PREDICTIVE POWER IN BEHAVIORAL WELFARE ECONOMICS

Elias Bouacida
Lancaster University Management School, LA1 4YX, Bailrigg, Lancaster, United Kingdom

Daniel Martin
Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, United States of America

Abstract
When choices are inconsistent due to behavioral biases, there is a theoretical debate about whether the structure of a model is necessary for providing 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 "model-free" approaches to behavioral welfare analysis. We find they typically have high predictive power, which means there is little ambiguity about what should be selected from each choice set. We also identify properties of revealed preferences that help to explain the predictive power of these approaches. (JEL: I30, C91, D12)

Keywords: Welfare economics, behavioral economics, predictive power, revealed preferences.

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 et al. (2007, 2014)) and observational (e.g., Blundell et al. (2003); Dean and Martin (2016)). Because inconsistencies generate reversals in the

The editor in charge of this paper was Juuso Välimäki.

Acknowledgments: We thank Jean-Marc Tallon and the other members of Elias's thesis defense committee (Eric Danan, Georgios Gerasimou, Olivier l'Haridon, and Stéphane Zuber), for their many helpful comments, and the editor Juuso Välimäki and three anonymous referees for their generous and productive guidance.

E-mail: e.bouacida@lancaster.ac.uk (Bouacida); d-martin@kellogg.northwestern.edu (Martin)

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 (see Bernheim (2009); Rubinstein and Salant (2012); Manzini and Mariotti (2014); Bernheim (2016)). While there are other important criteria for policymakers besides the precision of welfare guidance, if a behavioral welfare approach offers little guidance 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 "model-free" 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).[1] 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 both x and y are available.[2] However, the revealed preference relation P is unsuitable for welfare analysis if it contains a cycle: if there exists $x_1, x_2, \ldots, x_n$ such that $x_1 P x_2, \ldots, x_n P x_1$. To be free of such cycles, SUCR and TC retain only some elements of P. SUCR retains a relation element xPy if and only if y is never chosen when x and y are available (denoted as xP^*y). Alternatively, TC retains a relation element xPy if and only if it has a transitive relationship with every relation element involving x or y.

We evaluate whether these behavioral welfare relations offer precise welfare guidance by determining their *predictive power*, which is their ability to make sharp predictions.[3] 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 (unable to compare some alternatives), they do not always pin down what an agent would select from a set of alternatives, so their predictive power is in question.

1. Masatlioglu et al. (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 an introduction to revealed preference, see Varian (2006) and Adams and Crawford (2015). Note that the relation P is *strict* in the sense that it precludes indifference. See Bouacida (2019) for a method that allows for revealing indifference.

3. For other applications of predictive power in empirical revealed preference analysis, see Manzini and Mariotti (2006); Beatty and Crawford (2011); Andreoni et al. (2013); Dean and Martin (2016); Boccardi (2018).

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 choice sets such as $\{x,a\}$, SUCR predicts that any alternative 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 alternative should 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 alternative 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 for $\{x,y,z\}$ as the only individual welfare optimum for that choice set is x, but it offers no welfare guidance for $\{x,a\}$.

We study 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 (2006).[4] The latter is composed of choices from budget sets for 1,193 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. 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 (which we will call *full observability*),[6] 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.

We measure the predictive power of SUCR and TC for these data sets using the average value of Selten's index (Selten (1991)). With Selten's index, the proportion of choices that a theory predicts successfully is reduced by the *size of the area*, which is

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. De Clippel and Rozen (2020) provide warnings and guidance on how behavioral theories should be tested when there is not full observability.

the fraction of outcomes that are consistent with a theory. Thus, Selten's index has a value closer to 1 when a theory predicts successfully even though few choices would be consistent with the theory (when predictions are sharp). On the other extreme, it has a value closer to -1 when a theory fails to predict successfully even though most choices would be consistent with the theory (when predictions are weak). In this application, we calculate the size of the area by determining the fraction of options in a choice set that are predicted to be chosen by a relation. Beatty and Crawford (2011) provide an axiomatic characterization of Selten's index, and it has been used previously to determine predictive success in revealed preference testing (Manzini and Mariotti (2006); Beatty and Crawford (2011); Dean and Martin (2016)).

Using this measure, we find that SUCR and TC have high predictive power on average within-sample for both data sets.[7] Even when we restrict our attention to individuals with choice inconsistencies, the average value of Selten's index in the experimental data is 0.46 for SUCR and 0.44 for TC (relative to a maximum of 0.58 in this setting), and the average value of Selten's index in the consumption data is 0.95 for SUCR and 0.94 for TC (relative to a maximum 0.96).[8] These relatively high average values for Selten's index reflect the fact that only a small number of alternatives are predicted to be selected on average from a choice set. For individuals with inconsistent choices, the average number of options predicted to be selected from a choice set is 1.32 for SUCR and 1.38 for TC in the experimental data and 1.31 for SUCR and 1.60 for TC in the consumption data.

To learn when and why SUCR and TC have high predictive power, we break the relationship between choice and predictive power into two stages and identify the important factors in each stage. The first stage is the link between choice and the completeness of a relation (how often the relation allows us to compare two alternatives), and the second stage is the link between the completeness of a relation and its predictive power.

To illustrate the separation between these two stages, consider a simple example with three alternatives, A, B_1, and B_2, and two consumers. When offered a choice between B_1 and B_2, both consumers can be influenced by ancillary factors into choosing either alternative. However, consumers differ in their preferences: the first one prefers A to either B_i, whereas the second prefers either B_i to A. With full observability of choices, SUCR and TC consist of relation elements between A and either B_i for both consumers, and thus the completeness of these relations is identical for both consumers. However, these relations have a higher predictive power for the first consumer because when all three alternatives are available, the relations predict that only A is selected, whereas for the second consumer, both B_1 and B_2 can be selected.

7. We find that SUCR and TC have high average predictive power out-of-sample as well, as shown in Appendix E.

8. The maximum value of Selten's index is higher in the consumption data in part because choice sets are larger on average.

To understand more generally what factors drive the completeness and predictive power of these relations, we first identify properties of revealed preferences that are important for explaining the link between choice and completeness. The completeness of SUCR is driven entirely by the number of *direct* RP cycles (where x is revealed preferred to y and vice versa) because SUCR only excludes RP relation elements involved in such cycles.[9] The number of direct RP cycles also helps determine the completeness of TC because RP relations elements involved in such cycles violate transitivity, so are excluded from TC as well.[10] In addition, the completeness of TC is related to fraction of RP cycles of length 2 and 3 that are direct (of length 2). This fraction, which we call the *directness index*, correlates highly with the number of RP relation elements in cycles of length 3 that are excluded from TC because they violate transitivity. It is worth noting that while cycle length plays a critical role in this application, it is typically not considered in revealed preference analysis beyond distinguishing between violations of the Weak Axiom of Revealed Preference and the Strong Axiom of Revealed Preference.[11]

Next, we identify a property of revealed preferences that is important for explaining the predictive power a relation given its completeness: the "revealed quality" of incomparable alternatives. We measure the revealed quality of an alternative by looking how often it is revealed preferred to other alternatives instead of the reverse,[12] and our *incomparability index* is the average of this measure for all pairs of alternatives that are missing a relation element (that cannot be compared). The revealed quality of incomparable alternatives matters for predictive power because if two alternatives in a choice sets cannot be compared, then both will be predicted to be selected if they are not dominated by another alternative in the choice set.

Taken together, these factors explain a substantial portion of the variation in the predictive power of SUCR and TC for both data sets. For the experimental data, they explain 99% of the variation in the predictive power of SUCR and 90% of the variation in the predictive power of TC. For the consumption data, they explain 65% of the variation in the predictive power of SUCR and 85% of the variation in the predictive power of TC.

To the best of our knowledge, this paper is the first to use predictive power as a tool for evaluating the welfare guidance provided by behavioral welfare relations. We also believe that it provides the first non-parametric empirical applications of SUCR and

9. Without full observability, as in our consumption data, it is necessary to also control for the completeness of the RP relation because when there is not a RP relation element between two alternatives, there will not be a relation element for SUCR and TC between those alternatives either.

10. A relation with directly conflicting elements also violates the property of antisymmetry, so the number of direct RP cycles is proportionate to the number of violations of antisymmetry.

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

12. In network/graph terminology, we find the difference between the out-degree and in-degree of a node.

TC, two conservative "model-free" behavioral welfare relations.[13] Using this tool and these relations, we provide an answer to the question of how much model structure is necessary to provide precise welfare guidance in practice. 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 addition, we help to explain when and why behavioral welfare relations have higher predictive power. These factors are relatively quick to calculate, so 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.

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. In Section 5, we explain why certain properties drive completeness and predictive power. We conclude with a brief discussion in Section 6.

2. Behavioral Welfare Relations

2.1. Frames, Behavioral Biases, and Welfare

Bernheim and Rangel (2009) and Salant and Rubinstein (2008) separately proposed the idea of using frames to make welfare assessments in light of the inconsistencies in choice produced by behavioral biases. In both cases, the key data expansion is to consider, in addition to the choice itself, the ancillary conditions present when the choice is made.[14] While these ancillary conditions can impact choice, it is assumed that they do not affect the alternatives themselves or the welfare derived from them.[15]

Bernheim and Rangel (2009) also allow the econometrician to decide which frames are "welfare-relevant" and only use the choices made under welfare-relevant frames when assessing welfare. Because choices made in one welfare-relevant frame have the same weight as choices made in another welfare-relevant frame, the technical role of welfare relevance is to exclude some choices when assessing welfare. We wish to determine the very limits of predictive power for "model-free" approaches to behavioral welfare analysis in our data sets, so we consider all choices in our data sets to be welfare-relevant, as restrictions on welfare relevance would only serve to increase predictive power. In addition, there are potential downsides to imposing welfare-relevance in an ad hoc manner, as discussed in Gul and Pesendorfer (2009).

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

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

15. Rubinstein and Salant (2012); Benkert and Netzer (2018); Caplin and Martin (2020) also provide ways to use frames when assessing welfare.

2.2. SUCR, TC, and Revealed Preference

To provide welfare guidance from the set of welfare-relevant choices, Bernheim and Rangel (2009) propose using the strict unambiguous choice relation (SUCR). Formally, 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. Bernheim and Rangel (2009) assume that choices are observed from all possible subsets of choice options (full observability), and with this assumption, xP^*y only if x is chosen from a data set that includes y (specifically $\{x, y\}$). Because we study data sets both with and without full observability, we follow Manzini and Mariotti (2014) in explicitly adding this additional implication to the definition of P^*:

> An alternative x is in the relation P^* with y if x is sometimes chosen when y is available (that is, there is at least one choice set S for which S contains y and x is chosen), while y is never chosen when x is available.

By using this enhanced definition of SUCR, we avoid the possibility that SUCR has a direct cycle (xP^*y and yP^*x) merely because two alternatives x and y are never chosen in the presence of the other.

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.

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. In practice, we find that these relations offer trade-offs in terms of the degree of predictive power and the extent of acyclicity. On the one hand, TC is always nested in SUCR for the choice settings we study, so it is (weakly) less complete and has (weakly) less predictive power. In the experimental data, we find Selten's index is on average 0.02 higher for SUCR, and in the consumption data, it is on average 0.006 higher for SUCR. On the other hand, TC has an advantage over SUCR in terms of acyclicity. While both relations are guaranteed to be acyclic with full observability,[16] our consumption data does not have full observability. However, we find that TC never contains a cycle in that data set – even when SUCR does for the same individual. As a result, the rate of acyclicity is 19 percentage points higher for TC in the consumption data.

16. Nishimura (2018) shows that whenever choices are observed from all binary sets of choice options, TC is also guaranteed to be acyclic.

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 determine the number of alternatives that rank above each chosen alternative in a choice set according to the candidate order and weight that up by the frequency of facing that choice set and choosing that alternative. This approach uses choice set frequencies to overcome ambiguities, so is less conservative than SUCR and TC in making welfare assessments. In fact, 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.

2.4. Other Empirical Findings

Bernheim et al. (2015) provide the first empirical implementation of SUCR to choice data.[17] 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.

17. An application of concepts from Bernheim and Rangel (2009) also appears in Ambuehl et al. (2014).

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 reasons why the high predictive power of SUCR and TC for our consumption data might be 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 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 (2006) 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 et al. (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, 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.

TABLE 1. Two and three installment plans.

Delay	I2	D2	K2	J2	I3	D3	K3	J3
3 months	16	32	24	8	8	24	16	8
6 months					16	16	16	8
9 months	32	16	24	40	24	8	16	32
Total €	48	48	48	48	48	48	48	48

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 Table 1. Additional details are available in Manzini and Mariotti (2006).

A unique feature of the experiment of Manzini and Mariotti (2006) 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.[18] 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.[19]

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.[20]

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

18. Presenting all possible subsets is challenging because it requires many choice sets: for n alternatives, the number of choice sets is $2^n - n - 1$. To the best of our knowledge, only two other recent choice experiments present all possible subsets to decision makers (Costa-Gomes et al. (2019); Bouacida (2019)).

19. Choices were not incentivized for an additional 54 subjects, so we do not include them in our analysis.

20. As of March 2020, 178 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.

TABLE 2. Proportion of individuals that have cycles in RP, SUCR, and TC.

Data	Number of individuals	Percentage with cycles RP	SUCR	TC
Experimental	102	53%	0%	0%
Consumption	1,193	100%	19%	0%

previously, this panel contains several times more choices than existing papers that implement revealed preference tests on consumption data (e.g., Blundell et al. (2003); Dean and Martin (2016)).

In Appendix A, we provide a detailed description of our consumption data set. This includes our exclusion criteria for panelists, the manner in which bundles were constructed and prices were calculated, additional assumptions, and panelist demographics.

4. Results

In this section, we first determine the proportion of individuals who have choices that exhibit revealed preference (RP) cycles. For such individuals, we then determine the proportion that have cycles in SUCR and TC, the completeness of SUCR and TC, and finally the predictive power of SUCR and TC.

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. For both of our data sets, a majority of individuals have choices that generate at least one RP cycle, as shown in Table 2. In the experimental data, 53% of individuals have RP cycles for at least one installment plan. In the consumption data, 100% of individuals exhibit 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 et al. (2007), around 35% of subjects have RP cycles for choices from allocations over risky assets, and in the large-scale field experiment of Choi et al. (2014), around 90% of subjects exhibit RP cycles 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. A recent example is the 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 for data sets with full observability (choices observed from all subsets of alternatives). TC is also certain to be free of cycles when choices from all binary choice sets are observed because this condition is sufficiently strong to ensure that the underlying revealed preference relation will be complete.

The experimental data satisfies both conditions, so SUCR and TC are certain to be acyclic. On the other hand, neither condition is satisfied with the consumption data, so their acyclicity is in doubt in this data set. As shown in Table 2, SUCR has cycles for 19% of individuals in the consumption data, which 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, the remaining analyses will only consider individuals that have RP cycles. 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. If these subjects were included in the subsequent analyses of our experimental data, SUCR and TC would appear even more complete and would have even higher predictive power.

4.3. Completeness of SUCR and TC

A relation $\succ$ is *complete* if for all x and y in the grand set of alternatives X, either $x \succ y$ or $y \succ x$. We measure the completeness of a relation by dividing the number of relation elements it contains by the number of relation elements in a complete and acyclic relation.[21] If a relation is complete and is acyclic, there is $|X|(|X|-1)/2$ relation elements. Because RP can contain direct cycles, the number of RP relation elements can exceed this number.

Because SUCR excludes all relation elements that are a part in 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 9.6, which is 80% of the comparisons in a complete and acyclic relation. For

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

TC, the corresponding figures are 9.1 and 76%.[22] 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 one relation short with TC. An additional 13% are two relations short with SUCR and 9% with TC. Some individuals, however, have half or less of a complete and acyclic relation (9% with SUCR and 13% with TC).

In the consumption data, a complete and acyclic relation over the set of all bundles that an individual has chosen at some point would have 7,140 elements.[23] In theory, choices in the consumption data also generate revealed preference relation elements over bundles that are never chosen, but including them would inflate our assessment of the completeness of SUCR and TC. So, 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 7,082, which is 99.2% of the comparisons in a complete and acyclic relation. The corresponding figures for TC are 7,033 and 98.5%.[24] The maximal number of relation elements are 7,135 for both SUCR and TC, which is just 5 elements short of a complete and acyclic relation.

We also examine the completeness of these relations for those individuals with acyclic SUCR as a robustness check. Conditional on SUCR being acyclic, the average number of relation elements for SUCR is 7,089, which is 99.3% of the comparisons in complete and acyclic relation, and for the same subjects, the corresponding figures for TC are 7,053 and 98.8%.[25]

4.4. Predictive Power of SUCR and TC

The completeness of SUCR and TC gives us a sense of the precision of their welfare guidance. In fact, the completeness of these relations (when they are acyclic) tells us exactly how likely just one option is predicted to be selected from a random binary choice set. However, it does not tell us the precision of their welfare guidance for observed choice sets, so we also calculate the predictive power of SUCR and TC within-sample for both data sets.[26]

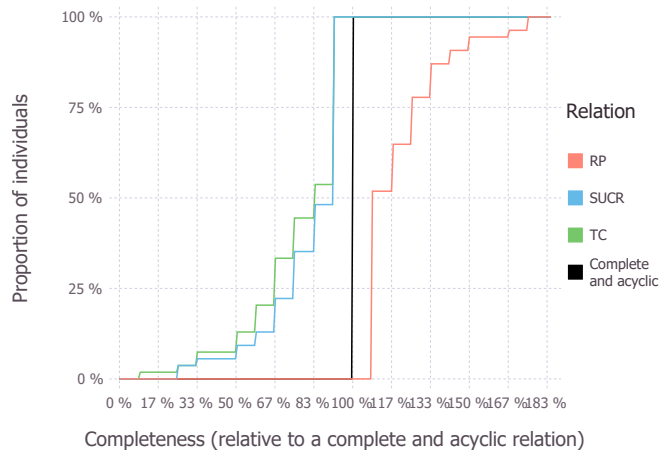
22. The difference between the average number of relations for SUCR and TC is significant (the two-sided paired t-test p-value is 0.0016), but a Mann-Whitney U-test (i.e., a Wilcoxon rank-sum test) of the two samples being drawn from the same distributions has a p-value is 0.35, so it cannot be rejected.

23. The set of chosen bundles can vary individual-by-individual, but its size does not.

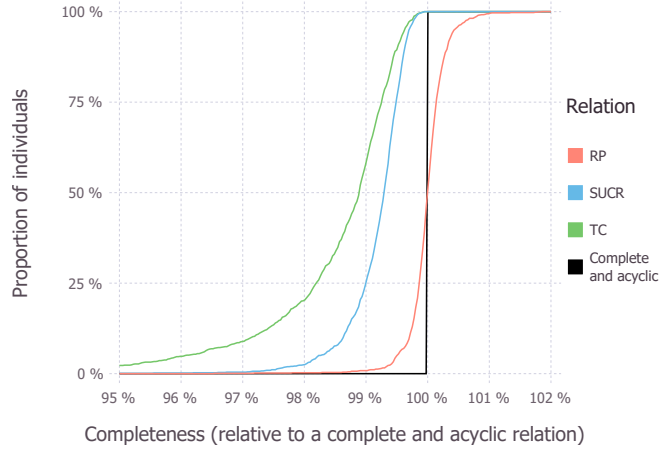
24. The difference between the average number of relations for SUCR and TC is significant (the two-sided paired t-test p-value is <0.001). For a U-test of the completeness of SUCR and TC being drawn from the same distribution, the p-value is <0.001.

25. 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 U-test of SUCR and TC completeness being drawn from the same distributions, the p-value is <0.001.

26. We calculate the predictive power of SUCR and TC out-of-sample in Appendix E.



(A) Experimental data.



(B) Consumption data.

FIGURE 1. CDF of the completeness of each relation (relative to a complete and acyclic relation) at the individual level (for individuals with cyclic RP).

To determine the predictions made by a relation for the observed choice sets, we follow Schwartz (1976); Ok (2002) in saying that the choice correspondence C induced by a (possibly incomplete) strict relation $\succ$ is $C_{\succ}(S) = \{x \in S \mid y \succ x \text{ for no } y \in S\}$.[27] The tightness of the predictions given by C is useful for studying the precision of welfare guidance because what is predicted to be selected

27. Bernheim and Rangel (2009) propose the same correspondence, which they denote as $m_{\succ}(S)$.

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 S.

We measure predictive power using the average value of Selten's index (Selten (1991)). 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 the fraction of all possible outcomes that are consistent with the theory.[28] 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 are designed to never directly contradict observed choices, they always predict successfully within-sample. As a result, r is always equal to 1, but a can vary by relation and choice set.[29] For a relation $\succ$ and choice set S, we define a as the proportion of alternatives that are predicted to be chosen, so that $a = |C_{\succ}(S)|/|S|$.

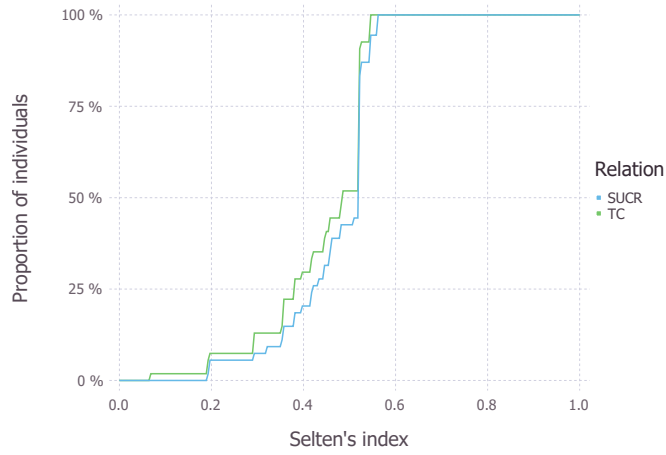
The choice set and its size are very straightforward to determine in the experimental data. In the consumption data, we take the choice set to be the set of all bundles that an individual has chosen at some point that are affordable at a given price and expenditure level. We could consider the choice 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, it 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 B.

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 for this setting is just 0.58, as shown in Table 3. To compute the theoretical maximum, we assume that in all choice sets exactly one alternative is predicted to be chosen, and then take the average value of Selten's index over all choice sets and all individuals.

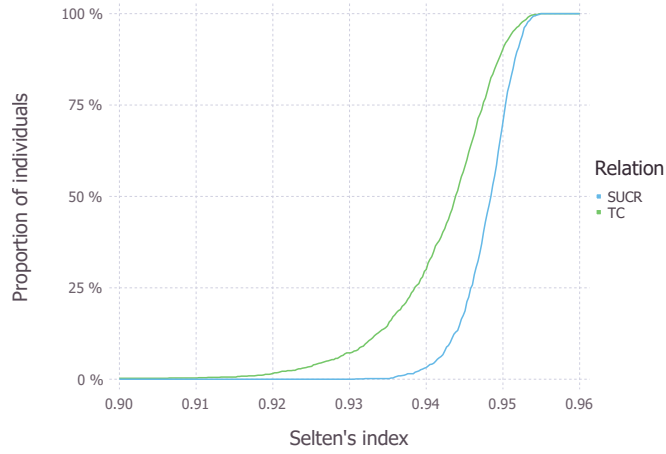
In the consumption data, the average Selten's index is 0.95 for SUCR and 0.94 for TC. The difference here is also significant, as the two-sided paired t-test p-value is <0.001. For the consumption data, the average theoretical maximum is 0.96. Selten's index is higher in the consumption data compared to the experimental data because

28. Selten's index has been used to answer other empirical questions revealed preference analysis (see Manzini and Mariotti (2006); 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" from rationality for observed choices.

29. In the analysis of Appendix E, some predictions are out-of-sample and r is not always equal to 1.



(A) Experimental data.



(B) Consumption data.

FIGURE 2. CDF of the average Selten's index at the individual level (for individuals with RP cycles).

choice set sizes are on average much higher on consumption data than they are on the experimental data.

Figure 2 provides the CDF of Selten's index for both SUCR and TC for both data sets. The distributions are significantly different for the consumption data, as the p-value of the Mann-Whitney U-test is <0.001, but the distributions are not significantly different in the experimental data (the p-value is 0.22).

One reason why the average values of Selten's index are relatively high, is that the average number of predicted options for a choice set is relatively small. In the

TABLE 3. Average value of Selten's index by choice set size in the experimental data (for individuals with RP cycles).

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

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. In the consumption data, the average number of predicted alternatives is 1.31 for SUCR and 1.60 for TC.[30] We provide more details about the predicted choice set sizes in Appendix B.

5. Explaining Predictive Power

In this section, we identify factors that help explain when and why SUCR and TC have high predictive power and provide empirical evidence of these relationships. We first show factors that help explain the completeness of SUCR and TC and then provide additional factors that help explain the predictive power of SUCR and TC given their completeness. Finally, we determine the combined explanatory power of these factors in both data sets using regression analyses.

5.1. Factors for Completeness

We start by showing that, in general, two properties of revealed preferences are important drivers of the completeness of SUCR and TC: the number of direct RP cycles and the fraction of direct and length 3 RP cycles that are direct. When looking at data sets with less than full observability, as with our consumption data, the completeness of the RP relation has to be taken into account too.

5.1.1. Factors for Completeness: SUCR . Given that x is revealed preferred to y (xPy) if and only if x is chosen in the presence of y in some choice set, there is a tight and clear relationship between the number of direct RP cycles and the number of SUCR relations. When there is a direct RP cycle between x and y, this means that both alternatives were at some point chosen when the other was available, so no relation is created by SUCR. The reverse is also true: when there is not a direct RP cycle between x and y, then a relation is generated by SUCR between x and y if x is revealed

30. Both differences are significant, as the respective two-sided paired t-test p-values are 0.0025 and <0.001.

preferred to y because the absence of a cycle means that y was never chosen when x was available.

Thus, the number of SUCR relations elements is equal to the number of RP relation elements minus two times the number of direct RP cycles.[31] So when the RP relation is complete, as with full observability, the completeness of SUCR is determined entirely by the number of direct RP cycles. With less than full observability, as in our consumption data, the completeness of SUCR is determined both by the number of direct RP cycles and the completeness of the RP relation.

5.1.2. Factors for Completeness: TC . TC is generated by excluding RP relation elements that are intransitive. The relation elements in a direct RP cycle (say xPy and yPx) are intransitive because x cannot be (strictly) revealed preferred to x, so the number of direct RP cycles also helps determine the completeness of TC. However, the completeness of TC also depends on the number of additional RP relations excluded from TC because they appear in length 3 RP cycles.

The ratio of direct RP cycles to direct and length 3 RP cycles (the directness index) helps to capture this additional dependency because holding fixed the number of direct RP cycles, a lower value of the directness index means not only more length 3 RP cycles, but also a lower likelihood that there are direct subcycles in each length 3 RP cycle. A *subcycle* is a cycle that is strictly shorter than another cycle and shares at least one relation element with it. It matters because when there are more subcycles contained in a length 3 RP cycle, there are (weakly) fewer additional relation elements excluded from TC on top of the relations elements excluded because they are in direct cycles.

First, with no subcycles, all three relation elements in a length 3 RP cycle are intransitive, so three additional relation elements are excluded from TC. If there is one just one subcycle, two additional relation elements are excluded. Without loss of generality, assume $xPyPzPx$ and $xPyPx$. On top of the relations elements removed because of the direct RP cycle, TC excludes yPz and zPx because we do not have xPz or zPy. If there are two subcycles, then one relation in the length 3 RP cycle can become transitive. Assume $xPyPzPx$, $xPyPx$, and $yPzPy$. At least amongst these alternatives, zPx is transitive, because we have zPx and xPy implies yPz and zPx implies yPx. Thus, one or possibly no additional relations are excluded depending on whether that relation element is transitive more generally. On the other extreme, when there are three subcycles, all three relation elements will be excluded for being in direct RP cycles, so there are no additional relations to exclude.

When there is full observability, as in our experimental data, all intransitivities are either due to direct or length 3 RP cycles because the RP relation is complete. However, when there is less than full observability, as in our consumption data, relation elements can be intransitive because of incompleteness: xPy and yPz but

31. The tightness of this relationship relies on the fact that our revealed preference P contains no indifference.

not xPz because there is no relation between x and z. We account for this additional source of intransitivities with two other factors. First, the completeness of the RP relation is clearly related to how often this can occur.[32] Second, all relation elements in cycles of length 4 and above without subcycles are excluded from TC because they are intransitive due to incompleteness. To see this, take any three alternatives x, y, z following each other in the cycle, so that $xPyPz$. This relationship must be intransitive due to incompleteness because if xPz, then it would form a subcycle with the remaining relation elements in the cycle, and if zPx, then there would be a subcycle of length 3 between those alternatives.

5.2. Factors for Predictive Power

If an option is dominated by another option in a choice set, then it will not be predicted to be chosen, so when SUCR and TC indicate most options in a choice set are dominated, they make sharp predictions for that choice set. Because a relation is more likely to identify that one option dominates another when it is more complete, the completeness of SUCR and TC is an important determinant of their predictive power.

However, completeness is not the only driver of the predictive power of these relations. It also matters which alternatives a relation is able compare. To see this, imagine a choice set of $\{x, y, z\}$. Knowing that $y \succ z$ allows us to rule out z from being chosen, so learning also that $x \succ z$ does not improve the predictive power of $\succ$. On the other hand, learning that $x \succ y$ improves the predictive power of $\succ$ because now y can be ruled out too.

In general, when relation elements are missing between two alternatives, predictive power is lower (than it would be if a relation element was present between those alternatives) if and only if those alternatives are otherwise undominated in a choice set in which they both appear. First, if those alternatives are otherwise undominated in a choice set in which they both appear, then adding a relation elements between them will necessarily make one dominated, increasing predictive power. Second, if predictive power is higher after adding a relation element between two alternatives, then it must be that one of those alternatives is newly dominated in a choice set, which is only possible if both alternatives appear in the same set together. Looking at the example above, predictive power was lower without a relation between x and y because having one would allow us to rule one out from $\{x, y, z\}$.

A pair of alternatives are undominated by other alternatives in a choice set if they are the "best" options in that choice set, so missing a relation element between alternatives that are best in many choice sets can be especially damaging to predictive power. We illustrate this point with four alternatives x, y, z, a, and a complete preference relation $x \succ y \succ z \succ a, x \succ z, x \succ a, y \succ a$, obtained from a data set with full observability. If we remove just one relation element from $\succ$, the impact on the

32. Also, like SUCR, the completeness of the RP relation matters for TC because it has no relations when RP has none either.

predictive power of $\succ$ depends on what relation element is removed. Imagine that $x \succ y$ is removed from the preference relation. Predictive power decreases in all sets where x and y are the best alternatives: $\{x,y\}$, $\{x,y,z\}$, $\{x,y,a\}$, and $\{x,y,z,a\}$. Predictive power decreases in these sets by $1/2$, $1/3$, $1/3$, and $1/4$ respectively. Imagine that $y \succ z$ is removed instead. As before, predictive power decreases in all sets where y and z are the best alternatives, but this is a smaller number of sets: $\{y,z\}$ and $\{y,z,a\}$. Predictive power decreases in these sets by $1/2$ and $1/3$.

In fact, removing a single relation element between two alternatives that are the best in many choice sets can have a bigger impact on predictive power than removing multiple relation elements between alternatives that are the best in few choice sets. Imagine now that both the relation elements $y \succ a$ and $z \succ a$ are removed. Predictive power decreases just in the sets $\{y,a\}$, $\{z,a\}$ and $\{y,z,a\}$, and the decrease in predictive power is $1/2$, $1/2$, and $1/3$ in these sets, which leads to a less of a drop in predictive power on average than just removing $x \succ y$. Thus, the variation in predictive power is not monotonic with the completeness of $\succ$.

To determine whether relation elements are missing between the best alternatives in choice sets, we construct a new index called the *incomparability index*. Roughly speaking, it measures the average "revealed quality" of alternatives that a relation cannot compare. We take the revealed quality of an alternative to be the difference between the number of other alternatives it dominates and the number of other alternatives that it is dominated by (according to the relation considered). For each pair of alternatives that are missing a relation element between them (that are not compared), we compute the average of this difference,[33] and we average this over all the pairs of alternatives that are missing relation elements. So, a higher incomparability index means that the alternatives with missing relation elements are on average more likely to be revealed preferred than the opposite.

5.3. Regression Analysis

Table 4 provides descriptive statistics for these factors for our two data sets. The standard deviation for each factor relative to its mean suggests substantial variation across individuals.

Table 5 presents the results of regressions that show the relationships between these factors and the predictive power of SUCR for both data sets.[34] For robustness, we also show the results for the subsample of individuals with acyclic SUCR. In all three regressions, both factors we have identified are significant at the 1% level.

In the experimental data, the factors we have identified explain almost all of the variation in predictive power, with an adjusted R^2 of 99%. On the consumption data, the factors we have identified explain around 65% of the variation in the predictive

33. In a graph theory terminology, we compute the difference between the outdegree and the indegree.

34. In Appendix C, we also examine the relationships between these factors and the completeness of each relation.

TABLE 4. Summary statistics for the factors identified that explain the completeness and predictive power of SUCR and TC (for individuals with cyclic RP).

	Experimental			Consumption		
	Mean	Med	Std	Mean	Med	Std
Direct RP cycles	2.43	1	2.08	29.17	25	19.73
Directness index	0.79	0.88	0.22	0.51	0.51	0.17
RP completeness (% acyclic, complete relation)	120	108	17	100	100	0.33
RP cycles of length$\geq$ 4 w/o subcycles				0.28	0	1.59
Incomparability index (SUCR)	0.67	1.00	0.59	2.51	2.68	11.84
Incomparability index (TC)	0.66	1.00	0.58	3.11	3.25	13.23

TABLE 5. Regressions of the average Selten's index for SUCR onto the completeness of RP, the number of direct cycles and the incomparability index (for individuals with cyclic RP).

VARIABLES	(1) Experimental	(2) Consumption	(3) Just acyclic SUCR
Number of direct RP cycles	-0.046***	-0.00015***	-0.00018***
	(0.00064)	(0.0000070)	(0.0000085)
Completeness of RP		0.033	0.063**
		(0.023)	(0.030)
Incomparability index	-0.028***	0.000067***	0.000064***
	(0.0023)	(0.0000052)	(0.0000053)
Constant	0.59***	0.92***	0.89***
	(0.0027)	(0.023)	(0.030)
Observations	54	1,193	963
Adjusted R^2	0.99	0.65	0.65

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

power of SUCR. For comparison, variation in the total number of RP cycles (for the 92% of individuals where we can count the exact number of RP cycles) explains 84% of the variation in the predictive power of SUCR for the experimental data (instead to 99%) and 4% of the variation in the predictive power of SUCR for the consumption data (instead to 67%).[35]

As expected, the number of direct RP cycles has a positive impact on the predictive power of SUCR (holding fixed the incomparability index in the experimental data and holding fixed the completeness of RP in the consumption data as well). Holding fixed the number of direct RP cycles, the incomparability index has a negative impact on predictive power in the experimental data, but a positive impact in the consumption data. This is because in the consumption data higher quality alternatives are actually less likely to be in choice sets together (given that they require high expenditure), so if

35. 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. For computational complexity reasons, we have capped the total number of RP cycles we count to 1,000,000. We have 94 individuals who have more than 1,000,000 RP cycles and for them we do not know the exact number of RP cycles.

TABLE 6. Regressions of the average Selten's index for TC onto the completeness of RP, the number of direct RP cycles, the incomparability index, the directness index and the number of RP cycles without subcycles of length 4 or greater (for individuals with cyclic RP).

VARIABLES	(1) Experimental	(2) Consumption
Number of direct RP cycles	-0.035***	-0.00040***
	(0.0054)	(0.000021)
Directness index	0.20***	0.0024*
	(0.048)	(0.0013)
Completeness of RP		0.97***
		(0.067)
RP cycles of length ≥ 4 without subcycles		-0.00014
		(0.00024)
Incomparability index	-0.031***	0.00011***
	(0.010)	(0.0000077)
Constant	0.39***	0.020
	(0.047)	(0.067)
Observations	54	1,193
Adjusted R^2	0.90	0.85

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

relation elements are missing between higher quality alternatives, the best alternatives in each choice set are actually more likely to be comparable.

Table 6 shows the regression results for the factors that explain the predictive power of TC for both data sets.[36] In both regressions, the factors we have identified are significant at the 10% level, except for the number of cycles of length 4 and above without subcycles.[37]

In the experimental data, the factors we have identified explain most of the variation in predictive power, with an adjusted R^2 of 90%. For the consumption data, the factors we have identified explain a similar fraction (85%) of the variation in the predictive power of TC. For comparison, variation in the total number of RP cycles (for the 92% of individuals where we can count the exact number of cycles) explains 84% of the variation in the predictive power of TC for the experimental data (instead to 90%) and 8% of the variation in the predictive power of TC for the consumption data (instead to 82%). Once again, we get the expected signs for the number of direct RP cycles and the directness index, and a change in the direction of the impact of the incomparability index between data sets.

36. Because all individuals have acyclic TC, there is no need to also look at a subsample of individuals for this relation.

37. However, this factor is significant in a regression of the completeness of TC onto these factors, as shown in Appendix C,

6. Discussion and Conclusion

For both of the standard data sets considered in this paper, we find that SUCR and TC have high predictive power on average, which means they typically offer precise welfare guidance. However, to provide a more comprehensive and general answer to when these relations have high predictive power, it would be necessary to look at other experimental and non-experimental data sets, such as those examined in the behavioral economics literature. Given the drivers of predictive power that we identify, the predictive power of these approaches could be lower in settings where framing is more actively manipulated and so choices are more directly conflicting.

That said, we feel that the data sets examined in this paper represent a valuable testing ground because behavioral biases are likely to influence choices made in both settings. For instance, when making grocery store purchases, consumers may be drawn to a product due to its flashy packaging or may buy products when they are hungry they would not have bought otherwise. When making experimental choices, subjects may select options presented at the top of the list more often. Given the many possible behavioral distortions in these settings, our prior belief was that SUCR and TC would not offer precise welfare guidance for these data sets.

Finally, this paper presents (to the best of our knowledge) the first non-parametric empirical application of SUCR and the first empirical application of TC. While our results speak to the precision of the welfare guidance provided by these relations, they leave open other important questions about these approaches, including the nature of the guidance they actually provide. When answering such questions, it would be important to also consider other "model-free" approaches to behavioral welfare analysis that have high predictive power, such as the welfare relation proposed by Apesteguia and Ballester (2015).

Appendix A: Consumption Data: Additional Details

A.1. 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 1,193 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.[38] 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 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 5,936,026 goods purchased during 1,328,712 store trips, accounting for 330,669 UPCs.

For the singles in our analysis sample, the average expenditure per month and per panelist on the goods we have kept is $248.47, whereas the average total expenditure per month and panelist is $427.27 for all households and goods in the NCP over these 10 years.

A.2. 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,[39] frozen food and packaged meat. Average budget shares for these product categories are given in Table A.1. 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 et al. (2013) study inflation with price indices built similarly.

A.3. 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

38. For an assessment of the representativeness of our sample, see Appendix A.5.

39. 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 A.1. Average budget shares (expenditure on a good category in proportion to total expenditure) in a month.

Product	Average	Standard deviation
Alcoholic beverages	6.41%	13.41%
Dairy products	13.99%	11.07%
Deli foods	3.36%	14.54%
Dry groceries	57.96%	19.40%
Frozen food	13.84%	11.11%
Packaged meat	4.44%	14.13%

each market are given in Figure A.1. 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.[40]

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.

A.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 et al. (2003); Dean and Martin (2016)).[41]

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

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

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.

A.5. Demographic Characteristics

All of the subjects who participated in the experiments of Manzini and Mariotti (2006) 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.[42]

As shown in table A.2, 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 (42.61% 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 A.2. Average hours worked per week.

Hours worked	< 30 hours	30-35 hours	> 35 hours	Not employed
Analysis sample over 10 years	9.33 %	3.48%	44.59%	42.61%
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 $30,000 and $35,000, which is lower than the median income of individuals in the US who were 30 years old and

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

TABLE A.3. Income quartiles.

Percentile	25th	50th	75th
Analysis sample over 10 years	$22,500	$32,500	$47,500
30+ years old 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 A.4. Level of education.

Education	College degree	No college degree
Analysis sample over 10 years	46.92%	53.08%
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.

above in 2004, as shown in table A.3. The level of education of our sample is slightly higher than this group, as table A.4 shows.

In the experiments of Manzini and Mariotti (2006), the subjects were a roughly even mix of men and women (see footnote 9 of Manzini and Mariotti (2006)). In the analysis sample of our consumption data, 741 out of the 1,193 panelists are women, a proportion of 62.11%. In the US population, the fraction of women among individuals aged 30 and older was 52.34% in 2004.[43]

43. US Census Bureau, CPS survey, Annual Social and Economic Supplement, 2004.

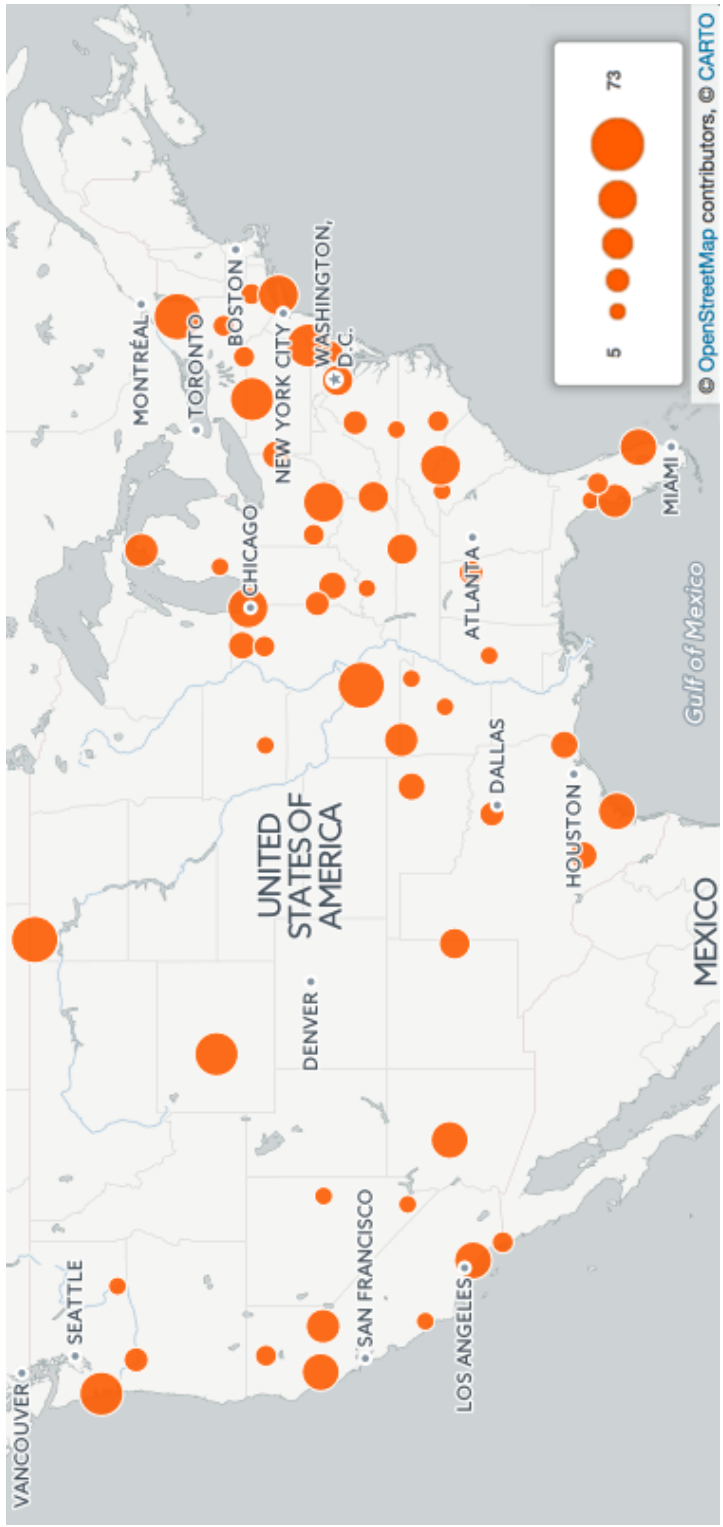


FIGURE A.1. Individuals in the consumption data by market. The size of a bubble is proportional to the number of individuals in a given market.

VARIABLES	(1) Experimental	(2) Consumption	(3) Just acyclic SUCR
Number of direct RP cycles	-0.083***	-0.00028***	-0.00028***
	(0.0)	(0.0)	(0.0)
Completeness of RP relation		1.0***	1.0***
		(0.00)	(0.00)
Constant	1.0***		
	(0.0)		
Observations	54	1,193	963
Adjusted R^2	1.0	1.0	1.0

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

TABLE C.1. Regressions of the completeness of SUCR onto the number of direct RP cycles (for individuals with cyclic RP).

Appendix B: Predicted Set Sizes and Choice Sets Sizes

Figure B.1 shows the number of alternatives that are predicted to be chosen from a choice set (the *predicted set size*) averaged at the individual level for SUCR and TC in both data sets. For each data set, we evaluate whether the distributions are statistically different between SUCR and TC using a Mann-Whitney U-test of the samples being drawn from the same distribution. For the experimental data the p-value is 0.20, and for the consumption data the p-value is < 0.001.

Figure B.2 shows that the distribution of choice set sizes for all individuals in the consumption data appears roughly uniform. However, an Anderson Darling-test of the sample being drawn from the uniform rejects this possibility with a p-value < 0.001.

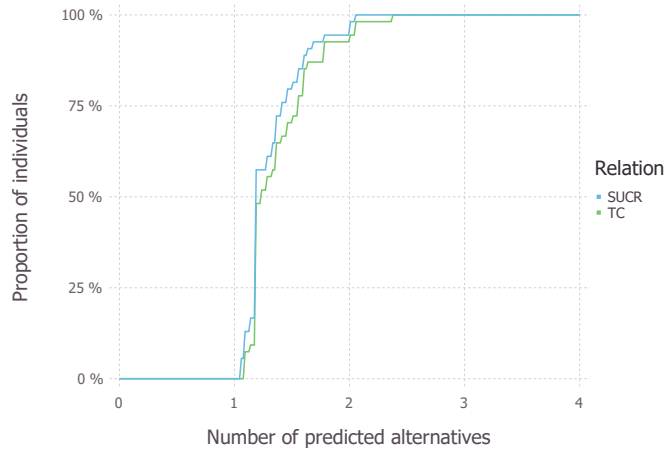
Appendix C: Regression Analysis: Completeness

C.1. Completeness of SUCR

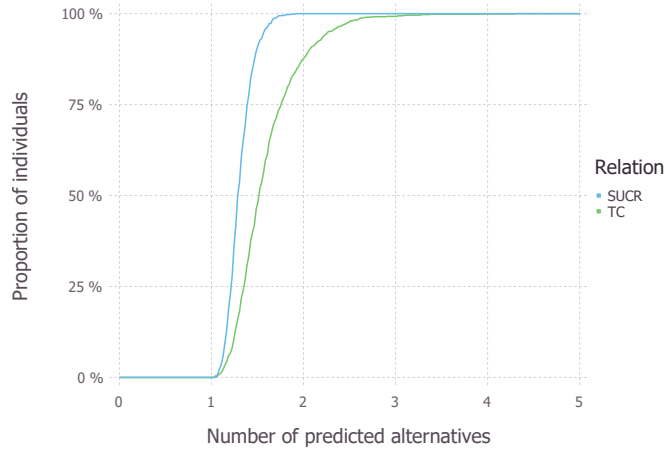
In Table C.1 we confirm the deterministic relationship between the completeness of SUCR and the number of direct RP cycles (as discussed in Section 5.1.1). The coefficients in the experimental and consumption data are different because we use the percentage of completeness, and not the absolute number of relation elements.

C.2. Completeness of TC

In Section 5.1.2, we identify several factors that explain the completeness of TC: the number of direct RP cycles, the directness index, and when there is not full observability, the completeness of RP and the number of cycles of length 4 or greater without subcycles. Table C.2 shows that these factors are all significant and have the expected signs. The number of direct RP cycles and the number of RP cycles of length



(A) Experimental data



(B) Consumption data

FIGURE B.1. CDF of the average number of predicted alternatives for TC and SUCR at the individual level (for individuals with RP cycles).

4 and above without subcycles have a negative relationship with the completeness of TC, whereas the directness index and the completeness of the RP relation have a positive relationship. Additionally, the variation in these factors explains over 90% of the variation in the completeness of TC.

For comparison, variation in the total number of RP cycles (for the 92% of individuals where we count the exact number of cycles) explains 93% and 86% of the variation in the completeness of SUCR and TC for the experimental data (instead

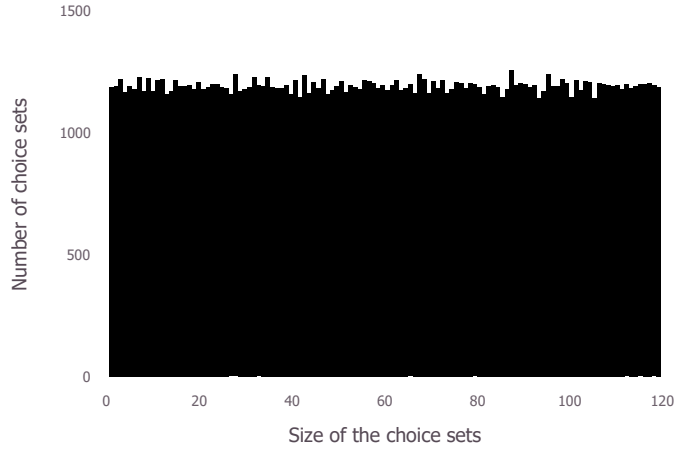


FIGURE B.2. Histogram of the size of choice sets in the consumption data.

TABLE C.2. Regressions of the completeness of TC onto the number of direct RP cycles, the directness index, and the number of cycles without subcycles of length 4 or greater (for individuals with cyclic RP).

VARIABLES	(1) Experimental	(2) Consumption
Number of direct RP cycles	-0.061***	-0.00064***
	(0.0075)	(0.000015)
Directness index	0.36***	0.0020**
	(0.062)	(0.00087)
Completeness of RP relation		2.4***
		(0.058)
RP cycles of length ≥ 4		-0.0013***
without subcycles		(0.00026)
Constant	0.62***	-1.4***
	(0.067)	(0.058)
Observations	54	1,193
Adjusted R^2	0.92	0.95

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

of 100% and 92%) and 9% and 10% of the variation in the completeness of SUCR and TC for the consumption data (instead of 100% and 95%).

Appendix D: Robustness Analysis: Uniform Random Demands

To check the robustness of the results found in Section 4, we ran an additional analysis using uniform random demands instead of observed demands. This indicates whether our results hold more generally beyond observed demands.

D.1. Uniform Random Demand 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. This yields a budget for this period. We then draw vectors of quantities that constitute random bundles. Each draw is made using uniform distribution on the simplex, using a flat Dirichlet distribution.[44] We then multiply 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 1,193,000 simulated individuals.

D.2. Uniform Random Demand Methodology: Experimental Data

Our primitives are the eleven choice sets for each installment plans, which means 22 choices sets in total. In each choice set, each alternative is chosen with equal probability. Here, all individuals faced the same choice sets and therefore are the same from a random distribution point of view. In total, we simulate a 1,000,000 individuals.[45]

D.3. Results

We find that uniform random demands are less consistent than observed demands. In the experimental data, 100% of the simulations have RP cycles. In the consumption data, 100% of the simulations have RP cycles, 73% have SUCR cycles, and 0% have TC cycles.

For these simulations, SUCR and TC are also less complete. In the experimental data, 93.79% of simulations are a half or less of a complete and acyclic relation with SUCR and 97.62% with TC. In the consumption data, SUCR is 97.90% of an acyclic and complete relation on average, whereas TC is 95.08%, which is lower than in the observed data.[46]

44. See https://en.wikipedia.org/wiki/Dirichlet_distribution.

45. There is in total 429,981,696 possible random individuals.

46. The difference between SUCR and TC in both data sets is statistically significant, as the p-values of the two-sided one-sample paired t-tests are <0.001).

TABLE D.1. Regressions of the average Selten's index for SUCR onto the completeness of RP, the number of direct cycles and the incomparability index (for simulations with cyclic RP).

VARIABLES	(1) Experimental	(2) Consumption	(3) Just acyclic SUCR
Number of direct RP cycles	-0.056***	-0.000098***	-0.00014***
	(0.000015)	(0.0000019)	(0.0000016)
Incomparability index	-0.085***	0.00014***	0.00012***
	(0.00020)	(0.0000029)	(0.0000028)
Completeness of RP		0.019***	0.076***
		(0.0058)	(0.0059)
Constant	0.69***	0.93***	0.87***
	(0.00028)	(0.0057)	(0.0058)
Observations	1,000,000	1,193,00	318,315
Adjusted R^2	0.98	0.61	0.61

Clustered standard errors in parentheses (at the individual level).
*** p<0.01, ** p<0.05, * p<0.1

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 0.94 for SUCR and 0.92 for TC.[47]

D.4. Factors Explaining Predictive Power

In the main analysis, we have identified several factors explaining predictive power. We assess the relevance of these factors on the randomly generated data. Tables D.1 and D.2 are similar to Tables 5 and 6. All important factors remain significant and keep the same signs. In addition, these factors have a similar explanatory power for uniform random demands and observed demands, as reflected the adjusted R^2.

Appendix E: Subsample Analysis

As an additional analysis, we reserve a portion of the sample as a "training" sample (keeping the remainder as a "test" sample), generate SUCR and TC using the training sample, and then examine the performance of each relation for the training sample, test sample, and full sample.

In general, Selten's index balances two features: predictive success and predictive power. A relation has high predictive success if the choices made from each choice set are consistent with the relation. A relation has high predictive power if few choices from each choice set would be consistent with the relation. SUCR and TC always "respect" observed choices (a chosen alternative is never dominated by any

47. Again, the difference in the between SUCR and TC in both data sets is statistically significant, as the p-values of the two-sided one-sample paired t-tests are <0.001).

TABLE D.2. Regressions of the average Selten's index for TC onto the completeness of RP, the number of direct RP cycles, the incomparability index, the directness index and the number of RP cycles of length 4 or above without subcycles (for simulations with cyclic RP).

VARIABLES	(1) Experimental	(2) Consumption
Number of direct RP cycles	-0.0037***	-0.00045***
	(0.000062)	(0.0000049)
Directness index	1.9***	-0.0037***
	(0.005)	(0.00087)
Incomparability index	-0.043***	0.00029***
	(0.00038)	(0.0000061)
Completeness of RP		1.3***
		(0.015)
RP cycles of length ≥ 4		0.000058***
without subcycles		(0.000012)
Constant	-0.70***	-0.34***
	(0.0028)	(0.015)
Observations	1,000,000	1,193,000
Adjusted R^2	0.68	0.88

Note: Clustered standard errors in parentheses (at the individual level).
*** p<0.01, ** p<0.05, * p<0.1

other alternative available in the choice set), so they always predict successfully within-sample. As a result, variation in Selten's index within-sample is always due to variation in predictive power. However, SUCR and TC may not always predict successfully out-of-sample, so variation in Selten's index is due both to variation in predictive success and variation in predictive power.

Second, by varying the size of training sample, we can determine how many observations it takes to get high predictive power. We generate variation in the size of the training sample by selecting either 1, 2, 3, or 4 choice sets from a data set and then removing the observations associated with those choice sets.[48] These removed observations then form the test sample. For a given size of the training sample, the actual training sample is generated by randomly selecting choice sets a thousand times for each individual.

Table E.1 and E.2 show the average predictive power of SUCR and TC obtained for the training sample (within-sample), test sample (out-of-sample), and full sample depending on the number of choice sets selected to have their observations removed. The main analysis corresponds to having 0 choice sets removed.

First, by looking at the test sample, we can determine the predictive power and predictive success of these relations out-of-sample. The level of Selten's index remains high in the test sample, though it is substantially lower than in the training sample for

48. In the experimental data, we remove the observations for both versions of a choice set, so we remove a total of between 2 and 8 observations.

TABLE E.1. Predictive power in the experimental data (for individuals with cyclic RP).

Variable	Sample	Number of choice sets removed 0	1	2	3	4
RP cyclic	all	52.94%	50.38%	47.39%	43.86%	39.61%
SUCR cyclic	all	0%	0.27%	0.48%	0.67%	0.95%
SUCR completeness	all	79.78%	79.51%	78.88%	77.69%	75.81%
TC completeness	all	76.08%	75.05%	73.06%	70.11%	65.75%
Direct RP cycles	all	2.43	2.30	2.16	2.01	1.85
Directness index	all	0.79	0.80	0.80	0.81	0.82
SUCR incomparability index	all	0.67	0.56	0.42	0.26	0.09
TC incomparability index	all	0.66	0.55	0.41	0.25	0.08
SUCR predictive success	training	1.0	1.0	1.0	1.0	1.0
	test		0.91	0.90	0.90	0.88
SUCR predicted set size	training	1.32	1.31	1.30	1.30	1.30
	test		1.30	1.30	1.31	1.33
SUCR Selten's index	training	0.46	0.47	0.47	0.47	0.47
	test		0.38	0.36	0.35	0.33
	all	0.46	0.46	0.45	0.44	0.42
TC predictive success	training	1.0	1.0	1.0	1.0	1.0
	test		0.94	0.93	0.93	0.93
TC predicted set size	training	1.38	1.38	1.40	1.42	1.46
	test		1.39	1.42	1.45	1.51
TC Selten's index	training	0.44	0.44	0.43	0.43	0.41
	test		0.37	0.35	0.33	0.30
	all	0.44	0.43	0.42	0.40	0.37

both relations and in both data sets. This decrease is mainly due to a loss of predictive success in the experimental data, whereas it is due to both a loss of predictive success and predictive power in the consumption data.

Second, by varying the size of training sample, we can determine how many observations it takes to get high predictive power. For both data sets, the predicted set size and the level of Selten's index remains high even after losing four observations, though the impact of decreasing the number of observations is strongest for TC and for the experimental data.

TABLE E.2. Predictive power in the consumption data (for individuals with cyclic RP).

Variable	Sample	Number of choice sets removed 0	1	2	3	4
RP cyclic	all	100.00%	100.00%	100.00%	100.00%	100.00%
SUCR cyclic	all	19.28%	18.89%	18.50%	18.13%	17.72%
RP completeness	all	100.00%	99.17%	98.34%	97.52%	96.71%
SUCR completeness	all	99.18%	98.36%	97.55%	96.74%	95.94%
TC completeness	all	98.50%	97.69%	96.88%	96.08%	95.29%
Direct RP cycles	all	29.17	28.69	28.21	27.74	27.27
Directness index	all	0.51	0.52	0.52	0.52	0.52
SUCR incomparability index	all	2.51	-21.30	-28.12	-30.97	-32.41
TC incomparability index	all	3.11	-15.40	-22.24	-25.69	-27.73
RP cycles length ≥ 4 without subcycles	all	0.28	0.27	0.26	0.25	0.24
SUCR predictive success	training	1.0	1.0	1.0	1.0	1.0
	test		0.69	0.69	0.69	0.70
SUCR predicted set size	training	1.29	1.31	1.31	1.30	1.30
	test		1.94	1.95	1.96	1.97
SUCR Selten's index	training	0.95	0.95	0.95	0.95	0.95
	test		0.61	0.61	0.61	0.62
	all	0.95	0.95	0.94	0.94	0.94
TC predictive success	training	1.0	1.0	1.0	1.0	1.0
	test		0.80	0.81	0.81	0.81
TC predicted set size	training	1.52	1.59	1.59	1.59	1.59
	test		2.35	2.36	2.36	2.37
TC Selten's index	training	0.94	0.94	0.94	0.94	0.94
	test		0.72	0.72	0.72	0.72
	all	0.94	0.94	0.94	0.94	0.93

References

Adams, Abi and Ian Crawford (2015). "Models of Revealed Preference." In *Emerging Trends in the Social and Behavioral Sciences*, pp. 1–15. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Aguiar, Mark and Erik Hurst (2007). "Life-Cycle Prices and Production." *American Economic Review*, 97(5), 1533–1559.

Aguiar, Victor and Roberto Serrano (2017). "Slutsky matrix norms: The size, classification, and comparative statics of bounded rationality." *Journal of Economic Theory*, 172, 163 – 201.

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." Working Paper 20618, National Bureau of Economic Research.

Andreoni, James, Ben Gillen, and William T. Harbaugh (2013). "The power of revealed preference tests: ex-post evaluation of experimental design.", URL https://econweb.ucsd.edu/~jandreon/WorkingPapers/GARPPower.pdf. Working paper.

Apesteguia, Jose and Miguel A. Ballester (2015). "A Measure of Rationality and Welfare." *Journal of Political Economy*, 123(6), 1278–1310.

Beatty, Timothy and Ian Crawford (2011). "How Demanding Is the Revealed Preference Approach to Demand?" *American Economic Review*, 101(6), 2782–95.

Benkert, Jean-Michel and Nick Netzer (2018). "Informational Requirements of Nudging." *Journal of Political Economy*, 126(6), 2323–2355.

Bernheim, B. Douglas (2009). "Behavioral Welfare Economics." *Journal of the European Economic Association*, 7(2-3), 267–319.

Bernheim, B. Douglas (2016). "The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics." *Journal of Benefit-Cost Analysis*, 7(01), 12–68.

Bernheim, B. Douglas, Andrey Fradkin, and Igor Popov (2015). "The Welfare Economics of Default Options in 401(k) Plans." *American Economic Review*, 105(9), 2798–2837.

Bernheim, B. Douglas and Antonio Rangel (2009). "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics." *The Quarterly Journal of Economics*, 124(1), 51–104.

Blundell, Richard, Martin Browning, and Ian Crawford (2003). "Nonparametric Engel Curves and Revealed Preference." *Econometrica*, 71(1), 205–240.

Boccardi, Maria Jose (2018). "Power of Revealed Preference Tests and Predictive (Un)Certainty.", URL https://drive.google.com/file/d/0B-SXcE9wvakcUmNmdFY5Wm54QU0/view. Working paper.

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

Caplin, Andrew (2016). "Economic Data Engineering.", URL https://www.newyorkfed.org/medialibrary/media/research/conference/2016/woodford/caplin_andrew_paper. Working Paper.

Caplin, Andrew and Daniel Martin (2020). "Framing, Information, and Welfare." Tech. rep., SSRN, URL http://dx.doi.org/10.2139/ssrn.3124194.

Chambers, Christopher P. and Takashi Hayashi (2012). "Choice and individual welfare." *Journal of Economic Theory*, 147(5), 1818–1849.

Choi, Syngjoo, Raymond Fisman, Douglas Gale, and Shachar Kariv (2007). "Consistency and Heterogeneity of Individual Behavior under Uncertainty." *American Economic Review*, 97(5), 1921–1938.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman (2014). "Who Is (More) Rational?" *American Economic Review*, 104(6), 1518–50.

Costa-Gomes, Miguel, Carlos Cueva, Georgios Gerasimou, and Matus Tejiscak (2019). "Choice, Deferral and Consistency." Discussion Paper Series, School of Economics and Finance 201416, School of Economics and Finance, University of St Andrews, URL https://ideas.repec.org/p/san/wpecon/1416.html.

De Clippel, Geoffroy and Kareen Rozen (2020). "Bounded rationality and limited datasets.", URL https://www.brown.edu/Departments/Economics/Faculty/Geoffroy_deClippel/lim-data.pdf. Working paper.

Dean, Mark and Daniel Martin (2016). "Measuring Rationality with the Minimum Cost of Revealed Preference Violations." *Review of Economics and Statistics*, 98(3), 524–534.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue (2002). "Time Discounting and Time Preference: A Critical Review." *Journal of Economic Literature*, 40(2), 351–401.

Gul, Faruk and Wolfgang Pesendorfer (2009). "A Comment on Bernheim's Appraisal of Neuroeconomics." *American Economic Journal: Microeconomics*, 1(2), 42–47.

Handbury, Jessie, Tsutomu Watanabe, and David E Weinstein (2013). "How Much Do Official Price Indexes Tell Us about Inflation?" Working Paper 19504, National Bureau of Economic Research, URL http://www.nber.org/papers/w19504.

Hinnosaar, Marit (2016). "Time inconsistency and alcohol sales restrictions." *European Economic Review*, 87, 108 – 131.

Koo, Anthony Y. C. (1963). "An Empirical Test of Revealed Preference Theory." *Econometrica*, 31, 646.

Manzini, Paola and Marco Mariotti (2006). "Two-Stage Boundedly Rational Choice Procedures: Theory and Experimental Evidence." Discussion paper 2341, IZA Institute for Labor Economics, URL https://ssrn.com/abstract=937874.

Manzini, Paola and Marco Mariotti (2014). "Welfare economics and bounded rationality: the case for model-based approaches." *Journal of Economic Methodology*, 21(4), 343–360.

Masatlioglu, Yusufcan, Daisuke Nakajima, and Erkut Y. Ozbay (2012). "Revealed Attention." *American Economic Review*, 102(5), 2183–2205.

Nishimura, Hiroki (2018). "The transitive core: Inference of welfare from nontransitive preference relations." *Theoretical Economics*, 13(2), 579–606.

Ok, Efe A. (2002). "Utility Representation of an Incomplete Preference Relation." *Journal of Economic Theory*, 104(2), 429 – 449.

Rubinstein, Ariel and Yuval Salant (2012). "Eliciting Welfare Preferences from Behavioural Data Sets." *The Review of Economic Studies*, 79(1), 375–387.

Salant, Yuval and Ariel Rubinstein (2008). "(A, f): Choice with Frames." *The Review of Economic Studies*, 75(4), 1287–1296.

Schwartz, Thomas (1976). "Choice functions, "rationality" conditions, and variations on the weak axiom of revealed preference." *Journal of Economic Theory*, 13(3), 414–427.

Selten, Reinhard (1991). "Properties of a measure of predictive success." *Mathematical Social Sciences*, 21(2), 153–167.

Varian, Hal R. (2006). "Revealed Preference." In *Samuelsonian Economics and the Twenty-First Century*, chap. 7. Oxford University Press.