0x for the same choice (denoted xP^0y). Second, x is revealed preferred to y (denoted xRy) if there exists $x, x_1, x_2, \ldots, x_n, y$ such that $xR^0x_1, x_1R^0x_2, \ldots, x_nR^0y$. Finally, R contains a cycle if xRy and yP^0x.

If a set of choices generates no revealed preferences cycles (a condition called acyclicity), the revealed preferences that correspond to those choices are suitable for conducting welfare analysis. However, if the choices generate revealed preference cycles, the decision maker can no longer be modeled as a utility maximizer, which calls for a different approach to welfare analysis.

2.1 Frames and Welfare

Bernheim and Rangel (2009) and Salant and Rubinstein (2008) separately proposed the idea of using frames, as defined by Tversky and Kahneman (1981), to make welfare assessments in light of inconsistencies in choice. In both cases, the key element is, in addition to the

[8] As discussed in section 2, there are existing parametric empirical applications of Strict UCR.

[9] For an introduction to revealed preferences, see Varian (2006) and Adams and Crawford (2015).

choice itself, an ancillary condition or frame present when the choice is made.[10] They assume that while a frame may impact choice, it does not affect the alternatives themselves or the welfare derived from them.[11]

Bernheim and Rangel (2009) allow the econometrician to decide which frames are "welfare-relevant" – that is, which frames should be considered when building a welfare relation. Among those frames deemed welfare-relevant, no frame takes precedence. In other words, choices in one welfare-relevant frame have the same weight as choices in another welfare-relevant frame. Thus, the technical role of frames is to exclude some choices when constructing the relation. Because welfare-relevance represents externally imposed structure and we wish to test predictive power with minimal additional structure, we do not make any assumptions about which frames are welfare-relevant in our baseline analysis.

2.2 Strict UCR and Revealed Preferences

Bernheim and Rangel (2009) define the following relation:

> x is strictly unambiguously preferred to y (denoted xP^*y) if whenever x and y are both available in some welfare-relevant frame, y is never chosen.

A fundamental difference between Strict UCR and the standard revealed preference relation is that with Strict UCR, multiple observations are considered jointly when inferring the ranking, whereas with revealed preferences, each observation is taken independently to do so.

Whenever choices are observed from all possible subsets of choice options (a condition we will call "full observability"), Strict UCR is guaranteed to be acyclic, which is useful property for performing welfare analyses.[12] Because all doubleton choice sets are observed, a necessary condition for xP^*y with full observability is:

> For some choice set (and for some welfare-relevant frame) where both x and y are available, x is chosen.[13]

As a consequence, with full observability xP^*y if and only if x is revealed preferred to y and y is never revealed preferred to x.

[10] A review of enhanced data sets, which include richer information than just final choices, is provided by Caplin (2016).

[11] Caplin and Martin (2012), Rubinstein and Salant (2012), and Benkert and Netzer (2016) also provide ways to use frames when assessing welfare.

[12] De Clippel and Rozen (2014) provide warnings and guidance on how behavioral theories should be tested when there is not full observability.

[13] This condition appears in Manzini and Mariotti (2014).

2.3 Other Welfare Relations

Several other welfare relations have been proposed in the literature that impose little *ad hoc* model structure. Unlike Strict UCR, the transitive core of Nishimura (2015) is generated from another relation R^* (in practice, the revealed preference relation). For this new relation, x is preferred to y (denoted $xcore(R^*)y$) if for all other options z, zR^*x implies zR^*y and yR^*z implies xR^*z. Nishimura (2015) shows that the transitive core makes recommendations that do not rely on arbitrary decisions from a modeler. This conservative approach is in the same spirit as Bernheim and Rangel (2009), but there are differences between Strict UCR and the transitive core. For the data sets we examine, the transitive core is slightly coarser than Strict UCR, but Nishimura (2015) presents theoretical examples where Strict UCR is coarser than the transitive core, specifically for models of time preferences with relative discounting and regret preferences. In addition, the transitive core has one substantial advantage over Strict UCR for empirical applications: even with less than full observability, the transitive core is always acyclic.

In a recent paper, Apesteguia and Ballester (2015) suggest a welfare relation based on a measure of rationality called the "swaps index". They provide a behavioral foundation for their index by identifying the axioms that characterize it. The corresponding welfare relation is found by choosing the preference order that is closest to (empirically) observed choices. To assess the closeness of a preference order, they look at the number alternatives that must be ignored in each choice set to match the preferences implied by choices to the candidate preference order. This approach uses choice set frequencies to overcome ambiguities, so is less conservative than Strict UCR in making welfare assessments.

An additional axiomatization of welfare inference was suggested by Chambers and Hayashi (2012). Broadly speaking, they introduce an individual welfare functional, which is a function from a choice distribution to a relation on alternatives, and they provide axioms to characterize the individual welfare functional. Like Apesteguia and Ballester (2015), this approach uses frequencies to overcome ambiguities, which enables them to generate a linear order. However, unlike Apesteguia and Ballester (2015), the frequencies they use are stochastic choice probabilities.

2.4 Other Empirical Findings

Bernheim, Fradkin, and Popov (2015) provide the first empirical implementation of Strict UCR to choice data.[14] They study the impact of making one retirement savings option the default, and because individuals appear to make inconsistent choices as the default option

[14] An application of concepts from Bernheim and Rangel (2009) also appears in Ambuehl, Bernheim, and Lusardi (2015).

changes, they use Strict UCR to identify the welfare impacts of such a change. However, to generate these welfare judgments, they make additional assumptions about the parametric form of utility and how different aspects of the choice correspondence relate to frames. We make no such additional assumptions, so our results are better situated to address the question of whether precise welfare assessments can be made with limited model structure.

Apesteguia and Ballester (2015) present an empirical application of the swaps index as a measure of rationality, but they do not provide results on the corresponding welfare relation. One challenge in empirically assessing the swaps welfare relation is that it may not be uniquely identified for data sets that do not have full observability, unlike the relations suggested by Bernheim and Rangel (2009) and Nishimura (2015). However, Apesteguia and Ballester (2015) formally prove that the mass of datasets for which the swaps welfare relation is not unique has mass zero, and when the welfare relation is not unique, the different welfare relations are likely to be very close to each other and coincide in the upper part of the rankings.

Finally, the results for our consumption data are not entirely unexpected, as Dean and Martin (2016) show for a panel of grocery store scanner data that households are "close" to being rational in the sense that few elements need to be removed to make a revealed preference relation acyclic. However, there are three ways in which the completeness of Strict UCR for our consumption data is surprising, even in light of their findings. First, Dean and Martin (2016) consider the minimal number of elements that need to be removed to make a revealed preference relation acyclic, whereas Strict UCR removes all ambiguous comparisons, which is in general much more conservative. Second, our panel is 8 years longer than theirs, so it provides a much tougher testing ground as it contains *5 times* more observations. Third, and most importantly, even if only a few elements need to be removed from a relation to make it acyclic, there is no guarantee that such a relation would be anywhere near complete.

3 Data

We use two very different data sets for our non-parametric applications of Strict UCR. The first one comes from an experiment carried out by Manzini and Mariotti (2010) and consists of choices among different sequences of delayed payments. The second one comes from the Nielsen Consumer Panel (NCP) and consists of grocery purchases recorded by the marketing firm Nielsen over 10 years. Among the many differences between these data sets are the subject pools (students versus shoppers), the choice setting (lab versus field), and the choice alternatives (choices from menus versus choices from budgets).

Despite these differences, both are representative of widely used types of data in the economic literature. Data from experiments in which subjects are asked to choose among delayed payments appear in many papers because they can be helpful when studying time-inconsistencies and time preferences (see Frederick, Loewenstein, and O'Donoghue 2002). Grocery store scanner data appears in a number of papers in the economics literature because it offers both price and quantity information at the UPC level across a wide range of households living in different markets with varying demographic characteristics. For instance, Aguiar and Hurst (2007) use grocery store scanner data to study the purchasing habits of retirees.

3.1 Experimental Data

The task that subjects undertook in this experiment was a simple choice task: subjects were asked to pick their preferred payment plan from a list of options. All payment plans were sequences of installment payments that were delayed by 3, 6 or 9 months. In each choice that a subject made, all of the listed plans had either two or three installments. In general, there were four types of plans, which were called the increasing plan (I), the decreasing plan (D), the constant plan (K), and the jump plan (J). For all plans, the total payment was 48€. The exact payments and delays for both sets of options are presented in tables 1 and 2. Additional details are available in Manzini and Mariotti (2010).

A unique feature of the experiment of Manzini and Mariotti (2010) is full observability: subjects were asked to choose from all possible subsets of choice options, which can be interpreted as eliciting the entire choice function.[15] Data with the property of full observability are appealing for two reasons. First, Strict UCR is guaranteed to be acyclic for such data. Second, such data provide a stringent test of the predictive power of Strict UCR.

Because subjects were asked to choose from all subsets for two sets of four plans, they made a total of 22 choices (each set of four plans corresponded to 11 choices). In the treatment where choices were incentivized, 102 individuals completed the experiment.[16]

3.2 Consumption Data

This data set is a balanced panel of purchases for single-person households that we have extracted from Nielsen's National Consumer Panel (NCP). NCP was formally known as the Homescan Consumer Panel because these grocery purchases are recorded using a scanner.

[15] Presenting all possible subsets is combinatorially challenging. For n alternatives, the number of choice sets is $2^n - n - 1$.

[16] Choices were not incentivized for an additional 54 subjects, so we do not include them in our analysis. However, the results do not qualitatively change if we also include those subjects in our analysis.

Table 1: Two installment plans.

Delay	I2	D2	K2	J2
3 months	16	32	24	8
9 months	32	16	24	40
Total €	48	48	48	48

Table 2: Three installment plans.

Delay	I3	D3	K3	J3
3 months	8	24	16	8
6 months	16	16	16	8
9 months	24	8	16	32
Total €	48	48	48	48

There are a growing number of papers that analyze NCP data.[17]

A unique feature of NCP is the duration of the panel. For the single-person households we study, the data set contains information on grocery purchases over a 10 year period. The length of this panel means that we have many observations, which allows us to perform a stringent test of the predictive power of Strict UCR.

3.2.1 Analysis Sample: Panelists

To construct our analysis sample, we start with purchases made by 140,827 households during a 10 year window (from 2004 to 2013). The full data set contains records for purchases of 565,583,696 goods from 98,684,440 store trips, and the purchases correspond to 3,692,767 Universal Product Codes (UPCs).

From these observations, we extracted a balanced panel of 1,183 singles who satisfy the following criteria over the entire 10 years:

1. Made purchases in every month;
2. Stayed single;
3. Did not move to a different market area (as defined by Nielsen);
4. Did not go into retirement.

While these restrictions may reduce the representativeness of our sample, the motivation for using such criteria is to keep preferences as stable as possible within each household over the 10 years we study.[18] For instance, we look at singles who stayed single because Dean and Martin (2016) find that singles and married couples have different levels of choice inconsistency. Also, we look at singles who do not go into retirement because Aguiar and Hurst (2007) find that retirement influences consumption patterns.

[17] As of February 2017, 49 working papers released by the Kilts Center use NCP. The current list of such papers can be found at http://www.ssrn.com/link/Chicago-Booth-Kilts-Ctr-Nielsen-Data.html.

[18] For an assessment of the representativeness of our sample, see section 6.2.1 of the appendix.

Nielsen registers purchases for a wide variety of products. To avoid products that can be stored for long periods, we have restricted ourselves to purchases of edible grocery products. This restriction reduces the original data to 365,014,702 goods purchased during 55,670,551 store trips and with 1,436,818 different UPCs. By further restricting the data of our balanced panel to singles, we end up with 5,897,440 goods purchased during 1,317,467 store trips, accounting for 329,753 UPCs.

For the singles in our analysis sample, the average expenditure per month and per panelist on the goods we have kept is $235.05, whereas the average total expenditure per month and panelist is $426.98 for all households and goods in the NCP over these 10 years.

3.2.2 Analysis Sample: Bundles

For a given month, each panelist has a corresponding bundle, made of 6 goods with quantities expressed in ounces. In order to construct bundles, we aggregate all purchases made during a month and aggregate the purchases into 6 categories given by Nielsen: alcoholic beverages, dairy products, deli foods, dry groceries,[19] frozen food, and packaged meat. Average budget shares for these product categories are given in table 3. Aggregation over a month is done for two reasons: first, to compensate for the fact that panelists do not in general shop every day; and second, to assuage concerns about the storage of products. Because the units of measure are not necessarily the same between UPCs, we have first converted every product quantity into ounces (either fluid or solid), so that each aggregated good is quantified in ounces.

Building bundles by aggregating over categories and time periods is common in the literature that uses scanner data. For instance, Dean and Martin (2016) build similar bundles to perform a revealed preference analysis using scanner data; Hinnosaar (2016) aggregates beer into one homogeneous good; and Handbury, Weinstein, and Watanabe (2015) study inflation with price indices built in a similar fashion.

3.2.3 Analysis Sample: Prices

The panelists are divided by Nielsen in 58 markets, which correspond roughly to large metropolitan areas of the United States. These markets and the number of panelists in each market are given in figure 11 of the appendix. For each market, we have built a price vector, which is a unit price for each aggregated good expressed in dollars per ounce. To

[19] The category dry grocery has a subcategory of pet food which we have removed. First, it is not edible, and second, there should be little substitution between pet food and human food.

Table 3: Average budget shares (expenditure on a product/good category in proportion to total expenditure) in a month.

Product	Average	Standard Deviation
Alcoholic beverages	5.71%	14.26%
Dairy products	16.25%	14.13%
Deli foods	2.56%	4.59%
Dry groceries	62.33%	19.89%
Frozen food	10.33%	10.15%
Packaged meat	2.82%	4.51%

build this price vector, we use a "Stone" price index:

$$P_{Jt} = \sum_{i \in J} w_{it} p_{it}$$

where P_{Jt} is the price index for good category J in period t, w_{it} is the budget share for UPC code i in period t, and p_{it} is the mean price for UPC code i in period t.[20]

We know that there is measurement error in prices, in particular because panelists sometimes enter prices themselves. Indeed, Nielsen uses the following data collection methodology: each panelist has a scanner at home and scans all purchases once home. Nielsen matches a price to the UPC by linking these purchases to a database of store prices. If a price is missing, the panelist is required to input the price by hand. To incentivize the panelists to make correct entries, Nielsen has different cash reward programs, but some price entry errors are inevitable. To reduce the impact of these and other price measurement errors, we take two steps. First, we use purchases from the entire panel to construct market prices, not just purchases from our analysis sample. Second, we do not consider entries in the upper 2.5% and lower 2.5% of the price distribution for a product category in a period.

3.2.4 Additional Considerations

Of course, grocery purchases are just one component of a household's regular expenditures. An implicit assumption made when considering the consistency of these choices is separability between grocery purchases and the rest of a household's expenditures. A justification for separability is that households may have a separate grocery budget. While strong, separability is a standard assumption in applications of revealed preference techniques to consumption data (for instance, see Koo 1963; Blundell, Browning, and Crawford 2003; Dean and Martin 2016).

[20] Dean and Martin (2016) do not find significant differences in revealed preference violations when using Stone, Laspeyres, or Paasche indices.

Another standard assumption is that all panelists from the same market face the same prices in a given period. This assumption is necessary, because if a household does not buy from a product category in a given period, prices are not identified for that category. Because we are using market prices, our analyses capture the impact of sustained and widespread price changes, not very temporary and local ones. Once again, this is a standard assumption in the applied revealed preference literature.

The last important assumption made for empirical testing is the stability of preferences over time, which is needed to make comparisons across periods. If preferences were to change, then having violations of revealed preferences would only mean that preferences have changed and would not be informative *per se*. While this assumption is also standard in the applied revealed preference literature, we recognize that it could potentially impact our results. However, even if preferences are indeed unstable over time, this should work against the precision of Strict UCR, which would make the test of predictive power even tougher.[21]

4 Results

In this section we study the empirical performance of Strict UCR for both choice settings: menus of payment plans (the experimental data) and bundles of grocery purchases (the consumption data). We also compare the performance of Strict UCR to the performance of other approaches, both model-free and model-based. For this analysis, we focus primarily on five questions:

1. Do demand functions exhibit choice inconsistencies?
2. Does Strict UCR produce clear welfare guidance?
3. If yes, what kinds of alternatives are welfare optimal?
4. Does Strict UCR differ from other "model-free" approaches?
5. How does Strict UCR perform compared to model-based approaches?

To answer these questions, we examine both observed demands as well as hypothetical ones. Hypothetical demands are generated by making random selections from the choice sets that individuals face in each choice setting, which allows us to explore a wide set of possible

[21] Also, as mentioned previously, the restrictions we apply to the subject pool are designed to minimize the degree of preference instability over time. Indeed, in our robustness checks we find that the predictive power of Strict UCR does not improve by considering a shorter time horizon.

demands. For the experimental data, we simulate 1,000,000 times an individual choosing uniform random from each choice set, and for the consumption data, we simulate 100 times the 120 bundle choices for each individual.[22]

4.1 Are Choices Inconsistent?

A standard marker for choice inconsistency is the presence of cycles in the preferences revealed by choice. To understand whether revealed preference cycles are prevalent in our data sets, we examine several dimensions of cyclicity: the breadth, the depth, and the length of revealed preference cycles. Along all three dimensions, we find that choice inconsistencies are prevalent in observed demands, but are far more prevalent in hypothetical demands.

First, we examine the "breadth" of revealed preference cycles: how many individuals exhibit at least one cycle. For both of our data sets, a majority of individuals have revealed preference cycles, as shown in figure 1. In the experimental data, 52.94% of individuals violate acyclicity for at least one set of options.[23] In the consumption data, all individuals violate acyclicity.

This wide breadth of revealed preference cycles is consistent with findings in the empirical literature on revealed preference testing.[24] In the laboratory experiments of Choi, Fisman, Gale, and Kariv (2007), around 35% of subjects have revealed preference cycles for choices from allocations over risky assets. In the large-scale field experiment of Choi, Kariv, Müller, and Silverman (2014), around 90% of subjects exhibit revealed preference cycles for a similar choice task. For consumption data, there is a long history of papers that detect revealed preference cycles. In one of the earliest computer-based studies of consumption data, Koo (1963) examined a panel of food purchases from 1958 for 215 Michigan households and concluded: "In an empirical study, it is not likely that one will find many individuals who are either entirely consistent or inconsistent." In a recent paper, Dean and Martin (2016) find that around 71% of households exhibit revealed preference cycles in a two-year balanced panel of grocery purchases.

For the choice settings we study, how likely is it for a demand function to exhibit choice inconsistencies? The answer is quite likely. For the experimental data, almost all of the 1,000,000 of the hypothetical demands we generated contain cycles (less than 0.01% are acyclic), and for the consumption data, almost 100% of hypothetical demands contain cycles (again less than 0.01% are acyclic).

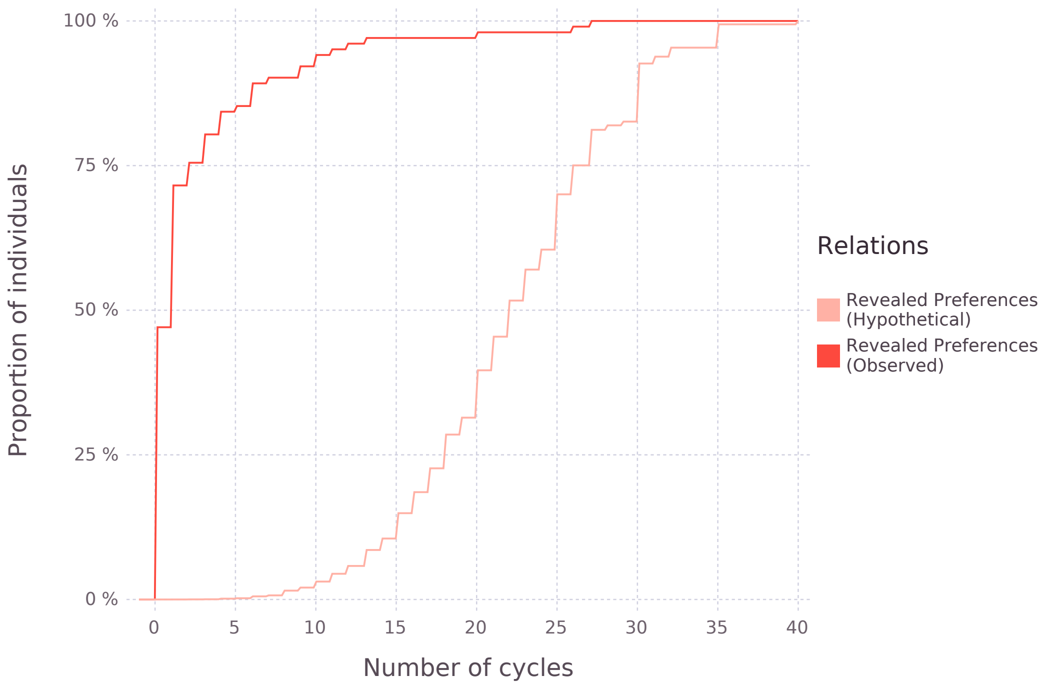
[22] For this simulation, we follow the procedure given in the appendix.

[23] Recall that subjects are asked to choose from every subset of two different sets of options.

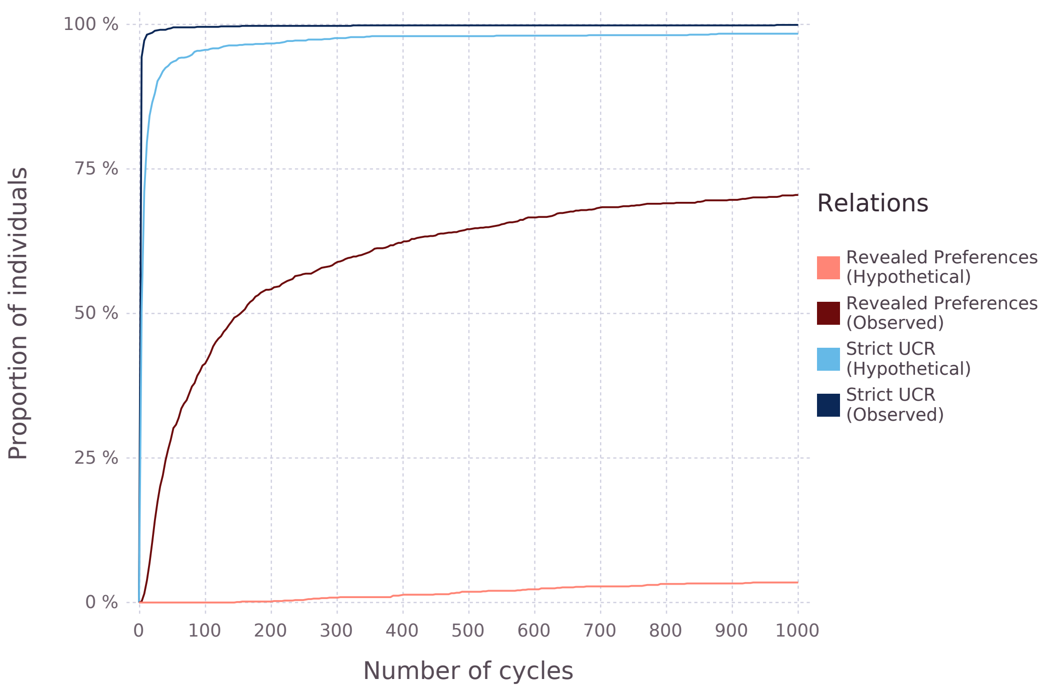
[24] See Crawford and De Rock (2014) and Adams and Crawford (2015) for a review of the empirical literature on revealed preference testing.

Figure 1: Number of cycles at the individual level.

(a) CDF of the number of cycles in the experimental data.



(b) CDF of the number of cycles in the consumption data (for <1,000 cycles).



Second, we examine the "depth" of revealed preference cycles: how many cycles each individual exhibits.[25] While a single cycle is enough to violate acyclicity, there is a sense in which the more cycles that a set of choices generate, the more inconsistent are those choices.

In the experimental data, revealed preferences have a median number of cycles of 2, conditional on the individual being cyclic. As shown in figure 1a, while 75.49% of subjects have 2 cycles or less, some subjects have a large number of cycles – up to a maximum of 27. In the consumption data, revealed preferences have a median of 153 cycles. For this data, 41.50% of individuals have 100 cycles or less and 70.5% have a 1,000 cycles or less. Here again, as figure 1b shows, some individuals have a large number of cycles, with 18.09% having at least 10,000 cycles and 6.42% of them having more than 10,000,000 cycles.[26]

Hypothetical demands have substantially more depth in their revealed preference cycles, as shown in table 4. For the experimental data, the median number of cycles jumps to 22, and for the consumption data, the median number of cycles is more than 1,000,000.

Table 4: Summary statistics on revealed preference cycles.

	Experimental data		Consumption data	
	Observed	Hypothetical	Observed	Hypothetical
Percentage acyclic	47.06%	<0.01%	0%	<0.01%
Median number of cycles	2	22	153	$> 10^6$
Maximum number of cycles	27	40	$> 10^6$	$> 10^6$
Median number of length 2 cycles	1	8	27	67

Last, we examine the "length" of revealed preference cycles: how many alternatives it takes to create a cycle. It could be argued that cycles of shorter lengths are more problematic than ones of longer length because they reflect more directly conflicting information about preferences. The shortest possible cycles (of length 2) correspond to violations of the Weak Axiom of Revealed Preference (WARP).

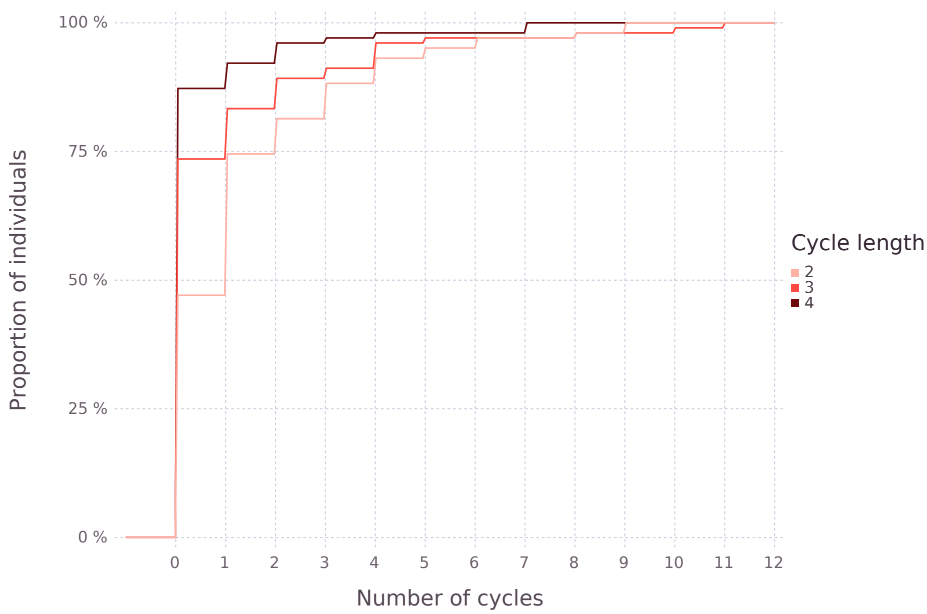
Because in the experiment there are just four alternatives in both sets of options, cycles can be of length 2, 3, or 4. Most cycles are of length 2 and 50% of cyclic individuals only have cycles of length 2. However, 24.07% of cyclic individuals have at least one length 4 cycle. Figure 2a shows that some subjects have a relatively large number of length 3 and length 4 cycles (up to 11 cycles of length 3 per individual and up to 7 cycles of length 4 per individual).

[25] When multiple choices reveal that x is preferred to y, we count that relation just once when counting revealed preference cycles. For instance, if two choices reveal that x is preferred to y and one choice reveals that y is preferred to x, those choices generate just one revealed preference cycle.

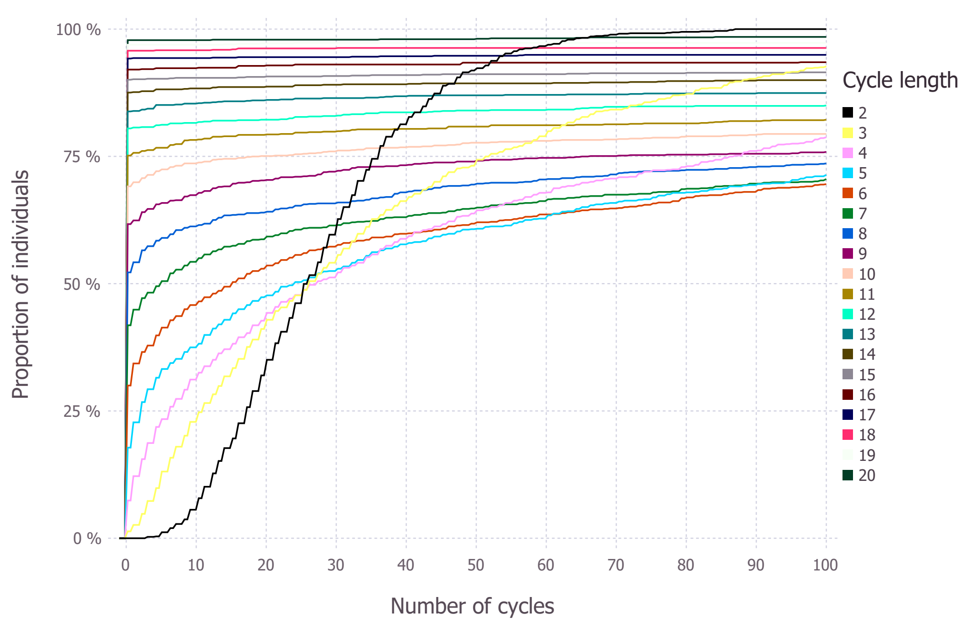
[26] For computational reasons, we just look at 10,000,000 cycles per individual for observed demands and 1,000,000 cycles per individual for hypothetical demands, so in several cases we just present lower bounds.

Figure 2: Length of revealed preferences cycles at the individual level for observed demands.

(a) CDF of cycle lengths in the experimental data.



(b) CDF of cycle lengths in the consumption data (for length <20).



For our consumption data, the picture is more complicated, but there are similarities with the experimental data. As in the experimental data, cycles of length 2 are widespread. In fact, all individuals have cycles of length 2, sometimes a large number of them, as the CDF of length 2 cycles given in figure 3 shows. However, revealed preference cycles can be up to a length 120 in the consumption data, so as a consequence, cycles of length 2 represent a smaller proportion of the total cycles. The median number of length 2 cycles is 27 (12.92% of the median number of cycles).

Two interesting patterns are revealed by figure 2b. First, most individuals have smaller cycles, but their number is limited. Second, fewer individuals have longer cycles, but when they have such cycles, these cycles are numerous. For instance, the maximal number of cycles of length 3 is 244, whereas the same number for cycles of length 10 is 1,118,618. On the other hand 98.40% of individuals have cycles of length 3, whereas only 34.57% of individuals have cycles of length 10. This stems from the underlying structure of cycle generation.[27] Loosely speaking, when several short cycles are connected, they create a large number of larger cycles. Because of this, while the number of cycles can become very large, the number of relation elements involved in those cycles can stay small, as shown later in section 4.2.2 and in particular figure 6.

4.2 Does Strict UCR Produce Clear Welfare Guidance?

For Strict UCR to provide clear welfare guidance, it needs to be free of cycles and have high predictive power. In this section, we examine the acyclicality, the coarseness, and the predictive power of Strict UCR for both observed demands and hypothetical demands. In both choice settings, we find that Strict UCR is relatively free of cycles and has high predictive power for observed demands. For the experimental data, Strict UCR is also free of cycles for hypothetical demands, but has much lower predictive power. For the consumption data, Strict UCR has high predictive power when acyclic, but almost always contains cycles.

4.2.1 Cyclicality of Strict UCR

As we did for revealed preferences, we also look at the breadth, the depth, and the length of cycles for Strict UCR in both data sets, and we find that the prevalence of cycles is dramatically lower with Strict UCR along all three dimensions for observed demands. However, Strict UCR contains many cycles for hypothetical demands in the consumption data, which renders it unable to provide clear welfare guidance in that choice setting.

Again we first look at the breadth of cycles. For the experimental data, all 52.94% of

[27] See Johnson (1975) for a formula of the number of cycles of given length.

Figure 3: CDF of the number of cycles of length 2 in the consumption data.

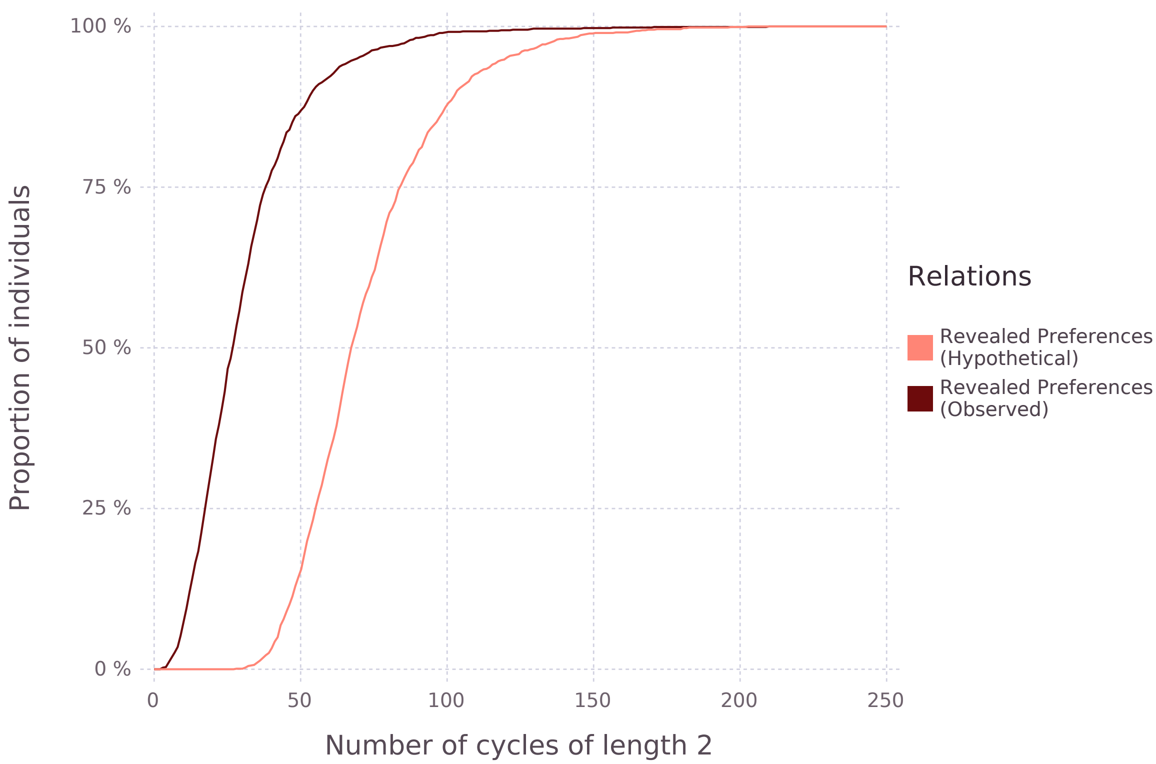


Table 5: Summary statistics for Strict UCR cycles.

	Experimental data		Consumption data	
	Observed	Hypothetical	Observed	Hypothetical
Percentage acyclic	100%	100%	79.54%	7.69%
Median number of cycles	0	0	0	3
Maximum number of cycles	0	0	57,662	$> 10^6$

Table 6: Median number of cycles.

		Revealed Preferences	Strict UCR
Experimental data	Observed	2	0
	Hypothetical	22	0
Consumption data	Observed	153	0
	Hypothetical	$> 10^6$	3

subjects who have preference cycles with the revealed preference approach have no cycles at all with Strict UCR. In fact, Strict UCR is always acyclic for demands in the experiment (both observed and hypothetical) because the full observability condition is satisfied. On the other hand, Strict UCR is not guaranteed to be acyclic for the consumption data because full observability is not satisfied. Indeed, Strict UCR violates acyclicity for 20.46% of individuals in that data set. However, this still represents a substantial reduction from the 100% of individuals who violate acyclicity with revealed preferences, which is presented graphically in figure 1b. On the other hand, Strict UCR violates acyclicality for 92.31% of hypothetical demands for the consumption data.

Next we look at the depth of cycles. A comparison of Strict UCR and revealed preference along this dimension is available in table 6. Observed demands have many fewer cycles than hypothetical ones, and Strict UCR has many fewer cycles than revealed preferences. For the 20.46% of individuals who violate acyclicity for Strict UCR in the consumption data, this difference is even starker: the median number of cycles drops from 4,897 with revealed preferences to 1 with Strict UCR. Also, the depth of cycles is quite shallow: 89.67% have less than 10 cycles and 97.93% less than 100.[28] Figure 4 demonstrates that for individuals with Strict UCR cycles, the depth of Strict UCR cycles is less than the depth of revealed preference cycles, even if we just look at those subjects with few cycles. A similar pattern is observed for the 92.31% of hypothetical demands that contain Strict UCR cycles, so despite the breadth of these cycles, there is not much depth.

Finally, we look at the length of Strict UCR cycles. Figure 5 shows that for longer cycles

[28] The maximal number of cycles that are possible with Strict UCR is 135,954.

Figure 4: CDF of the number of cycles in the consumption data (for <100 cycles) for observed demands when Strict UCR contains cycles.

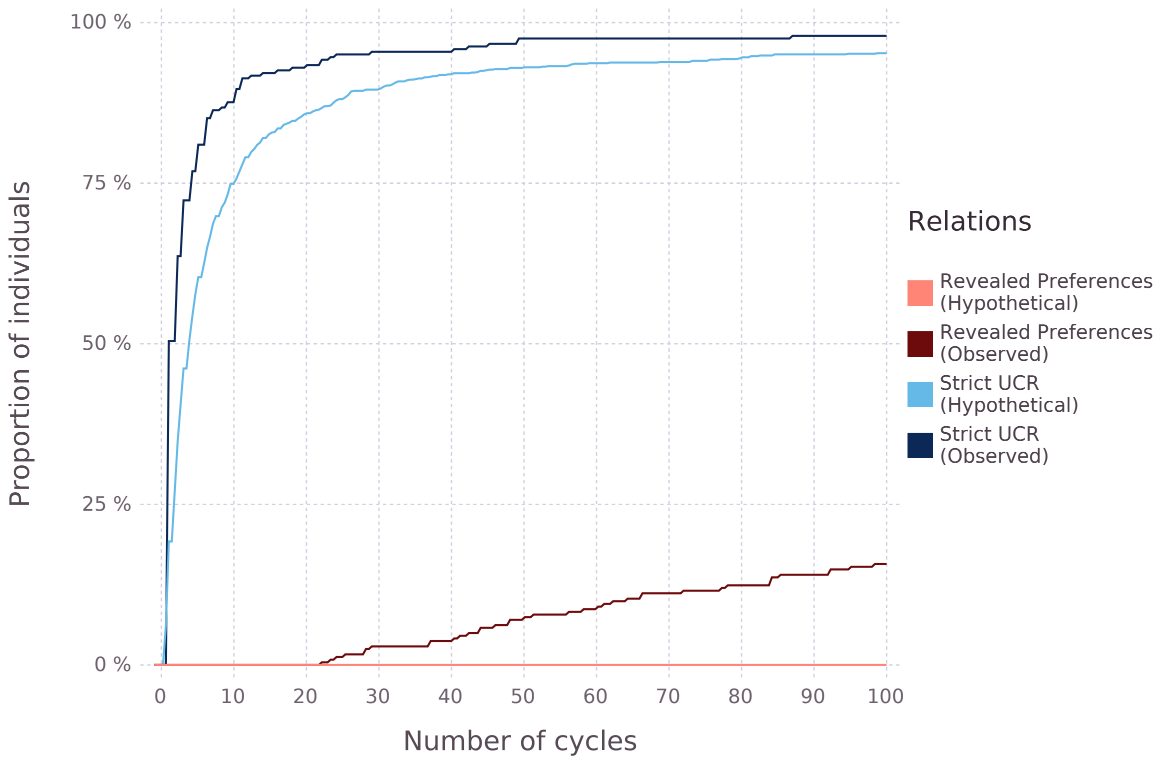
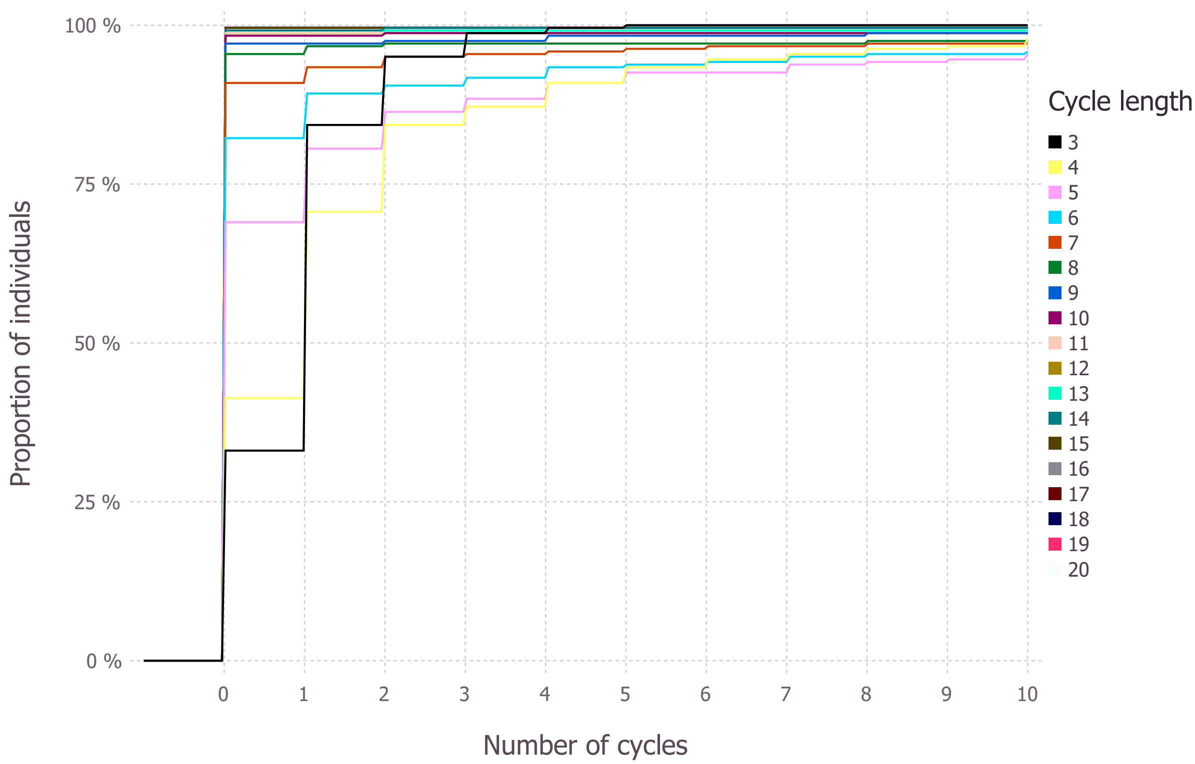


Figure 5: CDF of Strict UCR cycle lengths in the consumption data (for length <20) for observed demands when Strict UCR contains cycles.



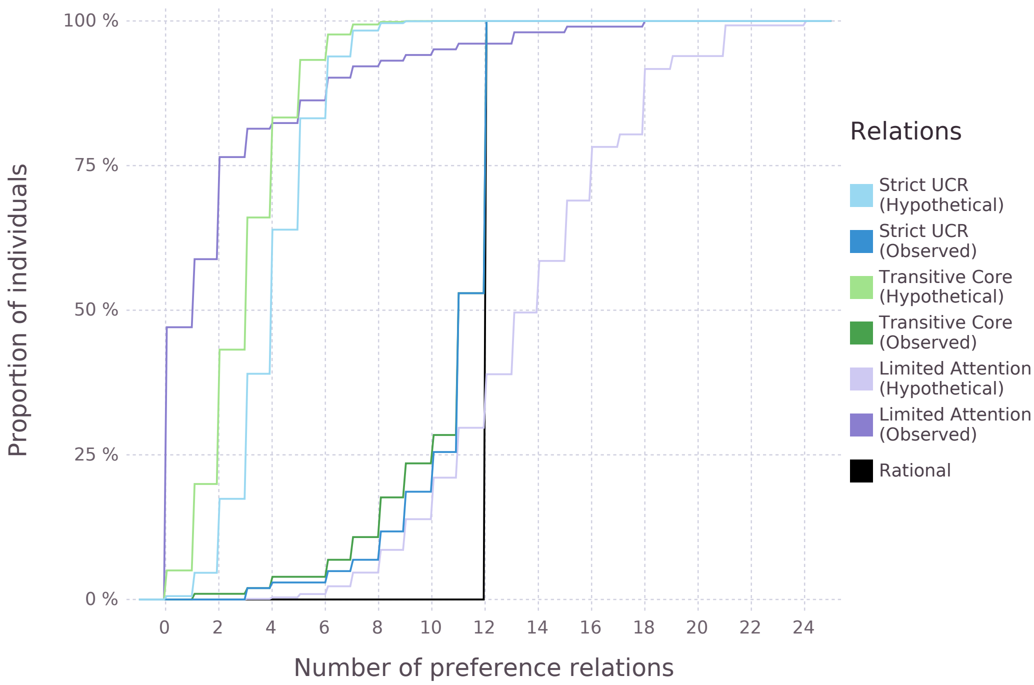
lengths, Strict UCR exhibits a similar pattern as revealed preferences, though with many fewer cycles. Once again, more individuals have smaller cycles, but the number of smaller cycles per individual is relatively small. For instance, the maximal number of cycles of length 3 is 5, whereas the same number for cycles of length 10 is 4,896. On the other hand, 99.17% of individuals with cyclical Strict UCR have cycles of length 3, whereas 66.12% have of cycles of length 10.

4.2.2 Coarseness of Strict UCR

For observed demands, we find that Strict UCR is far from coarse. This is despite the presence of many choice inconsistencies, which could produce ambiguities that Strict UCR cannot resolve. For the experimental data, 47.06% of subjects have a complete and acyclic Strict UCR for both sets of payment plans, and 81.37% of subjects have a complete and acyclic Strict UCR for at least one set of payment plans. 57.84% have a complete and acyclic relation for plans with two installments, and 70.59% have a complete and acyclic relation for plans with three installments.

Figure 6: Number of relation elements at the individual level.

(a) CDF of the number of relation elements in the experimental data.



(b) CDF of the number of relation elements in the consumption data.

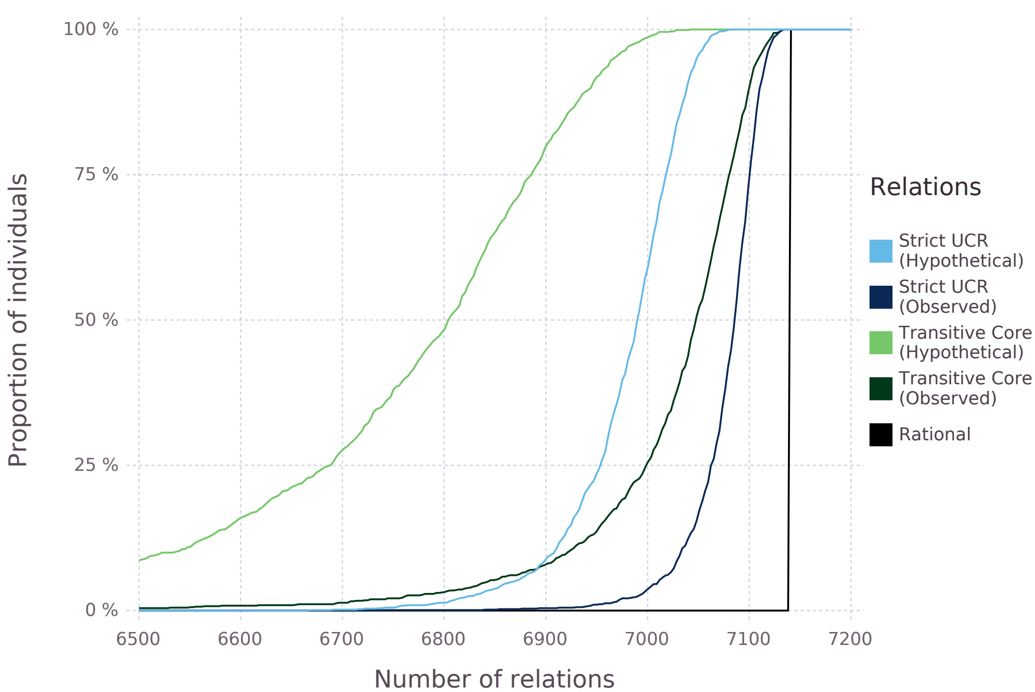


Figure 6a presents the number of relation elements for Strict UCR for the experimental data at the individual level. Because individuals make choices separately from two sets of four options, an individual with complete and acyclic preferences has 12 elements. As the figure shows, of the 52.84% of subjects who do not have a complete Strict UCR, 51.85% are just one element short, and 12.96% are just two short.

However, for hypothetical demands in the experimental data, we instead find evidence that Strict UCR is "a coarse binary relation that becomes even more so as the behavioural data set grows" (Salant and Rubinstein 2008). Figure 6a also shows that hypothetical demands produce many fewer Strict UCR elements, with no individual having a complete relation and a median number of relation elements of 4.

Unlike the experimental data, Strict UCR is never complete for observed demands in the consumption data. However, the average number of relation elements for Strict UCR is 7,076.95, which is 99.12% of the comparisons in complete and acyclic preferences over chosen bundles.[29] Conditional on Strict UCR being acyclic, the average number of relation elements for Strict UCR is 7,084.95, which is 99.23% of the comparisons in complete and acyclic preferences. For the small percentage of hypothetical demands that have acyclic Strict UCR, the average number of relation elements is lower, but still high, reaching 97.74% of completeness.

Figure 6b presents the number of relation elements for Strict UCR for the consumption data. As the figure shows, 99.28% of individuals have at least 7,000 relation elements and 24.09% have at least 7,100 (99.44% of a complete and acyclic relation). These numbers drop when we consider hypothetical demands, as the proportion of demands with more than 7,000 relations is 40.83% and none of them has more than 7,100 relations.

4.2.3 Predictive Power of Strict UCR

The welfare recommendations (or predictions) for a relation are the alternatives from a choice set that are consistent with that relation. To assess the clarity of the welfare recommendations made by Strict UCR, we determine the predictive power of Strict UCR for choices in both data sets.

Our approach is simple: to generate a prospective "virtual" choice set, we select a random subset of alternatives.[30] If only a few alternatives in that choice set are consistent with the

[29] Because there are 120 choice options, a complete and acyclic preference relation contains 7,140 comparisons.

[30] Specifically, we first randomly draw a set size from the uniform distribution and then randomly draw options (each equally likely) without replacement until we get a virtual choice set of that size. We randomly draw set sizes because if we draw uniformly from all possible subsets of choices alternatives, then we are much more likely to get subsets of a particular size. For the experimental data, the most frequent subset

relation, then the predictions of the model are tight. For the experimental data, every virtual choice set also appears in the data set, so the choice sets are "in-sample". On the other hand, for our consumption data, virtual choice sets do not appear in the data set, so they are "out-of-sample".[31]

Several papers use predictive power for a different aim: to determine the "predictive success" of revealed preference tests (Manzini and Mariotti 2010; Beatty and Crawford 2011; Andreoni, Gillen, and Harbaugh 2013; Dean and Martin 2016). These papers determine the fraction (or area) of demands that would pass a revealed preference test and then subtract this from an indicator for whether or not the observed choices passed the test. Their goal is to determine whether or not it is difficult for a set of choices to pass a revealed preference test for a given data set.

For the experimental data, figure 7 shows the number of alternatives from the choice set that are consistent with Strict UCR for each size of choice set. For all choice sets, at least one alternative is guaranteed to be consistent with Strict UCR, so the minimum number of "predicted" alternatives is 1. Unsurprisingly, the predictive power for choice sets of size 3 is higher than the predictive power for choice sets of size 4. Averaging across choice set of different sizes, the average number of predicted alternatives is 1.17 and 84.36% of sets have only one predicted alternative. The predictive power is slightly higher with 3 installment plans, with an average of 1.14 predicted alternatives and 86.63% of sets with only one predicted alternative against 1.19 and 82.09% for two installment plans.[32] Predictions are much looser with hypothetical demands: the average number of predicted alternatives is 1.93 and only 23.36% of choice sets have only one predicted alternative.

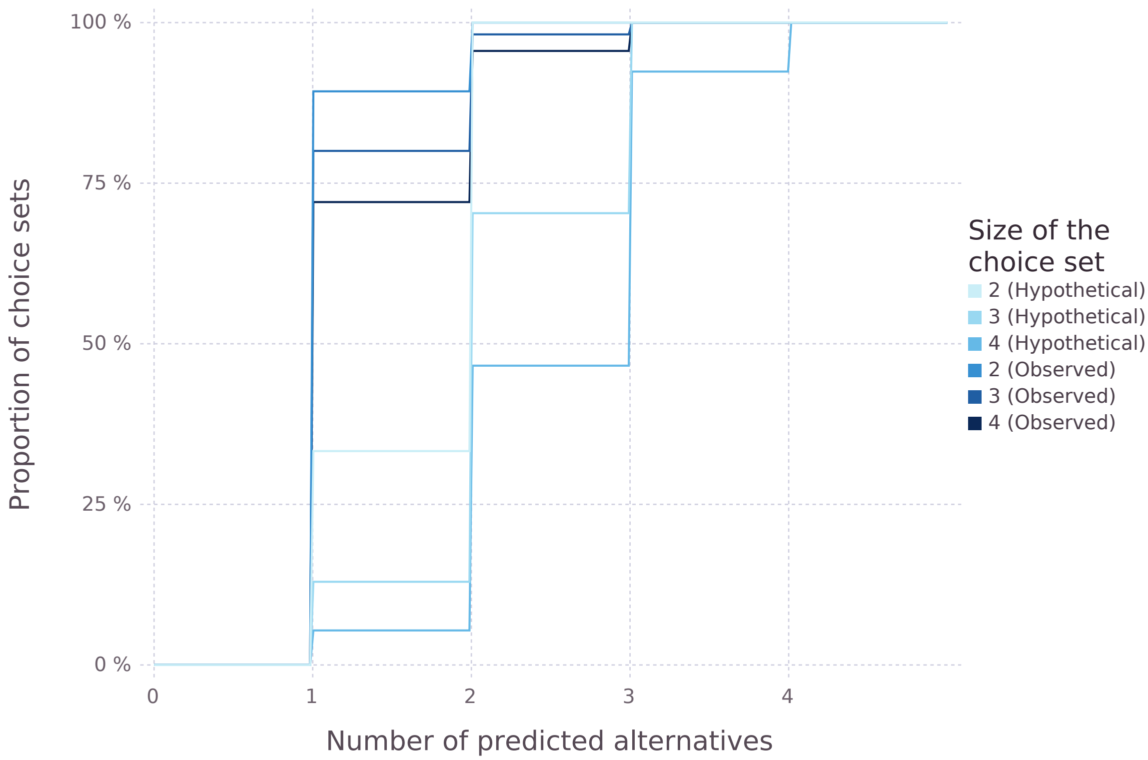
For the consumption data, looking at observed demands with acyclic Strict UCR, all choice sets have 4 or fewer predicted alternatives, even for those choice sets of size 120. Averaging across choice set of different sizes, the average number of predicted alternatives is 1.03 for observed demands with acyclic Strict UCR and 1.10 for hypothetical ones with acyclic Strict UCR. With observed demands with acyclic Strict UCR, the vast majority of sets (96.85%) have only 1 predicted alternative. For hypothetical demands with acyclic Strict UCR, this is still true for a majority of virtual sets, but many fewer, as sets with only 1 predicted alternative represent 61.92% of the total.

size is 2, and for the consumption data, the most frequent subsets size is 60.

[31]For the consumption data, we just draw sets of previously chosen bundles because standard revealed preference tests just consider chosen bundles (see Varian 1982). Such menus are not observed in practice.

[32]The difference in the average number of predicted alternatives is significant as the p-value is 10^{-4} for a two-sided t-test.

Figure 7: CDF of the number of predicted alternatives per choice set for Strict UCR in the experimental data.



4.3 What is Welfare Optimal?

Because Strict UCR is cyclic or has little predictive power for most hypothetical demands, it is not able to offer much in the way of welfare guidance for these demands. On the other hand, because Strict UCR is acyclic and provides tight predictions for observed demands, we can investigate what kinds of welfare guidance those predictions provide. For the experimental data, we find that this guidance points largely towards plans with sooner payments, which is consistent with delay aversion.

Recall that in the experiment, subjects chose from all subsets of four payment plans, which were of different types: a decreasing plan (D), a constant plan (K), an increasing plan (I), and a jump plan (J). They did this for both four two-installment plans and four three-installment plans (see tables 1 and 2).

One feature of the experimental data we have not leveraged so far is the natural ranking in terms of welfare for the different plans. Looking at tables 1 and 2, it is clear that an impatient value maximizer individual will have the preference ordering $D \succ K \succ I \succ J$. Do we observe behavior consistent with that pattern?

One way to answer this question is to look at the number of times that the maximal alternative (according to an impatient value maximizer individual) is predicted to be chosen for a given virtual choice set. For Strict UCR, the maximal alternative of an impatient and value maximizer is in the predicted set 84.54% of the time. Looking at the individual level, 62.75% of individuals are always consistent with value impatient value maximization, while only 43.14% are individuals are always consistent according to revealed preferences

Another answer is provided by table 7, which shows the relations for Strict UCR and revealed preferences that are generated by subject choices in the experiment. Rows dominate the columns, so that in the first row of the table, we have the relations $D \succ K$, $D \succ I$, and $D \succ J$. Unlike previous analyses, for this analysis we have pooled relations of similar types for two-installment plans and three-installment plans. For instance, if $D \succ K$ for either installment plan length, then an individual would be counted as having a relation for $D \succ K$.[33]

Table 7a shows us that, according to Strict UCR, D often dominates the other alternatives and that in general the plans that give money later do not provide higher welfare than plans that give money earlier (looking at the lower left triangle of the table). On the other hand, there are more ambiguities for revealed preferences, as shown in table 7b.

[33]Because Strict UCR is not always complete, the two boxes $D \succ K$ and $K \succ D$ do not add up to a exactly 100% in table 7a. Also, because the relations are pooled across installment plan lengths, a fully consistent individual could have both $D \succ K$ and $K \succ D$ if those relations came from different installment plan lengths.

Table 7: Proportion of individuals who have a Strict UCR relation where $row \succ column$.

(a) Strict UCR.

	D	K	I	J
D		52.45%	78.92%	80.88%
K	25.98%		84.80%	85.78%
I	14.71%	4.41%		86.76%
J	9.80%	5.39%	5.88%	

(b) Revealed Preferences.

	D	K	I	J
D		74.02%	85.29%	90.20%
K	47.55%		95.59%	94.61%
I	21.08%	15.20%		94.12%
J	19.12%	14.22%	13.24%	

4.4 Another "Model-Free" Approach: Transitive Core

As discussed in section 2, Strict UCR is just one of several "model-free" approaches proposed in the literature. Another proposition is the transitive core of Nishimura (2015), which like Strict UCR, takes a conservative approach to constructing a welfare relation. For both data sets, we find that Strict UCR and transitive core perform similarly well, although the transitive core is slightly coarser. Importantly, both approaches provide similar welfare guidance. One substantive difference is that unlike Strict UCR, the transitive closure is always acyclic.

Figure 6 shows that the CDF of the number of relations is similar in both scale and shape between Strict UCR and the transitive core. For the experimental data, 47.06% of subjects have a complete and acyclic transitive core for both sets of payment plans, and 81.37% of subjects have a complete and acyclic transitive core for at least one set of payment plans. Like Strict UCR, the transitive core is never complete for the consumption data, but is far from coarse (as illustrated in figure 6b). The average number of relation elements for the transitive core is 7,022.73, which is 98.36% of the comparisons in a complete and acyclic preference relation. As given by table 8, the number of relations at the individual level is highly correlated between Strict UCR and transitive closure.

Table 8: Correlation in the number of relations at the individual level between the transitive core and Strict UCR.

	Observed demands	Hypothetical demands
Experimental data	94.77%	62.71%
Consumption data	96.71%	98.75%

However, the transitive core is slightly coarser than Strict UCR, as shown in figure 6. In fact, using a double-sided Kolmogorov-Smirnov test, we reject the null hypothesis that the empirical CDFs are drawn from the same distribution at a 95% confidence level.

When looking at the number of predicted alternatives for both real and hypothetical demands in the experiment, figure 7 and 8 reveal similar distributions for both approaches. The relatively lower coefficient of correlation for hypothetical demands in the experiment is due to the fact that there are almost no relations, as figure 6a indicates. Indeed, there is median of just 4 relations with Strict UCR and of just 3 relations in the transitive core for hypothetical demands.

In fact, the predicted alternatives are exactly the same for Strict UCR and the transitive core 97.64% of the time. This proportion drops to 87.18% for hypothetical demands, which is still quite high. Not surprisingly, the welfare recommendations of Strict UCR and transitive core are similar, as suggested by a comparison of 7a and 9.

Table 9: Proportion of individuals who have a transitive core relation where *row* $\succ$ *column*.

Alternative	D	K	I	J
D		52.45%	77.94%	79.90%
K	25.49%		80.39%	85.78%
I	14.21%	4.41%		82.35%
J	9.80%	5.39%	5.88%	

However, unlike Strict UCR, the transitive core is always acyclic for our consumption data, even without full observability (as given by Axiom 2 in Nishimura 2015). This is particularly striking when looking at hypothetical demands, where cycles are quite prevalent with Strict UCR and absent with the transitive core.

4.5 A Model-Based Approach: Limited Attention

Unlike the "model-free" approach of Bernheim and Rangel (2009), Masatlioglu, Nakajima, and Ozbay (2012) evaluate welfare using a model in which a bias is explicitly taken into account (limited attention) when extracting preferences from choice. They show in particular how limited attention can lead to different welfare guidance for Strict UCR.

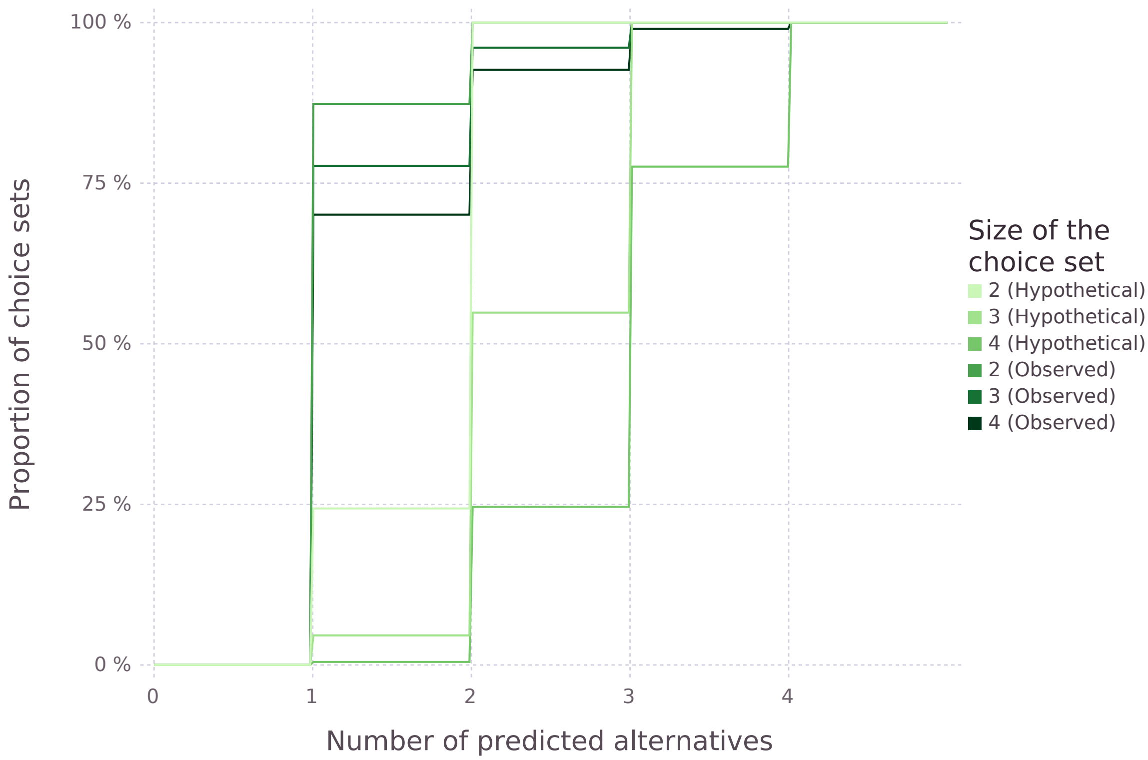
Masatlioglu, Nakajima, and Ozbay (2012) define a direct revealed preference relation P^0 in the following way:

$$xP^0y \text{ if there exists } T \text{ such that } c(T) = x \neq c(T \backslash y).^{34}$$

According to this approach, x is preferred to y if whenever we take y out of the choice set, x is no longer chosen. The reason is that the decision maker must have been paying attention to y in order for it to have impacted choice. The limited attention relation is then the transitive

[34] Where c is a choice correspondence, T is a set of alternatives.

Figure 8: CDF of the number of predicted alternatives per choice set for the transitive core in the experimental data.



closure of the relation obtained above. Two points should be kept in mind when using their model of limited attention:

1. If someone verifies WARP, her relation will be empty. Limited attention yields to relations only when WARP is violated. It means that for almost half of our experimental subjects, limited attention will produce an empty relation.

2. Because this approach requires menu choice, we only investigate limited attention in the experiment, as these observations are generally not available on consumption data.

A striking feature of figure 6a is the difference between observed and hypothetical demands in the number limited attention relations. Whereas the median number of relations is 1 for observed demands, the median number of relations is 14 for hypothetical demands. Because limited attention produces few relations for observed demands, it is not surprising that it offers limited welfare guidance for these demands, as demonstrated by table 10.

Table 10: Proportion of individuals who have a limited attention relation where *row* $\succ$ *column*.

Alternative	D	K	I	J
D		4.90%	11.76%	13.24%
K	6.86%		16.67%	13.24%
I	4.90%	3.92%		7.84%
J	9.31%	8.33%	7.84%	

The median number of relations for limited attention are in contrast to the corresponding numbers for Strict UCR, which are 11 for observed demands and 4 for hypothetical demands. In fact, the correlation in the number of relations between Strict UCR and limited attention at the individual level is -97.81% – they are almost perfectly negatively correlated. As a consequence, limited attention is meaningful when Strict UCR is not, in particular when Strict UCR is ambiguous and therefore close to empty of welfare guidance.

However, because a complete and acyclic relation has just 12 relations, the fact that hypothetical demands produce a median of 14 relations with limited attention suggests that many of these demands must be cyclic. Indeed, the median number of cycles for hypothetical demands is 5, whereas it is 0 for observed ones. Figure 9 shows the CDF of cycles with limited attention. Overall, 91.18% of observed demands produce an acyclic relation with observed demands, and 13.83% of hypothetical demands produce an acyclic relation. For Strict UCR, 100% of observed and hypothetical demands produce an acyclic relation.

Figure 9: CDF of the number of cycles with limited attention in the experimental data.

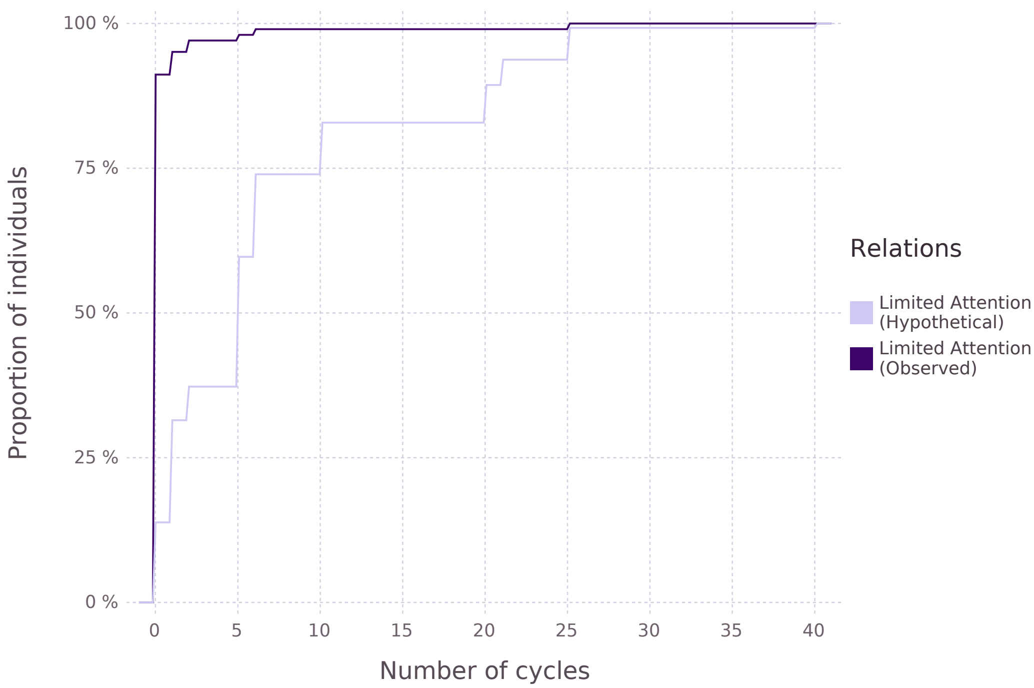
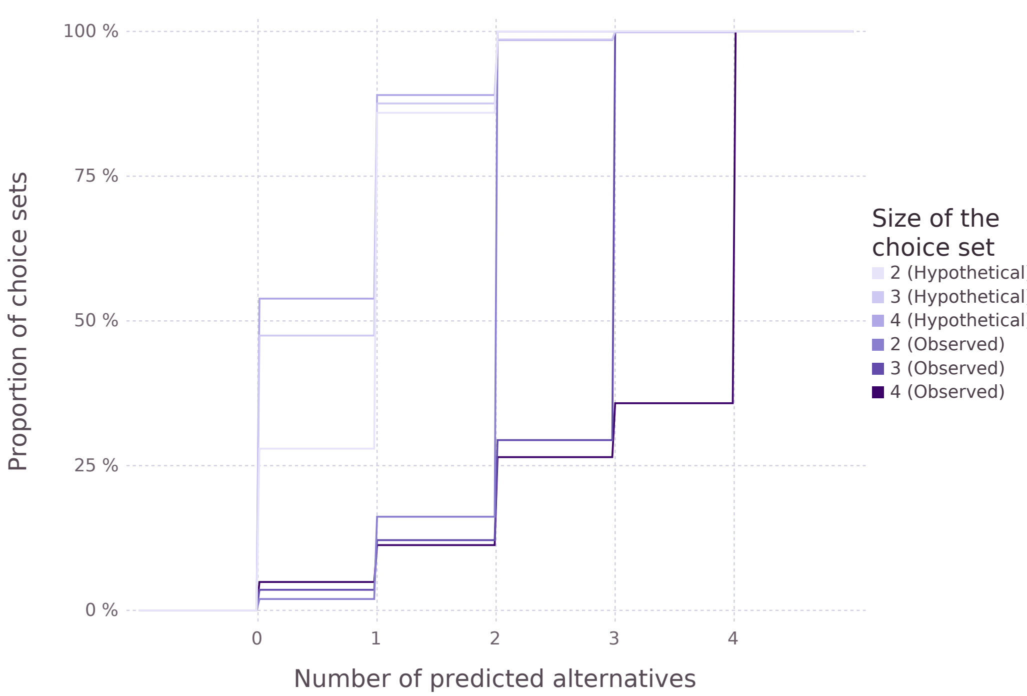


Figure 10: CDF of the number of predicted alternatives per choice set for limited attention in the experimental data.



4.6 Robustness Checks

Finally, we examine how the performance of Strict UCR changes if we vary the composition of the consumption panel in three different ways:

1. Dropping December. Because holiday spending can be considered a "frame", this robustness check can be considered as reducing the set of welfare-relevant frames. This change shrinks the number of observations from 120 to 110.

2. Dropping the last 5 years of data to reduce the impact of time variation in preferences. This change shrinks the number of observations from 120 to 60.

3. Aggregating purchases to the the annual level to reduce the impact of seasonality on purchases. This change shrinks the number of observations from 120 to 10.

Table 11 shows a summary of these robustness checks. On average, the predictive power is higher for all three robustness checks as compared to the full sample, but the increase is not substantial.

Table 11: Comparison of the full sample and robustness samples for the consumption data.

			Robustness samples		
		Full sample	Months 1-11	Years 1-5	Annual
Completeness	Revealed preferences	100.00%	99.97%	100.00%	99.98%
	Strict UCR	99.12%	99.18%	99.17%	99.98%
Acyclicity	Revealed preferences	0%	0%	1.43%	86.81%
	Strict UCR	79.54%	85.63%	97.38%	100%
Predictive power	Num. predicted (UCR)	1.03	1.03	1.03	1.01
	One predicted (UCR)	96.85%	97.08%	97.05%	98.79%

5 Discussion and Conclusion

In this paper, we provide the first non-parametric empirical applications of Strict UCR. The resulting analysis helps to provide an empirical answer to whether substantial model structure is needed to give strong welfare guidance when choices appear inconsistent.

We were also motivated to test Strict UCR non-parametrically because the original choice-theoretic formulation of P^* presented in Bernheim and Rangel (2009) is non-parametric. Related papers in the decision theory literature, such as Chambers and Hayashi (2012), Nishimura (2015), and Apesteguia and Ballester (2015), also provide a non-parametric approach.

For both data sets considered in this paper, we find that Strict UCR is acyclic and has high predictive power for observed demands, which means that it offers clear welfare guidance for those demands. Of course, to provide a more comprehensive and general answer to when Strict UCR is acyclic and has high predictive power, we would need to look at other experimental and non-experimental data sets, such as those examined in the behavioral economics literature. However, neither of our data sets were cherry-picked to produce a desired result. Our prior belief was that Strict UCR would be overly coarse for the data sets we examine in this paper. Instead, our results lead us to conclude something quite different for these data sets.

In addition, we feel that the data sets examined in this paper represent valid test sets because behavioral biases are likely to influence choices made in these settings. For instance, options presented at the top of the list in the experiment are likely to be picked more often. Also, in grocery store purchases, consumers may be drawn to a product due to its position in the aisle or special display case. Alternatively, individuals may be tempted to buy products that they do not actually want because they are hungry.

Finally, while the original formulation of Strict UCR was given for choices from menus, we felt it was important to examine the performance of Strict UCR for choices from budget sets. Not only are choices from budget sets a canonical revealed preference data set, but in many applications there are prices associated with goods, so to ignore budget set data is to ignore many real world settings.

References

Adams, Abi and Ian Crawford (2015). "Models of Revealed Preference". In: *Emerging Trends in the Social and Behavioral Sciences*. Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 1–15. ISBN: 9781118900772. DOI: 10.1002/9781118900772.etrds0227. URL: http://doi.wiley.com/10.1002/9781118900772.etrds0227.

Aguiar, Mark and Erik Hurst (2007). "Life-Cycle Prices and Production". In: *American Economic Review* 97.5, pp. 1533–1559. ISSN: 0002-8282. DOI: 10.1257/aer.97.5.1533.

Ambuehl, Sandro, B. Douglas Bernheim, and Annamaria Lusardi (2015). "Financial Education, Financial Competence, and Consumer Welfare". In: *NBER Working Paper Series*.

Andreoni, James, Ben Gillen, and William T. Harbaugh (2013). "The power of revealed preference tests: ex-post evaluation of experimental design". In: *Unpublished manuscript*. URL: http://econweb.ucsd.edu/%7B~%7Djandreon/WorkingPapers/GARPPower.pdf.

Apesteguia, José and Miguel A. Ballester (2015). "A Measure of Rationality and Welfare". In: 123.6, pp. 1278–1310.

Beatty, Timothy K. M. and Ian Crawford (2011). "How Demanding Is the Revealed Preference Approach to Demand?" In: *The American Economic Review* 101.6, pp. 2782–2795. ISSN: 0002-8282. URL: http://www.jstor.org/stable/23045658.

Benkert, Jean-Michel and Nick Netzer (2016). *Informational Requirements of Nudging*.

Bernheim, B. Douglas (2009). "Behavioral Welfare Economics". In: *Journal of the European Economic Association* 7.2/3, pp. 267–319. ISSN: 1542-4766. URL: http://www.jstor.org/stable/40282753.

— (2016). "The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics". In: *Journal of Benefit-Cost Analysis* 7.01, pp. 12–68. ISSN: 2194-5888. DOI: 10.1017/bca.2016.5.

Bernheim, B. Douglas, Andrey Fradkin, and Igor Popov (2015). "The Welfare Economics of Default Options in 401(k) Plans". en. In: *American Economic Review* 105.9, pp. 2798–2837. ISSN: 0002-8282. DOI: 10.1257/aer.20130907.

Bernheim, B. Douglas and Antonio Rangel (2009). "Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics". In: *The Quarterly Journal of Economics* 124.1, pp. 51–104.

Blundell, Richard W., Martin Browning, and Ian Crawford (2003). "Nonparametric Engel Curves and Revealed Preference". In: *Econometrica* 71.1, pp. 205–240. ISSN: 0012-9682. URL: http://www.jstor.org/stable/3082045.

Boccardi, Maria Jose (2017). "Predictive Ability and the Fit-Power Trade-Off in Theories of Consumer Behavior". In: pp. 1–34. URL: https://drive.google.com/file/d/0B-SXcE9wvakcUmNmdFY5Wm54QU0/view.

Caplin, Andrew (2016). "Economic Data Engineering". In: May, pp. 1–44.

Caplin, Andrew, Mark Dean, and Daniel Martin (2011). “Search and satisficing”. In: *The American Economic Review* 101.7, pp. 2899–2922. URL: http://www.jstor.org/stable/pdfplus/41408725.pdf.

Caplin, Andrew and Daniel Martin (2012). “Framing effects and optimization”. In: URL: http://www.econ.as.nyu.edu/docs/IO/23321/Martin%7B%5C_%7D04242012%7B%5C_%7D01.pdf.

Chambers, Christopher P. and Takashi Hayashi (2012). “Choice and individual welfare”. en. In: *Journal of Economic Theory* 147.5, pp. 1818–1849. ISSN: 00220531. DOI: 10.1016/j.jet.2012.01.013.

Choi, Syngjoo, Raymond Fisman, Douglas Gale, and Shachar Kariv (2007). In: *The American Economic Review* 97.5, pp. 1921–1938. ISSN: 0002-8282. URL: http://www.jstor.org/stable/30034591.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman (2014). “Who Is (More) Rational ?” In: *American Economic Review* 104.6, pp. 1518–1550.

Crawford, Ian and Bram De Rock (2014). “Empirical revealed preference”. In: *Annu. Rev. Econ.* 6.1, pp. 503–524. URL: http://www.annualreviews.org/doi/abs/10.1146/annurev-economics-080213-041238.

De Clippel, Geoffroy and Kareen Rozen (2014). “Bounded rationality and limited datasets”. In: *mimeo*. URL: http://www.econ.brown.edu/fac/Geoffroy_declippel/GK_BRLimitedData.pdf.

Dean, Mark and Daniel Martin (2016). “Measuring Rationality with the Minimum Cost of Revealed Preference Violations”. In: *Review of Economics and Statistics* 98.3, pp. 524–534. ISSN: 0034-6535. DOI: 10.1162/REST_a_00542.

Frederick, Shane, George Loewenstein, and Ted O’Donoghue (2002). “Time Discounting and Time Preference: A Critical Review”. In: *Journal of Economic Literature* 40.2, pp. 351–401. ISSN: 0022-0515. DOI: 10.1257/002205102320161311.

Handbury, Jessie, David E Weinstein, and Tsutomu Watanabe (2015). “How Much Do Official Price Indexes Tell Us About Inflation?” In: *NBER Working Papers* 24223003. DOI: 10.3386/w19504.

Hinnosaar, Marit (2016). “Time Inconsistency and Alcohol Sales Restrictions”. In: 2, pp. 1–46. ISSN: 00142921. DOI: 10.1016/j.euroecorev.2016.04.012.

Johnson, Donald B. (1975). “Finding all the elementary circuits of a directed graph”. In: *SIAM Journal on Computing* 4.1, pp. 77–84. URL: http://epubs.siam.org/doi/abs/10.1137/0204007.

Koo, Anthony YC (1963). “An empirical test of revealed preference theory”. In: *Econometrica: Journal of the Econometric Society*, pp. 646–664.

Manzini, Paola and Marco Mariotti (2010). “Revealed preferences and boundedly rational choice procedures: an experiment”. In: *University of St. Andrews and IZA Working Paper*, pp. 141–167. URL: http://www.st-andrews.ac.uk/-pm210/similEXPERIMENT.pdf.

Manzini, Paola and Marco Mariotti (2012). “Categorize Then Choose: Boundedly Rational Choice and Welfare”. en. In: *Journal of the European Economic Association* 10.5, pp. 1141–1165. ISSN: 1542-4774. DOI: 10.1111/j.1542-4774.2012.01078.x.

— (2014). “Welfare economics and bounded rationality: the case for model-based approaches”. In: *Journal of Economic Methodology* 21.4, pp. 343–360. ISSN: 1350-178X. DOI: 10.1080/1350178X.2014.965909.

Masatlioglu, Yusufcan, Daisuke Nakajima, and Erkut Y. Ozbay (2012). “Revealed Attention”. In: *American Economic Review* 102.5, pp. 2183–2205. ISSN: 0002-8282. DOI: 10.1257/aer.102.5.2183.

Nishimura, Hiroki (2015). “The transitive core : inference of welfare from nontransitive preference relations”. In: *Theoretical Ecomonics*, pp. 1–44. URL: http://www.hirokinishimura.net/files/TransitiveCore.pdf.

Rubinstein, Ariel and Yuval Salant (2012). “Eliciting welfare preferences from behavioural data sets”. In: *The Review of Economic Studies* 79.1, pp. 375–387.

Salant, Yuval and Ariel Rubinstein (2008). “(A, f): Choice with Frames”. In: *The Review of Economic Studies* 75.4, pp. 1287–1296.

Tversky, Amos and Daniel Kahneman (1981). “The framing of decisions and the psychology of choice”. In: *Science* 211, pp. 453–458.

Varian, Hal R. (1982). “The Nonparametric Approach to Demand Analysis”. In: *Econometrica* 50.4, p. 945. ISSN: 00129682. DOI: 10.2307/1912771.

— (2006). “Revealed preference”. In: *Samuelsonian economics and the twenty-first century*, pp. 99–115.

6 Appendix

6.1 Determination of Cycles

The most time-consuming step in determining preference cycles is in transforming observed choices into preference relation elements. Once we have completed that step, we use tools of graph theory for the rest of the analysis. Cycles are determined using the canonical graph theory algorithm of Johnson (1975).

Technically, all of the programs have been written using the software language Julia (versions 0.4.x and 0.5.x). The libraries used are DataFrames (versions 0.6, 0.7 and 0.8) to transform the original data, LightGraphs (versions 0.5, 0.6 and 0.7) to build the directed graph, Distributions (versions 0.10, 0.11 and 0.12) for the random draws, HypothesisTests (versions 0.x) for hypothesis testing, and Gadfly (versions 0.3, 0.4 and 0.5) and Colors (0.7) for producing plots. The graph algorithm is a custom implementation available here.

6.2 Random Choice Methodology

All explanations are given at the individual level. Our primitives are quantities (for each individual and each period) and prices (for each market and each period), which provide the actual expenditure for each individual.

1. We normalize prices by the expenditures for each individual, so that expenditure is 1 for each period and each individual.
2. We build vectors of quantities that constitute our random bundles, each one built this way:
    (a) For each good in the bundle, we count the number of times it was not purchased and divide this count by the number of periods, so that we get the frequency of periods where this good has been purchased. This give us a vector of 6 parameters for 6 Bernoulli distributions (one for each good). This step is done in order to mimic the real behavior of the panelists, who do not purchase all goods in all periods.
    (b) For each good, we determine whether it has been purchased or not at this period. If not, the quantity is zero. If yes, it is part of the next step.
    (c) For all goods purchased, we draw random budget shares using the Dirichlet distribution on the simplex (which allows us to get a total expenditure of 1).[35] We then

[35] The Dirichlet distribution with parameters of 1 is the proper distribution to perform uniform draws on the simplex (see Wikipedia).

transform the budget shares into quantities, for comparison with other periods.

(d) We then build the preference relation from these comparisons. We therefore have a preference relation on a set of alternatives (120 of them in our main analysis).

(e) To assess predictive power, we draw for each demand (observed or hypothetical) 1,000 set sizes from a uniform distribution over set sizes, draw random subsets of alternatives of the given set sizes, and then assess the predictive power on each of these sets.

6.2.1 Analysis Sample: Demographic Characteristics

All of the subjects who participated in the experiments of Manzini and Mariotti (2010) were Italian university students. On the other hand, the panelists in our consumption data are residents of the US, older, and largely working full-time or close to full-time.

For the analysis sample of our consumption data, the median age in 2004 is 56 years, and the youngest panelist in 2004 is 30 years old. Among individuals in the US who were 30 years old and above in 2004, the median age is 50.[36]

As shown in table 12, a majority of individuals in our analysis sample are working, and a plurality work more than 35 hours per week. There is, however, a substantial fraction that are not employed (42.76% on average over the 10 years), and this rate is higher than for individuals in the US who were 30 years old and above in 2004 (37.23%). This stems from a sample skewed towards people already retired. While we have excluded individuals that experience a change from employment to retirement, we have not removed those who are retired or inactive throughout the 10 year period.

Table 12: Average hours worked per week.

	< 30 hours	30-35 hours	> 35 hours	Not employed
Analysis sample over 10 years	9.52 %	3.55%	44.17%	42.76%
30+ year olds in US (2004)	10.72%	4.81%	47.23%	37.23%

Source: Table 19 of the CPS Labor Force survey.
http://www.bls.gov/cps/cps_aa2013.htm.

The median income of the analysis sample is between \$35,000 and \$40,000, which is slightly lower than the median income of individuals in the US who were 30 years old and above in 2004, as shown in table 13. The level of education of our sample is slightly higher than this group, as table 14 shows.

[36] Data from the Current Population Survey (CPS) for 2004. http://www.census.gov/population/age/data/2004comp.html.

Table 13: Income quartiles.

Percentile	25th	50th	75th
Analysis sample over 10 years	$22,500	$37,500	$65,000
30+ year olds in US (2004)	$26,250	$38,750	$56,250

Source: Annual Social and Economic (ASEC) Supplement of the CPS. http://www2.census.gov/programs-surveys/cps/tables/pinc-03/2005/new03_010.txt.
Note: The original data has income brackets, so the midpoint is used.

Table 14: Level of education.

Education	College degree	No college degree
Analysis sample over 10 years	46.50%	53.50%
30+ year olds in US (2004)	43.25%	56.75%

Source: Annual Social and Economic (ASEC) Supplement of the CPS. http://www2.census.gov/programs-surveys/cps/tables/pinc-03/2005/new03_010.txt.
Note: The degree considered is the highest received, so some individuals in the "no college" category might have been to college, but did not get their degree.

In the experiments of Manzini and Mariotti (2010), the subjects were a roughly even mix of men and women (see footnote 9 of Manzini and Mariotti 2010). In the analysis sample of our consumption data, 733 out of the 1,183 panelists are women, a proportion of 61.96%. In the US population, the fraction of women among individuals aged 30 and older was 52.34% in 2004.[37]

[37] US Census Bureau, CPS survey, Annual Social and Economic Supplement, 2004.

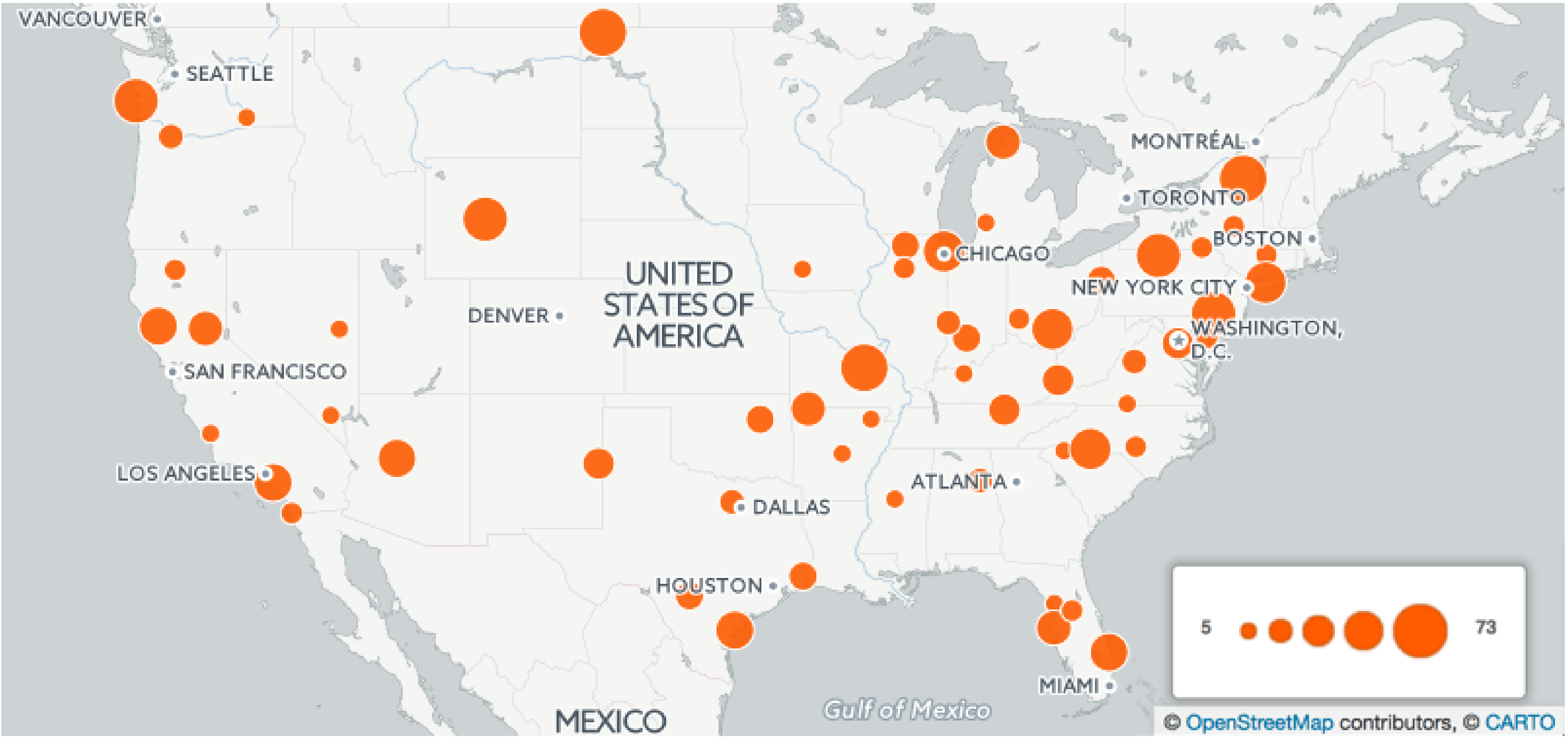


Figure 11: Individuals in the consumption data by market. The size of a bubble is proportional to the number of individuals in a given market.