Predictive Power in Behavioral Welfare Economics

Elias Bouacida, Daniel Martin

Predictive Power in Behavioral Welfare Economics

Elias Bouacida* and Daniel Martin†

May 18, 2019

Abstract

When choices are inconsistent due to behavioral biases, there is a theoretical debate about whether it is necessary to impose the structure of a model in order to provide precise welfare guidance based on those choices. To address this question empirically, we use standard data sets from the lab and field to evaluate the predictive power of two conservative "model-free" approaches to behavioral welfare analysis. We find that for most individuals, these approaches have high predictive power, which means there is little ambiguity about what should be selected from each choice set. We show that the predictive power of these approaches correlates highly with two properties of revealed preferences: the number of direct revealed preference cycles and the fraction of revealed preference cycles that are direct.

JEL Codes: I30, C91, D12

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

*Paris School of Economics, Université Paris 1 Panthéon Sorbonne, 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 choice inconsistencies identified in the behavioral economics literature, several recent papers have demonstrated widespread choice inconsistencies in standard data sets – both experimental (e.g., Choi, Fisman, Gale, and Kariv 2007; Choi, Kariv, Müller, and Silverman 2014) and observational (e.g., Blundell, Browning, and Crawford 2003; Dean and Martin 2016). Because inconsistencies generate reversals in the preferences revealed by choice, this means that individuals cannot be modeled *as if* they maximize a single, stable utility function in many standard 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. An alternative choice-based solution is to generate a relation from choices without imposing much *ad hoc* model structure and to use that relation to conduct welfare analysis (e.g., Bernheim and Rangel 2009; Chambers and Hayashi 2012; Apesteguia and Ballester 2015; Nishimura 2018).

Given the nature of this divide, a theoretical debate has emerged as to how much model structure is necessary to provide precise welfare guidance from inconsistent choices (Bernheim 2009; Rubinstein and Salant 2012; Manzini and Mariotti 2014; Bernheim 2016). It has been argued that "model-free" behavioral welfare approaches that are conservative in how they resolve the normative ambiguities produced by choice inconsistencies will have little to say about welfare in practice. While there are other normative criteria for policymakers besides the precision of welfare guidance, if an approach has little to say about welfare, then other considerations are likely to be moot.

We offer empirical evidence for this theoretical debate by determining, for standard data sets from the lab and field, the precision of welfare guidance offered by two behavioral welfare relations: the strict unambiguous choice relation (SUCR henceforth) proposed by Bernheim and Rangel (2009) and the transitive core (TC henceforth) proposed by Nishimura (2018). Both SUCR and TC are "conservative" in the sense that they do not attempt to resolve all of the normative ambiguities produced by choice inconsistencies, so are likely to be incomplete in their welfare guidance when behavioral biases impact choice.

Both of these behavioral welfare relations provide a loosening of revealed preferences by overlooking some inconsistencies in choice. The standard approach is to say that x is revealed preferred to y (denoted as xPy) if x is chosen when y is available and not chosen. However, this relation is unsuitable for welfare analysis if it contains a cycle: if there exists $x_1, x_2, \ldots, x_n$ such that $x_1 P x_2, \ldots, x_n P x_1$.

SUCR and TC aim to be free of such cycles by excluding revealed preferences relation elements that produce cycles. SUCR retains a relation between x and y when x is strictly unambiguously chosen over y (denoted as xP^*y) which holds if and only if y is never chosen when x and y are available.[1] Alternatively, TC retains a relation element between x and y if all pairwise comparisons with an option z are consistent with the ordering implied by that relation element.

We evaluate whether these behavioral welfare relations offer precise welfare guidance by determining their "predictive power", which is their ability to make sharp predictions.[2] When a theory does not offer unique predictions, predictive power indicates how loose or tight its predictions are. Because SUCR and TC can be incomplete, they do not always pin down what an agent would select from a set of options, so their predictive power is in question. As an example, imagine the choices of $\{x\}$ from $\{x, y\}$,$\{x\}$ from $\{x, y, z\}$, $\{x\}$ from $\{x, a\}$, and $\{a\}$ from $\{x, y, a\}$. From these choices, SUCR says that xP^*y, xP^*z, and aP^*y. For the choice set $\{x, y, z\}$, SUCR predicts that just x should be selected. On the other hand, for other choice sets, such as $\{x, a\}$, SUCR predicts that any option could be selected.

Predictive power is a useful way to evaluate the precision of welfare guidance because the predictions of a relation correspond to what is welfare optimal for that relation. For instance, if a welfare relation predicts that just one option could be selected from a choice set, then it has both maximal predictive power and offers the most precise welfare guidance. However, if a welfare relation predicts that any option could be selected from a choice set, then it has minimal predictive power and offers no welfare guidance. In the previous example, SUCR offers very precise welfare guidance from $\{x, y, z\}$ as the individual welfare optimum for that choice set is x, but it offers no welfare guidance from $\{x, a\}$.

One natural measure of predictive power is the number of options that are predicted to be chosen, where the highest possible predictive power corresponds to a value of 1 (a single option), and larger values represent less predictive power. However, we use Selten's index (Selten 1991) as our primary measure of predictive power instead because it has a theoretical

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

[2] For other applications of predictive power in empirical revealed preference analysis, see Manzini and Mariotti (2010), Beatty and Crawford (2011), Andreoni, Gillen, and Harbaugh (2013), Dean and Martin (2016), and Boccardi (2017).

grounding, has been used for related questions in the literature, and accounts for the number of available choice options. This index, axiomatized in Selten (1991), is designed explicitly for theories that predict a subset of possible outcomes. With Selten's index, the proportion of choices that a theory predicts successfully within-sample is reduced by the "size of the area", which is how many outcomes are consistent with a theory. We calculate the size of the area by determining the fraction of options in a choice set that are predicted to be chosen.[3]

We test SUCR and TC'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).[4] The latter is composed of choices from budget sets for 1,190 single-person households over 10 years, which comes from Nielsen's National Consumer Panel (NCP) – formerly known as the Homescan Consumer Panel.[5]

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, 53% of individuals make choices that generate revealed preference cycles, and for the consumption data, 100% of individuals exhibit revealed preference cycles.[6] 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 (an assumption about data made by Bernheim and Rangel 2009), and the consumption data contains a large number of individuals and observations per individual. Fourth, they are quite different from each other in terms of individual demographic characteristics, choice settings, and choice alternatives.

For both data sets, we find that SUCR and TC have a high level of predictive power. In the experimental data, the average number of predicted options is 1.32 for SUCR and 1.38 for TC, and in the consumption data, it is 1.33 for SUCR and 1.65 for TC. In the experimental data, the average value of Selten's index (which can range from 0 to 0.58 in this application) is 0.46 for SUCR and 0.44 for TC, and in the consumption data, the average value of Selten's index (which can range from 0 to 0.96 here) is 0.95 for SUCR and 0.94 for

[3]Because SUCR and TC always predict successfully within-sample, the value of Selten's index in our application is determined entirely by the average size of the area.

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

[5]Researcher(s) own analyses calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researcher(s) and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

[6]We restrict our subsequent analysis to those subjects with cycles in their revealed preferences.

TC.[7]

To learn when and why SUCR and TC have high predictive power, we study two properties of revealed preference (RP) that should correspond with their predictive power: the number of direct RP cycles and the fraction of all RP cycles that are direct. For the experimental data, xPy if x is chosen from a choice set that contains x and y, and for the consumption data, xPy if x is chosen from a budget set that contains x and y.[8] We say there is an RP cycle if there exists $x_1, x_2, \ldots, x_n$ such that $x_1 P x_2, \ldots, x_n P x_1$ and that an RP cycle is "direct" if xPy and yPx, that is, if it has a length of 2.[9]

It is worth noting that while cycle length is typically not considered in revealed preference analysis beyond distinguishing between violations of the Weak Axiom of Revealed Preference (WARP) and the Strong Axiom of Revealed Preference (SARP),[10] cycle length plays a critical role in this application.[11] For example, in determining the predictive power of SUCR and TC, the number of direct RP cycles is more important than the number of RP cycles of any other length or even the total number of RP cycles. The number of direct RP cycles is very informative about predictive power because SUCR and TC do not contain any relation elements that generate direct RP cycles, but can contain relation elements that are a part of longer length RP cycles.

That said, the number of longer length RP cycles does matter, but what matters most is their number in proportion to the number of direct RP cycles. If there are a lot of RP cycles of longer length relative to the number of direct RP cycles, then this is a problem for both SUCR and TC. For one, SUCR will be cyclical (P^* will contain cycles) if revealed preference cycles remain after removing all revealed preference relation elements that are in direct RP cycles, so SUCR is more likely to cyclical if there are many RP cycles of longer length relative to the number of direct RP cycles. In addition, TC excludes additional revealed preference relation elements beyond those that are in direct RP cycles, so its predictive power is likely to be lower if there are many RP cycles of longer length relative to the number direct RP cycles. Thus, for the predictive power of TC, the fraction of all RP cycles that are direct is likely to matter in addition to the number of direct RP cycles.

Because determining the number of RP cycles of longer length can become computation-

[7] Predictive power, as determined by Selten's index, is higher in the consumption data in part because choice sets are larger on average.

[8] We use the "strict" RP relation P because we never observe more than one option selected from a choice set and never observe two selected bundles that could have been purchased at the identical expenditure.

[9] When counting cycles, we avoid double-counting by requiring $x_1, x_2, \ldots, x_n$ to be distinct and assuming that any re-ordering of $x_1, x_2, \ldots, x_n$ is the same cycle.

[10] Direct RP cycles are violations of both WARP and SARP, whereas RP cycles of longer lengths are only violations of SARP.

[11] An exception is V. Aguiar and Serrano (2017), who consider the implications of cycles of different lengths relative to the Slutsky Matrix.

ally burdensome, we generate a bound on the fraction of all RP cycles that are of length 2 by dividing the number RP cycles of length 2 by the sum of all RP cycles of length 2 and 3. We call this measure the "directness index", and to the best of our knowledge, this measure is new to the literature.

As expected, we find that the *number* of direct RP cycles (RP cycles of length 2) is highly and negatively correlated with Selten's index for both SUCR and TC in our data sets. In the experimental data, the correlation for SUCR is -0.98, and for TC is -0.92, and in the consumption data, the correlation for SUCR is -0.75, and for TC is -0.81. We also find that the *fraction* of revealed preference cycles that are direct is highly and positively correlated with Selten's index for TC. In the experimental data, the correlation for TC is 0.78, and in the consumption data, the correlation for TC is 0.69. These relationships are also strong, positive, and significant in regressions that control for the number of direct RP cycles. In fact, as measured by R^2, variation in these two factors explains 89% of the variation in Selten's index for TC in the experimental data and 67% in the consumption data.[12]

In addition, we find that the *fraction* of direct RP cycles is strongly and positively correlated with the acyclicity of SUCR. In the consumption data, 79% of individuals have acyclic SUCR, and the correlation is 0.38 between the directness index and a dummy variable that takes a value of 1 when SUCR is acyclic.

In addition to offering an answer to when and why SUCR and TC perform well in practice, both measures (the number of direct RP cycles and the directness index) are relatively quick to calculate. When approaching a new data set, it is easy to assess whether conservative model-free approaches to behavioral welfare economics are likely to offer precise welfare guidance.

To the best of our knowledge, this paper presents the first non-parametric empirical application of SUCR and the first empirical application of TC.[13] In addition, we introduce predictive power as a tool for evaluating behavioral welfare relations. Based on a standard measure of predictive power, we help to provide an answer to the question of how much model structure is necessary to provide precise welfare guidance. For the standard choice data sets we consider, it appears that one can give precise 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 SUCR, TC, and alternative welfare relations. In Section 3, we describe the two data sets. In Section 4, we provide results for both data sets. We conclude with a brief discussion in Section 5.

[12] We do not have an *a priori* reason to believe that the directness index should matter for the predictive power of SUCR when controlling for direct RP cycles.

[13] As discussed in section 2, there are existing parametric empirical applications of SUCR.

2 Behavioral 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).[14] If a set of choices generate no RP cycles, then the revealed preferences that correspond to those choices are suitable for conducting welfare analysis. However, if the choices generate RP cycles, the decision maker can no longer be modeled as a maximizer of a standard utility function, 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 choice itself, an ancillary condition or frame present when the choice is made.[15] They assume that while a frame can impact the choice, it does not affect the alternatives themselves or the welfare derived from them.[16]

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 we wish to test predictive power with as conservative an approach as possible, we do not make any assumptions about which choices are welfare-relevant.[17]

2.2 SUCR, TC, and Revealed Preferences

Bernheim and Rangel (2009) define the following relation: x is (strictly) unambiguously chosen over y (denoted xP^*y) if whenever x and y are both available in some welfare-relevant frame, y is never chosen. Whenever choices are observed from all possible subsets of choice options (a condition we will call "full observability"), SUCR is guaranteed to be

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

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

[16] Caplin and Martin (2018), Rubinstein and Salant (2012), and Benkert and Netzer (2018) also provide ways to use frames when assessing welfare.

[17] The potential downsides of imposing welfare-relevance in an *ad hoc* manner are discussed in Gul and Pesendorfer (2009).

acyclic, which is a useful property for performing welfare analyses.[18]

A fundamental difference between SUCR and RP is that with SUCR, multiple observations are considered jointly when inferring the ranking, whereas, with RP, each observation is taken independently to do so. However, there is an important link between the two: SUCR does not include any RP relation elements that are involved in direct RP cycles. In other words, if xPy and yPx, then it is not possible that xP^*y or yP^*x because those revealed preference relation elements only exist if x is chosen when y is available and vice versa.

Unlike SUCR, TC is generated from another relation, which we take to be the (strict) revealed preference relation P. A relation element xPy is in the transitive core of P (denoted $xc(P)y$) if for all options z, zPx implies zPy and yPz implies xPz. Nishimura (2018) shows that TC makes recommendations that do not rely on arbitrary decisions from a modeler, so like SUCR, it does not use a model to resolve normative ambiguities. Also, like SUCR, TC does not include any RP relation elements involved in a direct RP cycle. For instance, if xPy and yPx, then it cannot be that $xc(P)y$ because for option y, yPx would imply yPy.

Despite these similarities, SUCR and TC can differ in their welfare guidance. Nishimura (2018) presents theoretical examples where SUCR is different from TC, specifically for models of time preferences with relative discounting and regret preferences. We find that on average TC is coarser than SUCR in our data sets. However, TC has a substantial advantage over SUCR for empirical applications. Whenever choices are observed from all binary sets of choice options, TC is guaranteed to be acyclic, but SUCR is not. Even though this condition does not hold for our consumption data, we find that TC is always acyclic – even when SUCR is not acyclic for the same individuals (although, as noted by Nishimura 2018, TC is not necessarily acyclic when the underlying relation is incomplete).

2.3 Other Welfare Relations

Other behavioral welfare relations have been proposed in the literature that impose little *ad hoc* model structure. 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 relations are found by choosing the complete linear order that is closest to (empirically) observed choices. To assess the closeness of an order, they look at the number of alternatives that must be ignored in each choice set to match the preferences implied by choices to the candidate order. This approach uses choice set frequencies to overcome ambiguities, so is less conservative than SUCR and TC in making welfare assessments. In fact,

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

because the set of welfare relations is complete, they always have full predictive power.

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 SUCR to choice data.[19] They study the impact of making one retirement savings option the default, and because individuals appear to make inconsistent choices as the default option changes, they use SUCR 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 a 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 (2018). However, Apesteguia and Ballester (2015) formally prove that the mass of data sets 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 the minimal cost to make a revealed preference relation acyclic is relatively small. However, there are three ways in which the high predictive power of SUCR and TC for our consumption data is surprising, even in light of their findings. First, Dean and Martin (2016) consider the minimal cost to make a revealed preference relation acyclic, whereas SUCR and TC remove all ambiguous comparisons, which is in general

[19] An application of concepts from Bernheim and Rangel (2009) also appears in Ambuehl, Bernheim, and Lusardi (2014).

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 revealed preference relation elements need to be removed from a relation to make it acyclic, there is no guarantee that such a relation will have high predictive power.

3 Data

We use two very different data sets for our non-parametric applications of SUCR. 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 individual demographic characteristics (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 several 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, M. 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 other words, subjects were asked when and how they would like to receive monetary payments.

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). These plans indicated how the size of their monetary payments would change over time. 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).

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

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.[20] Data with the property of full observability are appealing for two reasons. First, SUCR and TC are guaranteed to be acyclic for such data. Second, such data provide a stringent test of the predictive power of SUCR and TC.

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.[21]

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. There are a growing number of papers that analyze NCP data.[22]

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 10 years. The length of this panel means that we have many observations, which allows us to perform a stringent test of the predictive power of SUCR and TC. As mentioned previously, this panel contains several times more choices than existing papers that implement revealed preference tests on consumption data (e.g., Blundell, Browning, and Crawford 2003; Dean and Martin 2016).

[20] Presenting all possible subsets is combinatorially challenging. For n alternatives, the number of choice sets is $2^n - n - 1$. To the best of our knowledge, only two recent choice experiments present all possible subsets to decision makers: Gerasimou, Costa-Gomes, Cueva, and Tejiscak (2019) and Bouacida (2019).

[21] 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.

[22] As of May 2019, 132 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.

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 1,190 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 retire.

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.[23] 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 retire because M. Aguiar and Hurst (2007) find that retirement influences consumption patterns.

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 $427.27 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,

[23] For an assessment of the representativeness of our sample, see Section 6.1 of the appendix.

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%

dairy products, deli foods, dry groceries,[24] 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, Watanabe, and Weinstein (2013) study inflation with price indices built similarly.

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 7 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 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.[25]

[24] 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.

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

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).[26]

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 SUCR and TC, which would make the test of predictive power even tougher.

[26] However, this does impose some model structure, which is another reason SUCR and TC are not entirely "model-free" in our application.

4 Results

In this section, we first determine the proportion of individuals who have choices that exhibit revealed preference cycles. For such individuals, we then determine the proportion that has cycles in SUCR and TC, the completeness and predictive power of these relations, and the properties of revealed preferences that are correlated with the predictive power of these relations.

4.1 Inconsistencies in Revealed Preferences

As discussed previously, a standard marker for choice inconsistency is the presence of cycles in the preferences revealed by choice. We say that x is (strictly) revealed preferred to y (denoted xPy) if x is chosen when y is available and that there is an RP cycle if there exists $x_1, x_2, \ldots, x_n$ such that $x_1 P x_2, \ldots, x_n P x_1$. If such a cycle exists, then the choices that generated it cannot be rationalized with a single, stable utility function, which is the standard tool for conducting welfare analysis.

For both of our data sets, a majority of individuals have choices that generate at least one RP cycle, as shown in Table 4. In the experimental data, 53% of individuals have RP cycles for at least one installment plan. In the consumption data, all 1,190 individuals have RP cycles.

The wide breadth of RP cycles we observe is consistent with findings in the empirical literature on revealed preference testing. In the laboratory experiments of Choi, Fisman, Gale, and Kariv (2007), around 35% of subjects have RP cycles for choices from allocations over risky assets, and in the large-scale field experiment of Choi, Kariv, Müller, and Silverman (2014), around 90% of subjects exhibit RP for a similar choice task.

For consumption data, there is a long history of papers that detect RP 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." This prediction has held for subsequent studies, including a paper by Dean and Martin (2016) which finds that around 71% of households exhibit RP cycles in a two-year balanced panel of grocery purchases.

4.2 Inconsistencies in SUCR and TC

SUCR and TC are designed to produce welfare guidance that is free of cycles. However, they are only guaranteed to be acyclic under certain conditions. As mentioned previously,

Table 4: Percent of individuals that have cycles in RP, SUCR, and TC.

		Percent with cycles		
Data	Number of individuals	RP	SUCR	TC
Experimental	102	53%	0%	0%
Consumption	1,190	100%	21%	0%

a requirement that guarantees SUCR will be free of cycles is full observability (choices from all subsets of alternatives). TC is certain to be free of cycles when choices from all binary choice sets are observed, which is a weaker condition. However, this condition is sufficiently strong to ensure that the underlying revealed preference relation will be complete.

The experimental design meets both conditions, so SUCR and TC are certain to be acyclic in the experimental data. On the other hand, neither condition is satisfied wtih the consumption data, so their acyclicity is in doubt in this data set. As shown in Table 4, SUCR has cycles for 21% of individuals in the consumption data. This represents a substantial reduction from the 100% of individuals who have RP cycles in that data set. TC achieves an even larger reduction: even though acyclicity is not guaranteed for TC, it never contains cycles in this data set.

Because SUCR and TC are only needed for welfare guidance when individuals have cyclic revealed preference relations, we only consider individuals that have RP cycles in the remaining analyses. This does not restrict our consumption data at all, but it means that we keep only 53% of experimental subjects, which leaves a total of 54 subjects in our analysis sample.

As a baseline, we keep those individuals in the consumption data who have SUCR cycles. However, because these cycles could potentially distort our assessment of the completeness and predictive power of SUCR, we also provide results without those individuals as a robustness check.

4.3 Completeness of SUCR and TC

A relation $\succeq$ is "complete" if for all x and y in the grand set of alternatives X, either $x \succeq y$ or $y \succeq x$. One of the ways to measure the completeness of a relation is to measure the number of relation elements it contains, and this measure can be normalized by dividing it by the number of relations in a complete and acyclic relation.[27] If a relation is complete and

[27]It is well-known that a complete and transitive relation is acyclic, so that we could call this a complete and transitive relation instead.

does not contain cycles, then there will be $\frac{|X|(|X|-1)}{2}$ relation elements.[28]

Because SUCR excludes all relation elements that are a part of direct RP cycles, Rubinstein and Salant (2012) and Manzini and Mariotti (2014) have argued that SUCR has the potential to be quite incomplete. However, on average, we find that SUCR and TC are far from incomplete in our data sets, as shown in Figure 1.

In the experimental data, because individuals make choices separately from two sets of four options, a relation that is complete and acyclic would have 12 relation elements. For individuals with cyclic RP, the average number of relation elements for SUCR is 9.5, which is 80% of the comparisons in a complete and acyclic relation. For TC, the corresponding figures are 9.1 and 76%.[29] There is heterogeneity in the extent of completeness: none of these subjects have fully complete SUCR and TC relations, but 52% are one relation element short with SUCR and 46% are with TC. An additional 13% are two relations short with SUCR and 9% with TC. Some individuals, however, have a half or less of a complete and acyclic relation (9% of with SUCR and 13% with TC).

In the consumption data, a complete and acyclic relation would have 7,140 elements. To determine this number, we take the grand set of alternatives X to be the set of all bundles that an individual has chosen at some point.[30] In theory, choices in the consumption data also generate revealed preference content about bundles that are never chosen, but such revealed preference content will always be included in SUCR and TC, so including it would inflate our assessment of the completeness of SUCR and TC. To provide a tougher test of the completeness of these relations, we consider only relations over chosen bundles for our consumption data.

In this data, the average number of relation elements for SUCR is 7,076, which is 99.1% of the comparisons in a complete and acyclic relation. The corresponding figures for TC are 7,021 and 98.3%.[31] The maximal number of relation elements are 7,135 for both SUCR and TC, which is very close to the size of a complete and acyclic relation.

There is a sense in which the number of SUCR relation elements are over-counted in data sets without full observability, especially when the relation contains cycles. This may

[28] If a relation contains direct cycles, then this number can be exceeded (up to twice this number). In our analysis, RP will often exceed this number, but SUCR and TC, which do not contain direct cycles, will never exceed it.

[29] The difference between the average number of relations for SUCR and TC is significant (the two-sided paired t-test p-value is 0.0016). For a KS-test of equality of the distributions of SUCR and TC completeness, the p-value is <0.001.

[30] The contents of X can vary individual-by-individual, but its size does not.

[31] The difference between the average number of relations for SUCR and TC is significant (the two-sided paired t-test p-value is <0.001). For a KS-test of equality of the distributions of SUCR and TC completeness, the p-value is <0.001.

be a reason to focus more on the completeness of SUCR in the experimental data, which has full observability or to focus instead on the completeness of TC in the consumption data. Additionally, this is a reason to look at the completeness of SUCR for those individuals with acyclic SUCR. Conditional on SUCR being acyclic, the average number of relation elements for SUCR is 7,084, which is 99.2% of the comparisons in complete and acyclic relation, and for the same subjects, the corresponding figures for TC are 7,043 and 98.6%.[32]

4.4 Predictive Power of SUCR and TC

The completeness of SUCR and TC gives us a sense for the precision of their welfare guidance. However, it does not tell us exactly the precision of their welfare guidance for the observed choice sets, so we also calculate the predictive power of SUCR and TC "within sample" for both data sets.

To determine the predictions made by a relation for the observed choice sets, we follow Schwartz (1976) and Ok (2002) in saying that the choice correspondence C induced by a (possibly incomplete) strict relation $\succ$ is $C_{\succ}(Z) = \{x \in Z | y \succ x \text{ for no } y \in Z\}$.[33] The tightness of the predictions given by C is useful for studying the precision of welfare guidance because what is predicted to be selected from a choice set based on $C_{\succ}$ is what is welfare optimal for that choice set. In the language of Bernheim and Rangel (2009), the elements of $C_{\succ}$ are the "weak individual welfare optimum" of choice set Z.

A natural measure of the predictive power of $C_{\succ}$ for an individual is the average size of $C_{\succ}$ for that individual. For this measure, the highest possible predictive power corresponds to a value of 1 (a single alternative predicted from all choice sets), and larger values represent less predictive power (more alternatives predicted).

In the experimental data, SUCR predicts that an average of 1.32 alternatives could be chosen for individuals with cyclic RP, whereas TC predicts that on average, 1.38 alternatives could be chosen.[34] In the consumption data, the average number of predicted alternatives is 1.33 for SUCR and 1.65 for TC.[35] Figure 2 shows that SUCR makes tighter predictions than TC at the individual level. For a KS-test of equality of the distributions of SUCR and TC for the average number of predicted alternatives at the individual level is 0.0095 on the experimental data and <0.001 on the consumption data.

As our primary measure of predictive power, we use the average value of Selten's index

[32]For subjects with acyclic SUCR, once again the difference between the average number of relations for SUCR and TC is significant (the two-sided paired t-test p-value is <0.001). For a KS test of equality of the distributions of SUCR and TC completeness, the p-value is <0.001.

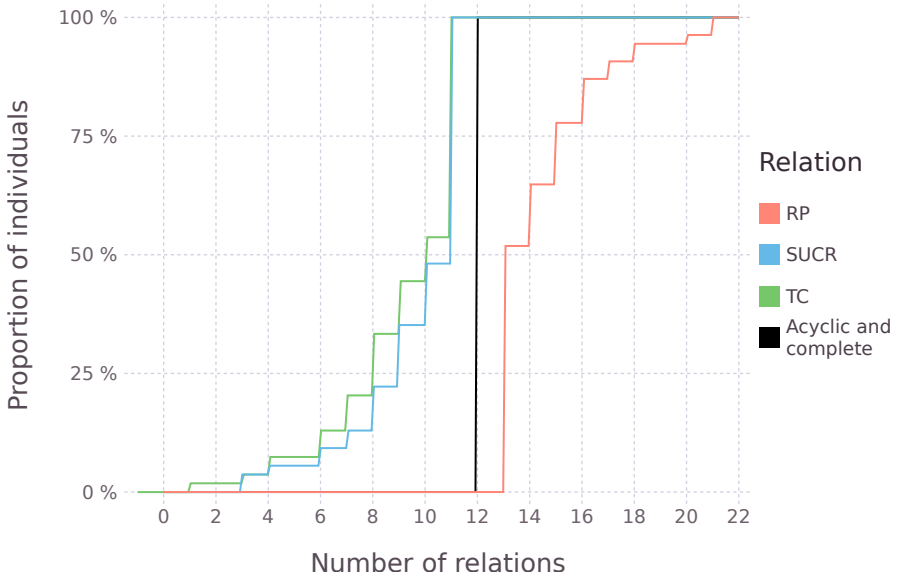
[33]Bernheim and Rangel (2009) propose the same correspondence, which they denote as $m_{\succ}(Z)$.

[34]This difference is significant, as the two-sided paired t-test p-value is 0.0025.

[35]This difference is significant, as the two-sided paired t-test p-value is <0.001.

Figure 1: CDF of the number of relation elements at the individual level (for individuals with cyclic RP).

(a) Experimental data.



(b) Consumption data.

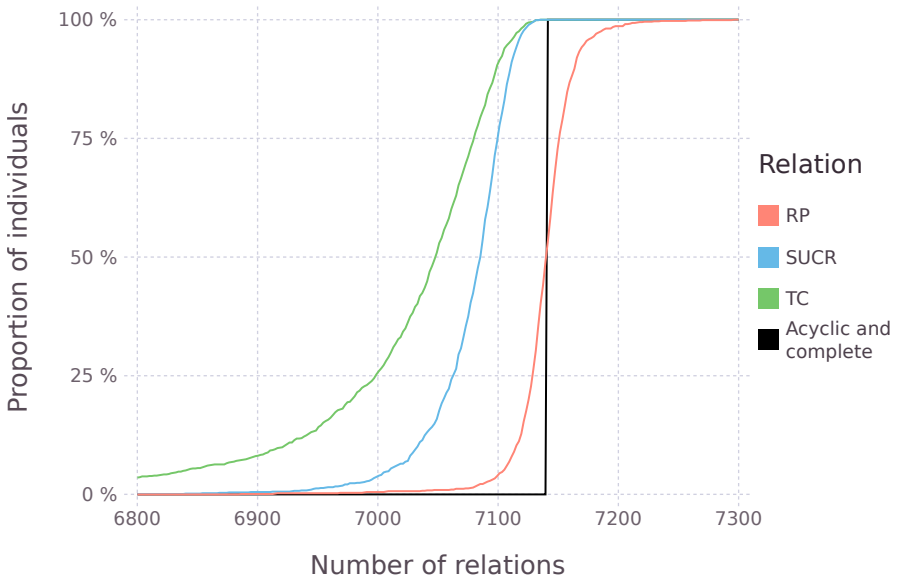
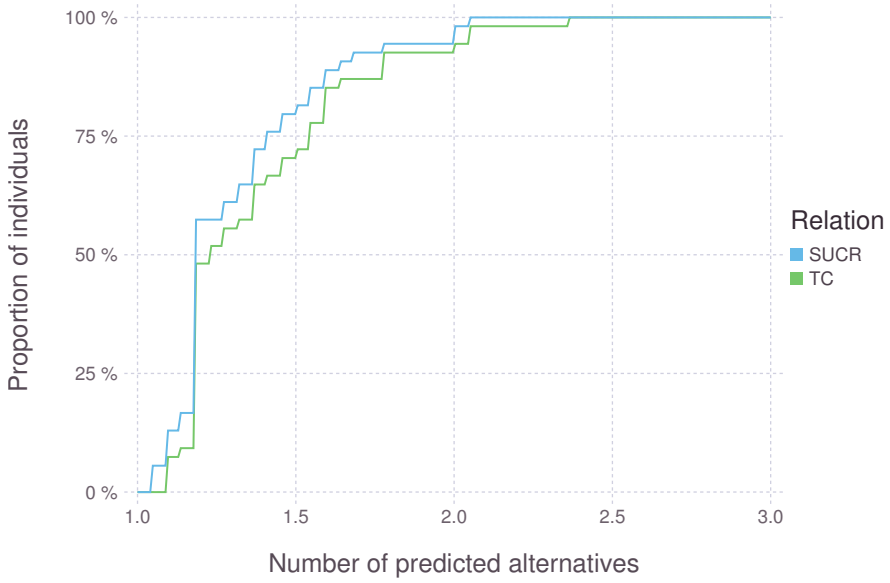
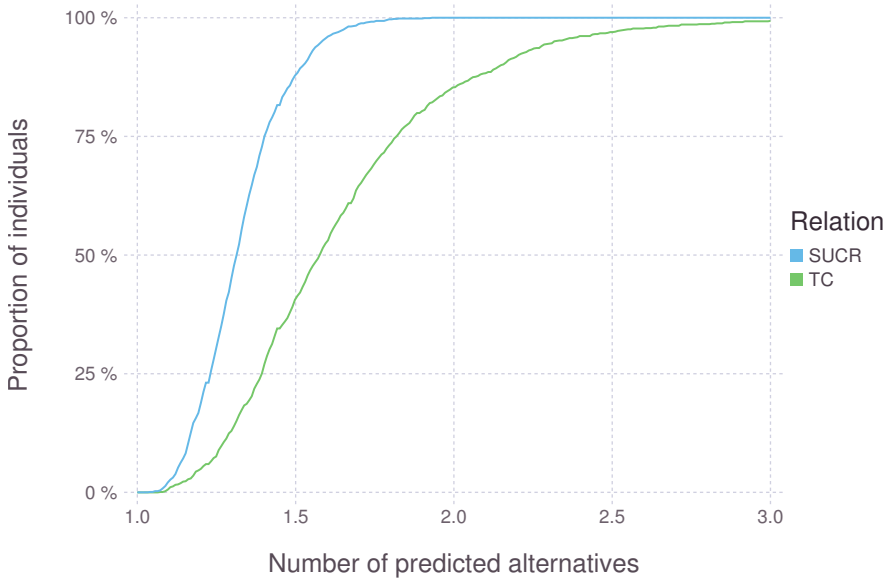


Figure 2: CDF of the average number of predicted alternatives for TC and SUCR at the individual level (for individuals with RP cycles).

(a) Experimental data



(b) Consumption data



(Selten 1991) instead because it has a theoretical grounding, has been used for related questions in the literature, and accounts for the number of available choice options. With Selten's index, the proportion of choices that a theory predicts successfully within-sample is reduced by the "area", which is the relative size of the predicted subset compared with the set of all possible outcomes. In the notation of Selten (1991), it is written as $m = r - a$, where r is the relative frequency of correct predictions and a is the area. Because SUCR and TC always successfully predict the chosen option, r is always equal to 1, but a can vary by relation and choice set. For a relation $\succ$ and choice set Z, we define a as the proportion of alternatives that are predicted to be chosen, so that $a = \frac{|C_{\succ}(Z)|}{|Z|}$.

The choice set and its size are very straightforward to determine in the experimental data. As mentioned previously, in the consumption data we take the grand set of alternatives X to be the set of all bundles that an individual has chosen at some point (in order to provide a tougher test of completeness). Analogously, we take the choice set to be the set of all bundles that an individual has chosen at some point (all bundles in X) that are affordable at a given price and expenditure level.

We could consider the choice set to be every possible bundle on the budget line (as in Beatty and Crawford 2011), but we consider this restricted space for three main reasons. First, it is far more computationally feasible to determine the set of predicted options given the large number of choices in our data set. Second, only bundles chosen elsewhere can generate inconsistencies. Third, this allows us to use the same metric across data sets. Fortunately, this approach provides a wide variety of choice set sizes, as shown in Appendix 6.3.

One reason that we use Selten's index is that it has an axiomatic foundation, and another is that it has been used elsewhere in the literature on empirical revealed preference analysis (Manzini and Mariotti 2010; Beatty and Crawford 2011; Dean and Martin 2016). For example, Beatty and Crawford (2011) determine the fraction 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 the revealed preference test for a given data set. Alternatively, Dean and Martin (2016) determine the average "distance" from rationality for all possible demands and then subtract this from the "distance" for observed choices.

In the experimental data, the average value of Selten's index for SUCR is 0.46 for individuals with cyclic RP, and for TC it is 0.44. The difference between the average Selten's index for TC and SUCR is significant, as the two-sided paired t-test p-value is equal to 0.0022. For the experimental data, the average theoretical maximum of Selten's index is just 0.58, as shown in Table 5. For a set size of 2, the maximum value of Selten's index is 0.5; for a set size of 3, it is 0.66; and for a set size of 4, it is 0.75.

Table 5: Average value of Selten's index by choice set size in the experimental data (for individuals with RP cycles).

	Choice set size			
Relation	2	3	4	Average
SUCR	0.40	0.53	0.60	0.46
TC	0.38	0.50	0.57	0.44
Acyclic and complete	0.50	0.66	0.75	0.58

In the consumption data, the average Selten's index is 0.95 for SUCR and 0.94 for TC. The difference here is significant, as the two-sided paired t-test p-value is <0.001. For the consumption data, the average theoretical maximum is 0.96.[36] Selten's index is higher in the consumption data compared to the experimental data because the size of the choice set appears in the denominator, and the choice set sizes are on average much higher on consumption data than they are on the experimental data.

Figure 3 provides the CDF of Selten's index for both SUCR and TC for both data sets. The KS-test for equality of distributions has a p-value of <0.001 between SUCR and TC for both data sets.

As these results show, SUCR and TC have high predictive power on average for both the experimental and consumption data, which means that they provide precise welfare guidance. It qualifies our results on completeness shown in Section 4.3: the relations induced by SUCR and TC are complete enough to provide precise welfare guidance for observed choice sets.

4.5 Predictive Power and Revealed Preference Properties

In this section, we indicate when and why SUCR and TC have high predictive power and provide empirical evidence of these relationships. Specifically, we show that there are two properties of revealed preferences that are especially important for the predictive power of SUCR and TC: the *number* of direct RP cycles and the *fraction* of RP cycles that are direct.

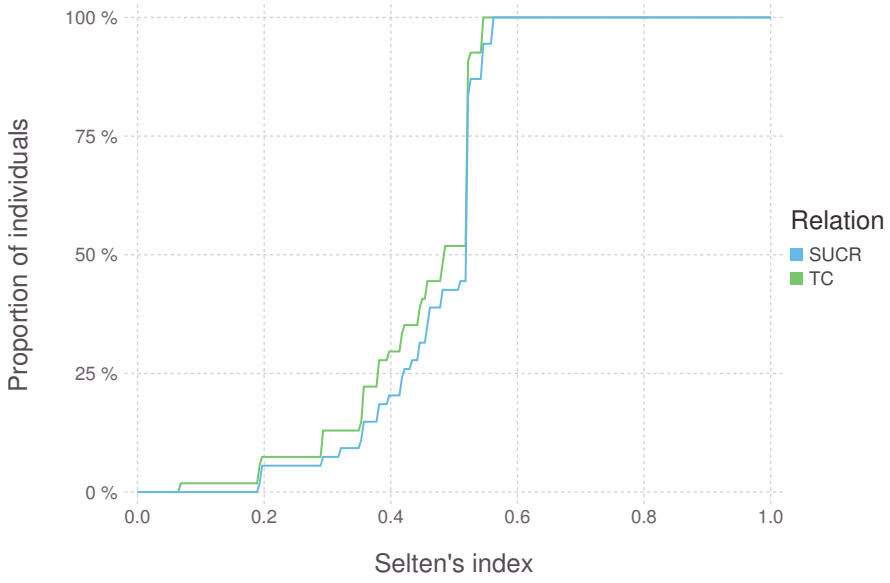
4.5.1 Number of Direct RP Cycles

As discussed previously, SUCR and TC remove all revealed preference relation elements that produce direct RP cycles, so their ability to make precise predictions should be linked to the number of such cycles. For individuals with RP cycles, the average number of direct RP

[36] In order to compute the theoretical maximum of Selten's index, we assume that in all choice sets exactly one alternative is predicted to be chosen, and then take the average over all choice sets.

Figure 3: CDF of the average Selten's index at the individual level (for individuals with RP cycles).

(a) Experimental data.



(b) Consumption data.

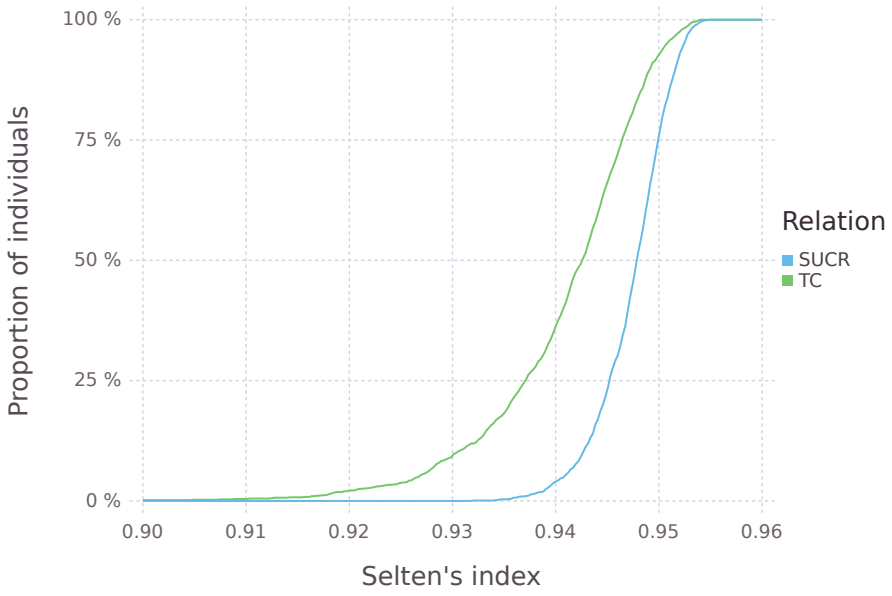


Table 6: Summary of correlations with the average Selten's index at the individual level (for individuals with RP cycles).

	Selten's index					
	Experimental		Consumption		Just acyclic SUCR	
	SUCR	TC	SUCR	TC	SUCR	TC
Number of relation elements	0.98	0.92	0.59	0.87	0.58	0.82
Number of direct RP cycles	-0.98	-0.92	-0.75	-0.81	-0.75	-0.78
Number of length 3 RP cycles	-0.89	-0.92	-0.49	-0.66	-0.48	-0.60
Number of length 2, 3 & 4 RP cycles	-0.92	-0.92	-0.35	-0.53	-0.31	-0.43
Directness index (DI)	0.58	0.78	0.56	0.69	0.52	0.65

cycles is 2.43 in the experimental data and 31.28 in the consumption data.[37] Figure 4 shows the distribution of the number of direct RP cycles for individuals with RP cycles, which indicates there is heterogeneity in the number of direct RP cycles. A majority of individuals have very few such cycles, but some exhibit comparatively more cycles.

Table 6 provides a summary of the correlations with the average value of Selten's index at the individual level. As expected, the predictive power is highly and negatively correlated with the number of direct RP cycles. The more direct RP cycles there are, the fewer relations there will be in TC and SUCR, and therefore the less predictive power they will have. Looking at correlation with Selten's index, we find that the number of direct RP cycles is more highly correlated than the number of RP cycles of length 3 or even the total number of cycles of length 2, 3, and 4.

4.5.2 Fraction of RP Cycles that are Direct

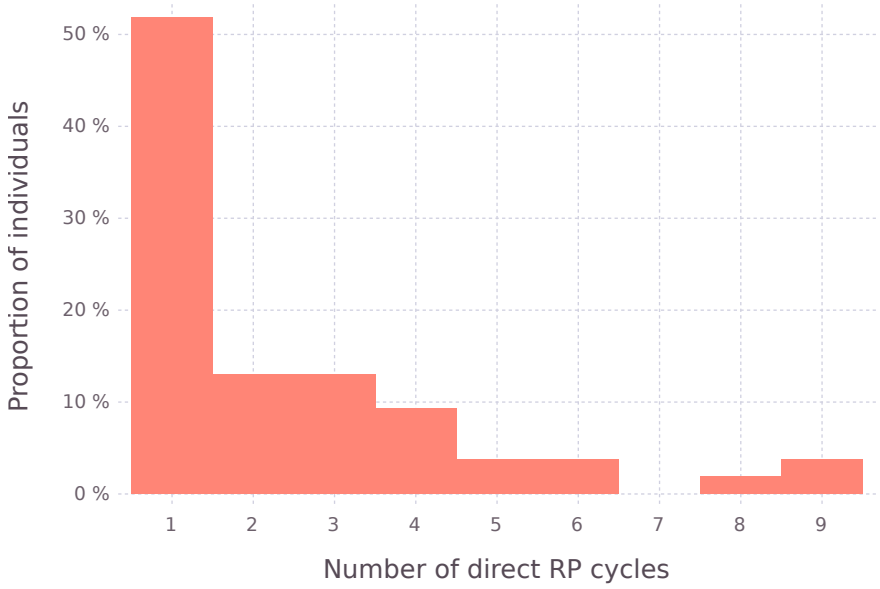
In addition to the number of direct RP cycles, if there are many cycles of longer length relative to the number of direct RP cycles, then this is a problem for both SUCR and TC. For one, SUCR will be cyclical if RP cycles remain after ignoring RP relation elements that generate direct RP cycles, so SUCR is more likely to cyclical if there are many cycles of longer lengths relative to the number of direct RP cycles. In addition, TC ignores RP relation elements beyond just the RP relation elements that produce direct RP cycles, so its predictive power is likely to be lower if there are many cycles of longer lengths relative to the number of direct RP cycles.

Also, the fraction of RP cycles that are direct has a possible interpretation in terms of

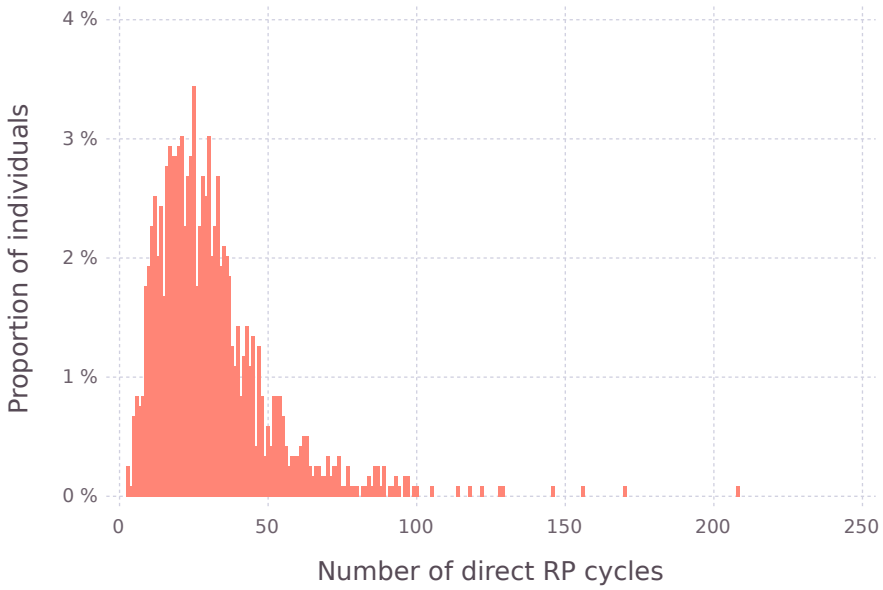
[37] To avoid double-counting, xPy and yPx count as a single cycle. Also, when multiple choices reveal that x is preferred to y, then we count the relation element just once when counting cycles. For instance, if two choices reveal xPy and one choice reveals yPx, then this counts as just one direct RP cycle.

Figure 4: Histogram of the number of direct RP cycles per individual (for individuals with RP cycles).

(a) Experimental data.



(b) Consumption data.



behavioral welfare. It could be argued that direct inconsistencies should be more obvious to decision makers, so are more likely to represent intentional violations. As such, this fraction could be interpreted as the fraction of RP cycles that represent "real" behavioral preference changes.

However, determining the fraction of RP cycles that are direct requires determining the number of cycles of all lengths, which can become computationally burdensome. Instead, we analyze an upper bound on this fraction that is much quicker and easier to calculate: the ratio of the number of length 2 cycles to the number of length 2 and length 3 cycles. We call this measure the "directness index" (DI henceforth).

Figure 5 shows the histogram of DI for individuals with RP cycles. For the experimental data, there is a large mass point at 1. The distribution for the consumption data (Figure 5b) is hump-shaped around .5, but there is also a mass point at 1. In the experimental data, the average DI is 0.79 for individuals with RP cycles, and the average is 0.58 if we exclude individuals with a DI of 1. In the consumption data, the average DI is 0.50, and it falls to 0.49 if we exclude individuals with a DI of 1.

The correlations with predictive power are given in Table 6. As expected, DI is highly and positively correlated with the value of Selten's index for TC. In the experimental data, the correlation is 0.78, and in the consumption data, the correlation is 0.69.

In addition, we find that as expected, the acyclicity of SUCR in the consumption data is highly and positively correlated with DI. The correlation between DI and a dummy variable that takes a value of 1 when SUCR has no cycles is -0.38. In fact, this correlation is even higher than the correlation between this dummy variable and the number of direct RP cycles, which is 0.36.

4.5.3 Regression Analysis

Because DI is highly correlated with the number of direct RP cycles (-0.68 in the experimental data and -0.76 in the consumption data), we also use regressions to examine the impact of DI while controlling for the number of direct RP cycles. As shown in specification 2 of Table 7, DI is positively and significantly related to Selten's index, even when controlling for the number of direct RP cycles.

5 Discussion and Conclusion

In this paper, we provide the first non-parametric empirical application of SUCR and the first empirical application of TC. The resulting analysis helps to provide an empirical answer

Figure 5: Histogram of the directness index (DI) at the individual level (for individuals with RP cycles).

(a) Experimental data.

(b) Consumption data.

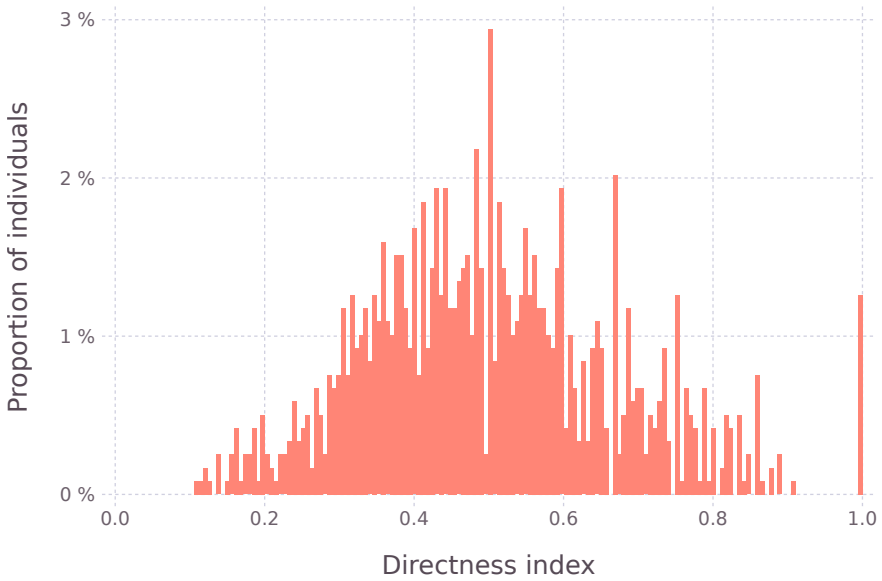


Table 7: Regressions of the average Selten's index for TC onto the number of direct RP cycles and the directness index (DI).

	(1)	(2)	(3)
VARIABLES	Experimental	Consumption	Just acyclic SUCR
Number of direct RP cycles	-0.038***	-0.00026***	-0.00027***
	(0.005)	(0.00003)	(0.00002)
Directness index (DI)	0.144***	0.00892***	0.00513***
	(0.046)	(0.00201)	(0.00141)
Observations	54	1190	946
R^2	0.8872	0.6723	0.6159

Robust standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

to whether a substantial model structure is needed to give precise welfare guidance when choices appear inconsistent due to behavioral biases.

For both data sets considered in this paper, we find that SUCR is most often acyclic and that both SUCR and TC have high predictive power, which means that they offer precise welfare guidance. Of course, to provide a more comprehensive and general answer to when SUCR 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.

It should be noted that neither of our data sets was cherry-picked to produce a desired result. Our prior belief was that SUCR and TC would not offer precise welfare guidance 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 want because they are hungry.

Finally, while the original formulation of SUCR was given for choices from menus, we felt it was important to examine the performance of SUCR 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.

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.

Aguiar, Victor and Roberto Serrano (2017). "Slutsky matrix norms: The size, classification, and comparative statics of bounded rationality". In: *Journal of Economic Theory* 172, pp. 163–201. ISSN: 0022-0531. DOI: https://doi.org/10.1016/j.jet.2017.08.007.

Ambuehl, Sandro, B. Douglas Bernheim, and Annamaria Lusardi (2014). "A Method for Evaluating the Quality of Financial Decision Making, with an Application to Financial Education". In: Working Paper Series 20618. DOI: 10.3386/w20618.

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, Jose and Miguel A. Ballester (2015). "A Measure of Rationality and Welfare". In: *Journal of Political Economy* 123.6, pp. 1278–1310. DOI: 10.1086/683838.

Beatty, Timothy and Ian Crawford (2011). "How Demanding Is the Revealed Preference Approach to Demand?" In: *American Economic Review* 101.6, pp. 2782–95. DOI: 10.1257/aer.101.6.2782.

Benkert, Jean-Michel and Nick Netzer (2018). *Informational Requirements of Nudging*. DOI: 10.1086/700072.

Bernheim, B. Douglas (2009). "Behavioral Welfare Economics". In: *Journal of the European Economic Association* 7.2-3, pp. 267–319. ISSN: 1542-4766. DOI: 10.1162/JEEA.2009.7.2-3.267.

— (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. ISSN: 0033-5533. DOI: 10.1162/qjec.2009.124.1.51.

Blundell, Richard, Martin Browning, and Ian Crawford (2003). "Nonparametric Engel Curves and Revealed Preference". In: *Econometrica* 71.1, pp. 205–240. DOI: 10.1111/1468-0262.00394.

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.

Bouacida, Elias (2019). "Eliciting Choice Correspondences: A General Method and an Experimental Implementation". In: working paper or preprint. URL: https://halshs.archives-ouvertes.fr/halshs-01998001.

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

Caplin, Andrew and Daniel Martin (2018). "Framing as Information Design". In: DOI: 10.2139/ssrn.3124194.

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). "Consistency and Heterogeneity of Individual Behavior under Uncertainty". In: *American Economic Review* 97.5, pp. 1921–1938. DOI: 10.1257/aer.97.5.1921.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman (2014). "Who Is (More) Rational?" In: *American Economic Review* 104.6, pp. 1518–50. DOI: 10.1257/aer.104.6.1518.

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.

Gerasimou, Georgios, Miguel Costa-Gomes, Carlos Cueva, and Matus Tejiscak (2019). "Choice, Deferral and Consistency". In: URL: https://www.st-andrews.ac.uk/~wwwecon/repecfiles/4/1416.pdf.

Gul, Faruk and Wolfgang Pesendorfer (2009). "A Comment on Bernheim's Appraisal of Neuroeconomics". In: *American Economic Journal: Microeconomics* 1.2, pp. 42–47. DOI: 10.1257/mic.1.2.42.

Handbury, Jessie, Tsutomu Watanabe, and David E Weinstein (2013). "How Much Do Official Price Indexes Tell Us about Inflation?" In: Working Paper Series 19504. 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. DOI: 10.1137/0204007.

Koo, Anthony Y. C. (1963). "An Empirical Test of Revealed Preference Theory". In: *Econometrica* 31, p. 646. DOI: 10.2307/1909164.

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.

— (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 (2018). "The transitive core: Inference of welfare from nontransitive preference relations". In: *Theoretical Economics* 13.2, pp. 579–606. DOI: 10.3982/TE1769.

Ok, Efe A. (2002). "Utility Representation of an Incomplete Preference Relation". In: *Journal of Economic Theory* 104.2, pp. 429–449. ISSN: 0022-0531. DOI: https://doi.org/10.1006/jeth.2001.2814.

Rubinstein, Ariel and Yuval Salant (2012). "Eliciting Welfare Preferences from Behavioural Data Sets". In: *The Review of Economic Studies* 79.1, pp. 375–387. ISSN: 0034-6527. DOI: 10.1093/restud/rdr024.

Salant, Yuval and Ariel Rubinstein (2008). "(A, f): Choice with Frames". In: *The Review of Economic Studies* 75.4, pp. 1287–1296. ISSN: 0034-6527. DOI: 10.1111/j.1467-937X.2008.00510.x.

Schwartz, Thomas (1976). "Choice functions, "rationality" conditions, and variations on the weak axiom of revealed preference". en. In: *Journal of Economic Theory* 13.3, pp. 414–427. ISSN: 00220531. DOI: 10.1016/0022-0531(76)90050-8.

Selten, Reinhard (1991). "Properties of a measure of predictive success". In: *Mathematical Social Sciences* 21.2, pp. 153–167. ISSN: 0165-4896. DOI: 10.1016/0165-4896(91)90076-4.

Tversky, A and D Kahneman (1981). "The framing of decisions and the psychology of choice". In: *Science* 211.4481, pp. 453–458. ISSN: 0036-8075. DOI: 10.1126/science.7455683.

Varian, Hal R. (2006). "Revealed preference". In: *Samuelsonian economics and the twenty-first century*, pp. 99–115. URL: http://people.ischool.berkeley.edu/~hal/Papers/2005/revpref.pdf.

6 Appendix

6.1 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.[38]

As shown in table 8, a majority of individuals in our analysis sample are working, and a plurality works more than 35 hours per week. There is, however, a substantial fraction that is not employed (42.99% 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 years.

Table 8: Average hours worked per week.

	$<$ 30 hours	30-35 hours	$>$ 35 hours	Not employed
Analysis sample over 10 years	9.48 %	3.52%	44.01%	42.99%
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.

Table 9: Income quartiles.

Percentile	25th	50th	75th
Analysis sample over 10 years	\$17,500	\$32,500	\$47,500
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.

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

Table 10: Level of education.

Education	College degree	No college degree
Analysis sample over 10 years	46.55%	53.45%
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.

The median income of the analysis sample is between $30,000 and $35,000, which is lower than the median income of individuals in the US who were 30 years old and above in 2004, as shown in table 9. The level of education of our sample is slightly higher than this group, as table 10 shows.

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, 740 out of the 1,190 panelists are women, a proportion of 62.18%. In the US population, the fraction of women among individuals aged 30 and older was 52.34% in 2004.[39]

6.2 Analytical Computer Programs

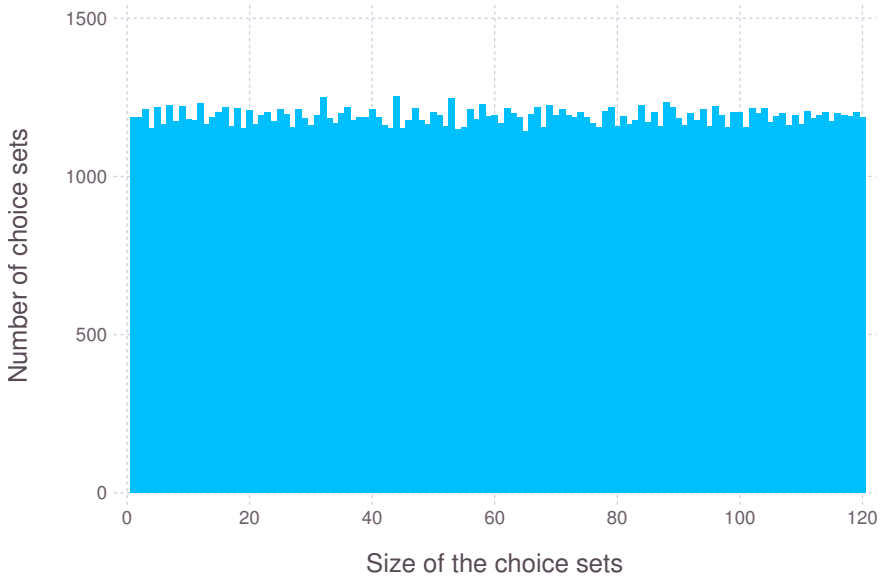
Our analytical computer programs are divided into two main parts:

1. Transforming the observed data into preference relations (either RP, SUCR or TC);
2. Determining the number of cycles and predictive power of these relations, and the relationship between them. The determination of cycles is built on Johnson (1975)'s canonical algorithm.

All programs were written using the software language Julia (versions 1.0.3), and the standard libraries included therein. The libraries used are: CSV (0.4.x), DataFrames (versions 0.18.x) to read and transform the original data, LightGraphs (versions 1.2.x), GraphIO (0.4.x) to build and save the directed graph, Distributions (versions 0.18) for the random draws, HypothesisTests (versions 0.8.x) for hypothesis testing, and Gadfly (versions 1.0.x) and Colors (0.9.x) for producing plots. The graph algorithm is a custom implementation available here. The revealed preference tests are included in a new package (under construction): RevealedPreferences. All regressions were run using Stata instead.

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

Figure 6: Histogram of the size of choice sets in the consumption data.



6.3 Choice Set Size

Figure 6 shows that the choice set sizes in the consumption data appear to be close to uniform. However, a KS-test rejects this possibility with a p-value of < 0.001.

6.4 Correlations for Uniform Random Demands

In order to check the robustness of the results found in Section 4, we ran a robustness check using uniform random demands.

6.4.1 Uniform Random Methodology: Consumption Data

Our primitives are quantities (for each individual and each period) and prices (for each market and each period). We first compute the observed expenditures for each period. We then build vectors of quantities that constitute our random bundles. Each bundle is built by drawing a vector from the simplex uniformly[40] and then multiplying it by the observed expenditures of each period to get (random) quantities. We now have observed prices and random quantities for each individual. We simulate each observed individual 1,000 times. In total, we have 1,190,000 simulated individuals.

[40]Using a flat Dirichlet distribution (see https://en.wikipedia.org/wiki/Dirichlet_distribution).

Table 11: Summary of correlations with the average Selten's index at the simulation level (for simulations with RP cycles).

	Selten's index					
	Experimental		Consumption		Just acyclic SUCR	
	SUCR	TC	SUCR	TC	SUCR	TC
Number of relation elements	0.96	0.97	0.58	0.88	0.56	0.83
Number of direct RP cycles	-0.96	-0.58	-0.64	-0.84	-0.63	-0.78
Number of length 3 RP cycles	-0.89	-0.74	-0.49	-0.75	-0.49	-0.68
Number of length 2, 3 & 4 RP cycles	-0.91	-0.69	-0.38	-0.61	-0.40	-0.57
Directness index (DI)	0.54	0.81	0.47	0.70	0.41	0.61

6.4.2 Uniform Random Methodology: Experimental Data

Our primitives are the twelve possible choice sets.In each choice set, each alternative is chosen with equal probability. For this choice setting, there is no difference between individuals, as they all faced the same choice sets. In total, we have 1,000,000 simulated individuals.

6.4.3 Results

We find that uniform random demands are much less consistent than observed demands. In the experimental data, all but 1 simulation has RP cycles. In the consumption data, 100% of the simulations have RP cycles, 75% have SUCR cycles, and 0% have TC cycles. Again, we now restrict the sample to simulations with RP cycles. For these simulations, SUCR and TC also less complete. In the experimental data, 93.85% of simulations are a half or less of a complete and acyclic relation with SUCR and 97.65% with TC. In the consumption data, on average, SUCR is 97.86% of an acyclic and complete relations, whereas TC is 94.98%, which is much lower than in the observed data.

The average predictive power drops as well. In the experimental data, the average value of Selten's index is 0.22 for SUCR and 0.16 for TC. In the consumption data, the average Selten's index is 0.98 for SUCR and 0.96 for TC. The correlation with our proposed measures are presented in Table 11, and the qualitative figures are similar to those of Table 6. This suggests that the relationship between the number and the fraction of direct RP cycles and the predictive power of SUCR and TC holds more generally than just the demands we observe in our data sets.

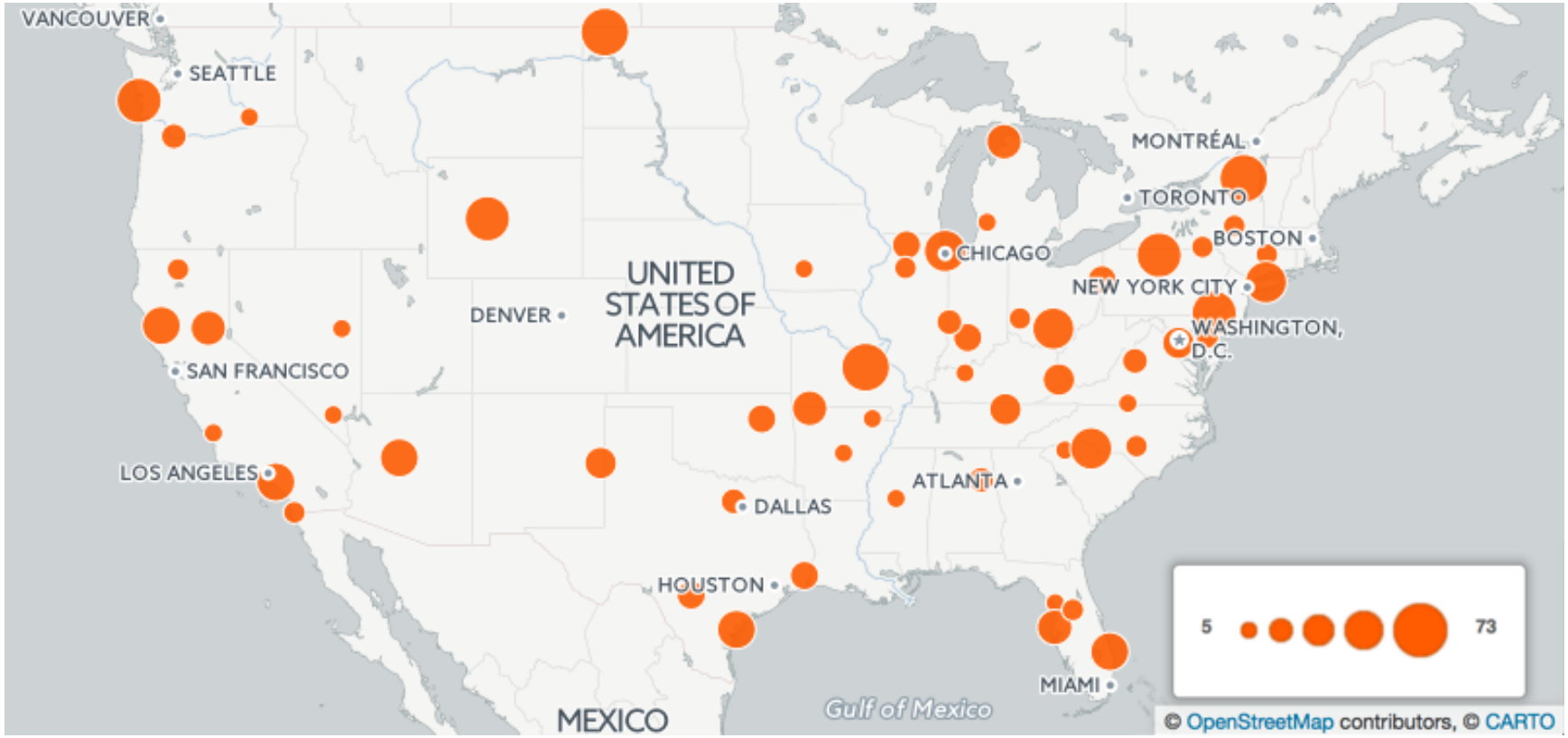


Figure 7: Individuals in the consumption data by market. The size of a bubble is proportional to the number of individuals in a given market.