MEASURING RATIONALITY WITH THE MINIMUM COST OF REVEALED PREFERENCE VIOLATIONS

Mark Dean and Daniel Martin*

Abstract—We introduce a new measure of how close a set of choices is to satisfying the observable implications of rationality and apply it to a large, balanced panel of household level consumption data. This new measure, the minimum cost index, is the minimum cost of breaking all revealed preference cycles found in choices from budget sets. Unlike existing measures of rationality, it responds to both the number and severity of revealed preference violations.

I. Introduction

ARGUABLY the most pervasive assumption in economics is that agents are 'rational' in the sense that they make choices as if they are maximizing some stable underlying utility function. Since the pioneering work of Samuelson (1938), Houthakker (1950), and Richter (1966), the necessary and sufficient condition for choices to be consistent with utility maximization have been well known: the preference relations revealed by choice must be acyclic, excluding cycles of indifference.

As Afriat (1967) and Varian (1982) showed, acyclicality provides a simple, elegant, and nonparametric way of testing whether a finite set of choices is in line with utility maximization. Unfortunately, it provides no information as to whether choices that contain revealed preference cycles are close to being rational. A single mistaken choice is enough to declare an entire data set incompatible with rationality, even if all other choices could be explained as resulting from utility maximization. In practice, most choice data sets contain some revealed preference cycles.[1] In order to gauge the extent of these violations, a number of goodness-of-fit measures for rationality have been proposed.[2]

In this paper, we propose a new measure of goodness of fit, the minimum cost index (MCI). This index is the minimum cost of breaking all revealed preference cycles in a data set, where the cost of removing a relation is determined by the money metric (Varian, 1982). Under this metric, if bundle x is chosen when y is available for \$100 less, the revealed preference for x over y is stronger than if x was chosen when y was available for \$1 less. While existing measures have appealing computational properties or intuitive interpretations, MCI's advantage is that it responds to both the number and severity of revealed preference violations. This is not the case for measures proposed by Afriat (1973), which assesses only the cost of the largest violation; Houtman and Maks (1985) which counts only the number of violations; and Echenique, Lee, and Shum (2011), which looks only at the cost of the mean or median violation. Apesteguia and Ballester (2015) discuss in detail the problems with existing indices. We discuss the relationship between our index and those currently in the literature in section IIA.

We also introduce an algorithmic method for calculating MCI and related indices.[3] Like many other goodness-of-fit measures, the calculation of MCI is an NP-hard problem, meaning that the solution time can grow quickly with the number of observations.[4] Our algorithm takes advantage of the fact that calculating MCI is equivalent to solving a minimum set covering problem (MSCP), which is well studied in the computer sciences and operations research literature. While MSCP is equally complex from a theoretical perspective, this is useful in calculating MCI because of the wide variety of methods that are extremely efficient in solving MSCP for practical cases and are included in standard solver software packages (see Caprara, Toth, & Fischetti, 2000).

We use our algorithm to calculate MCI for grocery scanner data of the type considered by Aguiar and Hurst (2007).[5] We find that while the cost of removing all cycles in the consumption data is small (on average 0.08% of a household's total expenditure), this is also true of uniform random choice (0.43% of total expenditure). We treat the difference between these two values (which we call the Selten score) as an additional goodness-of-fit measure that takes into account the predictive power of MCI. This approach is related to the predictive power adjustment proposed by Beatty and Crawford (2011).

Received for publication January 10, 2014. Revision accepted for publication February 4, 2015. Editor: Philippe Aghion.

* Dean: Brown University; Martin: Paris School of Economics.

This paper subsumes parts of our earlier working papers, "How Rational Are Your Choice Data?" and "Testing for Rationality with Consumption Data." We are grateful to Ian Crawford, Federico Echenique, Stefan Hoderlein, Martijn Houtman, Shachar Kariv, Hiroki Nishimura, Krishna Pendakur, Jesse Perla, Debraj Ray, Michael Richter, Joerg Stoye, Juan Carlos Suarez Serrato, Alistair Wilson, and participants in the CIREQ/CEMMAP seminar for helpful comments and to Adam Sachs and John McNeill for research assistance.

A supplemental appendix is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/REST_a_00542.

[1] See Koo (1963), Varian (1982), Famulari (1995), Andreoni and Miller (2002), Choi et al. (2007), Cherchye et al. (2008), Beatty and Crawford (2011), Echenique, Lee, and Shum (2011), and Crawford and Pendakur (2013).

[2] Several of the proposed methods will be discussed below, but see Koo (1963), Afriat (1973), Houtman and Maks (1985), Varian (1993), Famulari (1995), Beatty and Crawford (2011), Echenique et al. (2011), and Apesteguia and Ballester (2015). Varian (2006) and Cherchye et al. (2009) provide excellent reviews.

[3] See Choi et al. (2014) for an application of our approach to the index of Houtman and Maks (1985). Our approach cannot be used to calculate Echenique, Lee, and Shum's money pump index (MPI), but Cherchye et al. (2013) suggest variants of MPI that can be calculated in polynomial time.

[4] The complexity of various measures is discussed in Smeulders et al. (2014).

[5] This data set records the prices and quantities of all packaged food and beverage purchases made in any grocery store, convenience store, discount store, or drugstore for a sample of 977 households over a period of 24 months.

The Review of Economics and Statistics, July 2016, 98(3): 524–534

doi:10.1162/REST_a_00542

We find significant differences in the Selten score between demographic groups.[6] Perhaps our most interesting findings are that households of retirement age are more rational than younger households and that households with more than one "head of household" are more rational than those with one. The former finding may make sense in light of the findings by Aguiar and Hurst (2007), who show that seniors invest more time and effort in shopping than younger households do. The latter finding is puzzling in light of the fact that aggregation of preferences at the household level can lead to irrational choice (e.g., Cherchye et al., 2008).

Our results demonstrate the importance of considering predictive power when comparing behavior across demographic groups. For instance, while we find a significant relationship between the age of household heads and the Selten score, the relationship disappears when the dependent variable is the raw MCI value. This suggests that differences in irrationality can be obscured by differences in the ability of data sets to expose irrational behavior.

We also find that predictive power is largely unaffected by choice of price index or time period, but is significantly affected by the degree of product aggregation. The consumers in our choice set are actually less rational on average than a benchmark of uniform random choices if we place products into 38 product categories instead of 3 product categories (e.g., "soda" and "milk" instead of just "beverages"). When using more detailed product categories, a random choice benchmark based on actual choices, as proposed by Andreoni, Gillen, and Harbaugh (2013), provides (in our view) a more sensible predictive power adjustment, as uniform random choice provides unrealistically few revealed preference violations.

Section II describes the minimum cost index in detail. Section III applies our measure to the consumption data. Section IV discusses the related literature.

II. A New Measure of Rationality

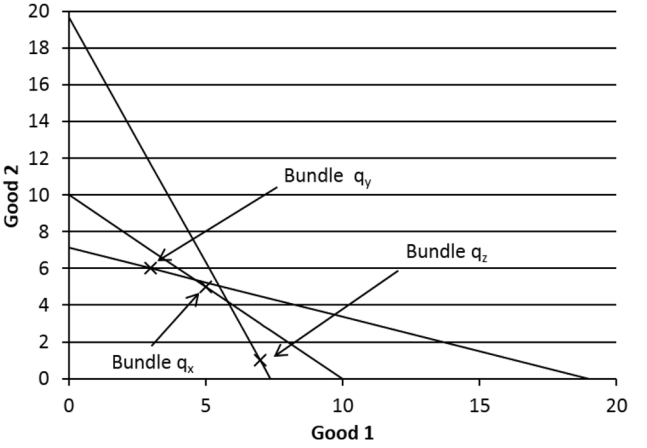
In this paper, we introduce, and then implement, a new measure of rationality for choices from budget sets. This measure, the minimum cost index, is the lowest-cost way of breaking all revealed preference cycles in a data set, divided by total expenditure. We say a revealed preference cycle is broken if a revealed preference relation is removed from that cycle and the cost of removing a revealed preference relation is measured by the monetary difference between the chosen and nonchosen bundle that generated the relation.

Formally, a data set $S = \{(p_t; q_t) \,|t = 1, ..., T\}$ is composed of T observed choices from budget sets, where p_t is the vector of prices and q_t is the vector of quantities (bundle) for observation t. The revealed preference relation R_0 is defined using the standard revealed preference concept: x is revealed preferred to y if y was affordable when x was chosen. Thus, for $x, y \in T$, $q_x R_0 q_y$ if $p_x q_x \geq p_x q_y$.

TABLE 1.—EXAMPLE DATA

	Prices		Quantities		Cost of bundle		
Observations	Good 1	Good 2	Good 1	Good 2	q_x	q_y	q_z
x	5	5	5	5	50	45	40
y	3	8	3	6	55	57	29
z	8	3	7	1	55	42	59

FIGURE 1.—BUDGET SETS FOR EXAMPLE 1



Definition 1. *For a data set S, the minimum cost index (MCI) is defined as*

$$W = \min_{B \subset R_0} \frac{\sum_{(x,y)\in B} p_x \left(q_x - q_y\right)}{\sum_{x=1}^{T} p_x q_x},$$

such that R_0/B is acyclic.

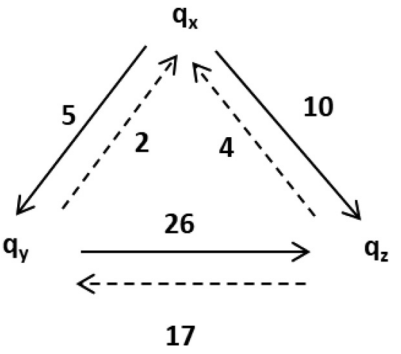
In this definition, B is the set of removed relations, and $p_x \left(q_x - q_y\right)$ is the cost of removing relation $q_x R_0 q_y$.[7] The following example illustrates MCI.

Example 1. Consider the (fictitious) data set described in table 1, consisting of observed choices of two-good bundles under different prices. The data set consists of three observations, so $T = \{x, y, z\}$. The prices in effect for each observation are given in the second and third columns of table 1, while the observed quantities purchased are given in the next two columns. Figure 1 shows the implied budget sets and chosen bundles for each observation.

[6] It should be noted that while we describe households that do not behave as if they are maximizing a stable utility function as being irrational, this is really a linguistic shortcut. There are many reasons that a household may have revealed preference cycles that are perfectly sensible—preference shocks, for example. Essentially, we are checking whether consumption choices can be modeled as resulting from maximizing a single, stable utility function, not whether departures from this model are rational.

[7] An alternative approach would be to consider removing observations rather than revealed preference relations. A difficulty with this approach is how to assign a cost to removing an observation. One natural way of doing so would be the cost of all revealed preference violations generated by an observation, but that risks applying a high cost to an observation even if it generates only low-cost revealed preference relations.

FIGURE 2.—REVEALED PREFERENCE INFORMATION AND ASSOCIATED COSTS FOR EXAMPLE 1



In order to extract the revealed preference information from this data set, we need to calculate the cost of each bundle q_x, q_y, and q_z under each set of prices p_x, p_y, and p_z. This is shown in the last three columns of table 1. In this example, each bundle is revealed preferred to all the others at the price it is purchased. This revealed preference information, along with the cost associated with each relation, is shown in figure 2. This cost is calculated as the cost difference between the chosen bundle and the unchosen bundle at the prices of the chosen bundle.

Clearly, these data have several cycles: $q_xR_0q_yR_0q_x$, $q_xR_0q_zR_0q_x$, $q_yR_0q_zR_0q_y$, $q_xR_0q_yR_0q_zR_0q_x$, and $q_zR_0q_yR_0q_xR_0q_z$. There are also several ways to remove revealed preference relations in order to render the remainder acyclic. The minimum-cost way of doing so is to remove the dashed relations in figure 2. MCI for these data is therefore equal to the total cost of the removed relations ($2 + 4 + 17 = 23$) over total expenditure (166), or 0.14.

The MCI measure is equal to 0 if and only if the data set satisfies the strong axiom of revealed preference (SARP). The index takes a high value when there is a large number of cycles for which all revealed preferences violations are based on significant monetary differences relative to total expenditure.

The motivation for this approach is to develop a measure that is responsive to both the severity and the number of revealed preference violations. In assessing severity, we assume that if a cycle can be broken by removing a low-cost relation, then it is less serious than if a high-cost relation has to be removed, because the decision maker has "thrown away" less money on this mistake. In example 1, we consider the cycle generated by the choice of q_x over q_y and q_y over q_x (which can be broken at a cost of 2) to be less serious than the choice of q_x over q_z and q_z over q_x (which can be broken at the cost of 4). As a result, the MCI does not allow a single small mistake to overly influence our impression of a consumer's rationality.

At the same time, our measure is based on the total cost of removing all cycles, and so it is also responsive to the number of mistakes that the decision maker makes. If, in example 1, we replaced observation x with x' such that $q_{x'} = \{7, 7\}$ (while p_x is still equal to $\{5, 5\}$) then it would no longer be the case that either q_y or q_z is revealed preferred to q_x. Thus, the only remaining cycle in the data would be $q_yR_0q_zR_0q_y$, which could be broken by removing the observation that q_z is revealed preferred to q_y. The MCI would be consequently lower for this new data set, falling from 0.14 to 0.09.

One issue with MCI is computational complexity. In the very simplistic case of example 1, there are 6 ways of removing one relation; 15 ways of removing two relations; and 20, 15, 6, and 1 ways of removing three, four, five, and six relations, respectively. This gives 63 different possibilities to be checked to identify the lowest-cost way of rendering the remaining data acyclic. This number increases rapidly with the number of revealed preference observations. With only twenty observations there are over 1 million combinations to check, and with forty, there are more than 1×10^{12} combinations. Brute-force attempts to calculate the MCI quickly become impractical as the number of observations increase.

To address the computational issue, we introduce an algorithm for calculating MCI that is based on finding the size of the largest subset of a choice data set that is consistent with acyclicality (detailed in the online appendix to this paper). This problem, which we call the maximal acyclical set problem (MASP), is NP-hard.[8] The key to our approach is to take advantage of the fact that MASP is equivalent to the minimum set covering problem (MSCP),[9] which is well studied in the computer sciences and operations research literature. While MSCP is also NP-hard, many method are extremely efficient in solving it for practical cases and are included in standard solver software packages (see Caprara et al., 2000). For any choice data set, we can therefore translate the associated MASP into an equivalent MSCP, which can then be solved using one of these software packages.[10] As we detail in the appendix, our approach can be applied to existing rationality measures in the literature, such as that proposed by Koo (1963) and Houtman and Maks (1985).

The MCI measure can be applied in a wide range of settings in which consumers choose from budget sets. For example, it can be applied in the case of kinked budget sets or indivisible goods. It can also be generalized to other choice settings (beyond choices from budget sets) in which a set of observations gives rise to a set of revealed preference relations. In order to apply the index, it is necessary to select a weighting function that indicates the cost of a particular revealed preference relation. A simple implementation would be to apply the same weight to every revealed preference relation.

[8] This means that there is no known algorithm with solution times that are certain to only increase polynomially with the number of choices or revealed preference relations.

[9] As shown in Garey and Johnson (1979).

[10] Off-the-shelf algorithms for solving MSCP are included in many software packages that perform optimization (such as Matlab). More powerful solvers are available for free over the Internet (such as SCIP, GLPK, and MINTO) or are available commercially (such as CPLEX).

TABLE 2.—COMPARISON OF MEASURES

Measure	Violation Severity	Violation Aggregation	Complexity
Minimum cost index	Minimum cost	Sum	NP
Afriat efficiency index	Minimum cost	Maximum	P
Houtman-Maks index	None	Sum	NP
Money pump index	Total cost	Median or mean	NP

A. *Other Measures of Rationality*

The literature includes many other measures of how close a data set is to satisfying rationality. In this section, we discuss MCI relative to three others: the Afriat efficiency index (proposed by Afriat, 1973), the Houtman-Maks index (proposed by Koo, 1963, and Houtman & Maks, 1985), and the money pump index (proposed by Echenique et al., 2011). A summary of these comparisons is presented in table 2.

One of the earliest and most widely applied measures of rationality was provided by Afriat (1973), using the concept of "revealed preferred at efficiency level e": if bundle x is chosen when y was available at a fraction e of the cost of x, then x is preferred to y at efficiency level e. The Afriat efficiency index (AEI) of a data set is the largest e^* such that there are no preference cycles revealed at that efficiency level. Koo (1963) and Houtman and Maks (1985) propose measuring rationality with the largest number of observations that are jointly acyclic, divided by the total number of observations, which is often called the Houtman-Maks index (HMI). A recent addition is the money pump index (MPI) of Echenique et al. (2011), which is the mean or median cost of all revealed preference cycles in a data set.

Given the existence of AEI, HMI, and MPI in the literature, our primary motivation was to develop a measure that takes into account both the number and severity of revealed preference violations that is also computationally convenient. Existing measures, such as the ones described in this section, focus on either the number or the severity of revealed preference violations.

The first dimension on which we compare measures is how or whether they account for the severity of revealed preference violations (column 2 of table 2). MCI accounts for the severity of a revealed preference violation in a manner similar to AEI: the cost of a revealed preference cycle is the weakest link in that cycle. The motivation for this approach is that cycles with low-cost relations can be interpreted as small "mistakes," which would disappear with a small perturbation of the budget set.

MPI accounts for the severity of a revealed preference violation differently from AEI and MCI. With this measure, the cost of a revealed preference cycle is the total cost of all links in a cycle. The difference is illustrated in the following example. Say that bundle x was chosen when y was available for \$10 less, y was chosen when z was available for 5¢ less, and z was chosen when x was available for 5¢ less. MPI would say the cost of this cycle was \$10.10, whereas MCI would say the cost was 5¢. Thus, MPI and MCI extract somewhat different information about preference violations.

HMI does not account for the severity of a violation; instead, it weights all violations equally. As such, it would treat a data set that contained only observations x and y from example 1 as equivalent to one that contained only x and z. We have argued above that the latter exhibits a worse violation of rationality.

The second dimension on which we compare measures is how they aggregate revealed preference violations (column 3 of table 2). MCI takes into account the number of revealed preferences violations as it sums the cost of all violations (HMI is similar in this regard). The impact of a single preference cycle therefore falls with the amount of data observed. AEI looks only at the worst violation of rationality, ignoring all others. Thus, a single bad choice can make AEI arbitrarily small, even if other violations are not large or do not exist. Furthermore, it does not take into account the number of violations. For example, the change from observation x to x' described in section 2 would not affect the AEI, despite the fact than the observations $\{x, y, z\}$ generate many more revealed preference cycles than $\{x', y, z\}$. In both cases an e of 0.71 is needed to remove the cycle $q_y R_0 q_z R_0 q_y$.

MPI looks at the mean or median cost of revealed preference cycles. This means that the cost of a single severe violation will be mitigated if the data contain a number of smaller violations. However, this is not the case if the data contain no other cycles. Moreover, the measure is invariant to the number of similarly sized cycles, so a data set with a single cycle will receive the same MPI score as one with many cycles of similar severity. To illustrate a potential issue with this approach, note that the observations $\{x', y, z\}$ generate a higher MPI than $\{x, y, z\}$, as the average cycle is more costly in the former case than the latter (the former only contains the most expensive cycle of the latter).[11] However, the former has just one revealed preference cycle, while the latter has five.

A third dimension on which we compare measures is computational complexity (column 4 of table 2). AEI is widely used in part because it is very easy to calculate. In fact, Smeulders et al. (2014) show that AEI can be solved in polynomial time. While calculating MCI is NP-hard, it reduces to an MSCP as described above, meaning that we can make use of off-the-shelf solvers to calculate the index. Computational constraints are particularly important when using power measures based on random choice, which can lead to very irrational data. Like MCI, HMI suffers from computational complexity. However, the algorithm we present in this paper can also be used to quickly calculate HMI (see Choi et al., 2014, for an application of our algorithm to HMI).

Echenique et al. (2011) get around the computational complexity of MPI by approximating their measure by using

[11] It should be noted that this is not per se a difference between the MCI and MPI approaches. One could consider an index based on the total cost of MPI violations.

only cycles of shorter lengths and by other methods. More recently, Cherchye et al. (2013) provided algorithms that can calculate the cost of the most severe and least severe preference cycles in the data. While these measures can be calculated in polynomial time, they cannot in principle provide information on the number of violations in the data.

Finally, it should be noted that one significant advantage of MPI over the MCI is that it can be interpreted as a statistical test based on some underlying measurement error on prices. Gross (1995) implements AEI as a statistical test by using bootstrapping to generate standard errors.

Apesteguia and Ballester (2015) provide axiomatic foundations for several classes of rationality indices. The most general class is the general weighted index, which they characterize with four axioms: continuity, rationality, concavity, and piecewise linearity. They show that HMI falls into this class but not AEI or MPI. As defined previously, our MCI is also not a general weighted index. However, if we did not divide by total expenditure, then our index would fall into this class.

B. Predictive Power

One issue with all rationality measures is that it can be hard to interpret what a raw value tells us about the underlying data. For example, consider a data set in which we observe choices from budget lines that never intersect. In this case, all the rationality measures discussed in the previous section would report a perfect score for any observed pattern of choice. In other words, such a data set does not offer a meaningful opportunity to measure rationality.

Beatty and Crawford (2011) address this issue by using Selten's adjustment for predictive success (Selten, 1991) to determine how successful the standard utility maximization model is in explaining a set of choice data. For each household, they determine all possible choices (from that household's budget sets) that would satisfy GARP, which they call a household's target area. For a household i, they then compare a_i, the size of the target (target area divided by total area), to a measure of rationality r_i. This measure r_i can be a $\{0,1\}$ indicator of whether the data for household i satisfied GARP or a measure of goodness of fit based on the Euclidean distance from the observed data and the target area.

One way to interpret a_i (the size of the target) is the percentage of times a consumer would pass GARP by choosing budget shares at random from a uniform distribution over possible budget shares. Through this lens, the Beatty and Crawford (2011) approach can be seen as a way to compare households to a benchmark consumer that chooses uniform randomly over the outcome space. Although uniform random choice is a relatively weak comparator, it is applicable to almost any choice setting so is widely used. The role of random choice in determining the statistical power of rationality measures is discussed by Bronars (1987).

In this paper, we use two methods to assess predictive power, each based on the distribution of goodness-of-fit values generated by a consumer faced with the same budget sets we observe in the data but chose budget shares at random. Our primary method is to subtract the average value of the goodness-of-fit measure for this comparison benchmark consumer from the actual value of goodness-of-fit measure for each household. We refer to this predictive power adjustment as the Selten score. This adjustment differs slightly from the one proposed by Beatty and Crawford (2011) in that we subtract off the average value of the goodness-of-fit measure for the comparison benchmark consumer while they subtract off the average pass rate for GARP for the comparison benchmark consumer. A second predictive power assessment we use is to read off the percentile that a household's goodness-of-fit measure falls in the distribution of goodness-of-fit measures for the comparison benchmark consumer.

While we base most of our analyses on the uniform random choice benchmark, we also perform a robustness check using a benchmark in which budget shares are drawn at random from the observed distribution of budget shares across all households and budget sets, which is similar to an approach taken by Andreoni et al. (2013). Thus, with this alternative benchmark, the benchmark consumer has an equal chance of choosing any observed budget shares rather than any feasible budget share.[12]

III. Measuring Rationality in Scanner Data

We now apply the minimum cost index and the Selten score to a set of consumption data. We analyze purchases of packaged foods and beverages for a balanced panel of 977 representative households in the Denver metropolitan area over two years (February 1993–February 1995). These records are derived from the data set used in Aguiar and Hurst (2007), in which participating households document the universal product code (UPC), price, date, store, and shopper characteristics for all packaged grocery products purchased across retail outlets. In addition, households maintain detailed demographic information that is updated annually (see the appendix in Aguiar & Hurst, 2007, for a more complete description of the data).

From the initial data set, we restrict our attention to purchases that we can place in three food and beverage product categories: beverages, meals, and snacks. This covers 384,964 beverage purchases, 307,391 meal purchases, and 132,499 snack purchases. We exclude products that do not have units that can be converted to ounces, which eliminates 8,156 beverage purchases.

From this data set, which includes purchases from 2,100 households, we keep households that participated for the entire 24-month period and had at least one purchase every month. For the remaining 977 households, we have an

[12] Alternatively, we could have generated a distribution of possible index values for a given choice environment using other error models or decision rules, as in Choi et al. (2007).

TABLE 3.—SUMMARY OF DEMOGRAPHIC CHARACTERISTICS

Category	Subcategory	Percentage
Age of household head(s)	<35	17
	≥ 35 & < 65	67
	≥ 65	16
Household composition	1–2 members	42
	3–4 members	41
	>4 members	17
Number of household heads	1 head	20
	2 heads	80
Household income	< $20,000	10
	≥ $20,000 and $45,000	39
	≥ $45,000	51
Education of household head(s)	No degree	5
	High school	65
	College	30

average of 20.5 purchases, 7 store trips, and $51.60 in expenditure per month. Table 3 summarizes the demographics of our sample households.[13]

We aggregate purchases at the monthly level to alleviate concerns about the fact that some items are storable. To construct monthly price indexes for each of the three product categories, we weight the mean price per UPC by the expenditure on that UPC in a given month,

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

where P_{Jt} is the price index for good category J in month t, w_{it} is the budget share for UPC code i in month t, and p_{it} is the mean price for UPC code i in month t. This approach is our baseline index, but we also examine three others in section IIIA. One limitation to all four indices is a potential bias from products entering and leaving the market (see Erickson & Pakes, 2011).

It should be noted that for utility maximization to imply acyclicality of revealed preference relations in this data set requires further assumptions. For example, food and beverages must be weakly separable from utility for other goods and services. This assumption is strong but standard in the literature (see, e.g., Echenique et al., 2011).

In addition, we follow much of the applied revealed preference literature in assuming that all households face the same prices and that these prices are constant within a period. Because prices do vary within the period and among stores, this assumption could lead to errors in the measurement of rationality. If a household faces prices that differ from imputed prices in an irregular fashion, then the household's true MCI could be smaller or larger than its computed MCI. However, if a household faces consistently lower (or higher) prices (of an equal percentage) across all products in all periods, then the true MCI would be exactly the same as the computed MCI. This is because the set of lowest-cost removals would be the same and the proportional decrease in the cost of the removal set would be the same as the proportional decrease in total expenditure.

Regardless, it is necessary to assume a single price because not all goods were bought in all periods by all households, even with just three product categories, and if a price is missing in a month, then it is not possible to do standard revealed preference testing.[14] Thus, if we had chosen to use a household-specific price index, it would have restricted our attention to households with complete price information for the entire period, resulting in a loss of almost 85% of households.[15] Looking only at this group would also have biased our predictive power assessment because these households never purchase at the corners of the budget set.

[13] When there are two household heads, "education of household head(s)" is the average education among household heads and "age of household head(s)" is the average age among household heads.

TABLE 4.—SUMMARY STATISTICS FOR THE MINIMUM COST INDEX AND AFRIAT EFFICIENCY INDEX

	Minimum Cost Index	Afriat Efficiency Index
Number of households	977	977
Perfectly rational	29%	29%
Value	0.08% (0.13%)	0.99 (0.02)
Selten score	−0.35% (0.25%)	0.03 (0.02)
Percentile	84 (21)	82 (23)

Standard deviations are in parentheses.

A. *Are Households Rational?*

Our first task is to calculate the minimum cost index for each household in our baseline data set.[16] Table 4 summarizes these results. "Perfectly rational" reports the proportion of households whose data generate no preference cycles, "value" is the average index value across households, "Selten score" reports the average difference between the index value for each household and the average of 100 simulated index values from a population that chooses at random from the same budget sets faced by that household, and "percentile" is the average across households of the percentile rank of simulated values of the index that are equal to each household's actual value.

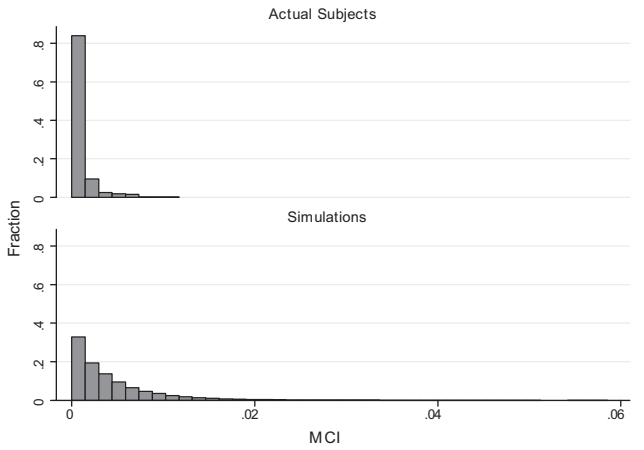
The summary statistics reported in table 4 suggest that in absolute terms, the behavior of households in our baseline sample is close to that of the paradigmatic rational agent. While only 29% of households have choices that are perfectly in line with rationality, the average cost of preference relations that need to be removed to make the data set consistent with rationality is very small—around $0.34, or about 0.08% of total expenditure. There is significant variation across households in their absolute degree of rationality: the

[14] Alternatively, we could use the approach detailed in Blow, Browning, and Crawford (2008), in which the standard GARP test is weakened to allow for missing price data, but this test would not allow us to measure the degree of violation for a household.

[15] We also attempted to create a price index for each household by using average prices in the stores where each household made its purchases. However, the households are spread across Denver in such a way that there is little overlap in the stores visited.

[16] In doing so, we implicitly treat each household as a single entity. Alternatively, Cherchye et al. (2008) implement a test of collective household consumption.

FIGURE 3.—DISTRIBUTION OF MINIMUM COST INDEX (MCI) VALUES IN THE BASELINE POPULATION (TOP) AND A SIMULATED POPULATION OF RANDOM CHOOSERS (BOTTOM)



MCI is measured as percentage of total expenditure.

maximum removal needed is around \$4.71, or about 1% of total expenditure. The top panel of figure 3 shows the distribution of MCI values in our sample population.

These raw values, however, tell us little about whether these results should be considered as providing strong support for the model of rationality. So far we know nothing of the predictive power of the measure on this data set, which is its ability to identify irrational choice. In order to answer this question, we employ the Selten score, a predictive power adjustment introduced in section IIB. We calculate this score at the household level using the 24 budget sets actually faced by that household. For each budget set, we generate a choice by drawing a bundle from a random distribution over budget shares. We then calculate MCI for these 24 simulated choices. This procedure is repeated 100 times to create a distribution of index values generated under random choice for that particular household. For a given index and household, the Selten score is the index value for the household minus the mean index value of the simulated data.

The Selten score row of table 4 reports the average Selten score across households using a benchmark of uniform random choice over budget shares. This score suggests that while our consumers do on average outperform the simulated random data, they do not do so to a great degree. On average, our simulated random choosers required removals totaling only around 0.43% of total expenditure to achieve rationality, giving an average Selten score of −0.35%. A comparison between the behavior of random choosers and that of our baseline population can be seen in figure 3. The bottom panel of the graph shows the distribution of index values from the simulated data for all households. The two distributions are statistically different from those generated by the households in our sample (at the 1% level using the Kolmogorov-Smirnov test).

While the data generated by our households seem close to rationality, random choosers also look relatively close to rationality, a result consistent with that of Beatty and Crawford (2011) and Echenique et al. (2011). Of course, it could be that while the mean index values for the simulated random choosers seem close to those of those of the actual households, it is very unlikely for random choosers to reach the values of our actual households. In other words, the variance of the index values of the simulated choosers could be small relative to the gap in means between simulated and actual values. To test this hypothesis, we calculate for each household the percentile of simulated values that the household index falls into—in other words, the proportion of simulated households that are no more rational than the actual household. The "percentile" row of table 4 reports the average of these percentiles across households. This measure supports the hypothesis that our observed households are on average only somewhat more rational than the random benchmark: the average percentile for our households is 84.

Finally, we repeated the analysis using the Afriat efficiency index. The correlation between MCI and AEI is high for our households. For raw values, the correlation is −0.8302, and for the Selten scores, the correlation is −0.7844 (there is a negative correlation because more consistency increases AEI but shrinks MCI). Thus, not surprisingly, AEI gives the same impression of the rationality of our households as does MCI: violations from rationality exhibited by our households are small in absolute terms but are not far from those exhibited by households that choose randomly. As shown in table 4, the average AEI is around 99%, meaning that there is around a 1% loss of efficiency on average, with a 4% average efficiency loss for random choosers. However, once again the average percentile is high (around 82).

Robustness checks. We provide a series of robustness checks on our results. In other words, we examine the extent to which the assumptions described at the start of section III affect our conclusions. First, we examine whether the number of product categories into which we aggregate goods matters for our results. To do so, we repeat our analysis, but rather than use the three aggregate product categories of our baseline data, we use the 38 product categories available in the data.[17] These results are shown in the second line of table 5. The value of MCI for our households is little changed (0.09%, as compared to 0.08% in the baseline data). However, the amount of rationality observed in the simulations changes dramatically: at the 38-product level, there are only very small violations of rationality in the random choice data. As a result, the household data are on average less rational than the simulated data at this level of product disaggregation, as indicated by a positive Selten score: the average cost of removing irrationality from the random data is 0.03%, giving a Selten score of 0.06%.

The lack of irrationality in the random choice data reflects the fact that with 38 product categories, the regions of the

[17] Two of the 38 good categories were dropped because they were not purchased enough to create a price index.

TABLE 5.—SUMMARY STATISTICS FOR ROBUSTNESS TESTS

	Households	Fully Rational	Minimum Cost Index	Selten Score	Percentile
Baseline	977	29%	0.08% (0.13%)	−0.35% (0.25%)	84 (21)
38 products	977	24%	0.09% (0.15%)	0.06% (0.14%)	41 (38)
2 weeks	397	<1%	0.25% (0.25%)	−0.59% (0.42%)	87 (20)
Paasche	977	44%	0.02% (0.04%)	−0.12% (0.08%)	84 (22)
Laspeyres	977	34%	0.04% (0.07%)	−0.12% (0.11%)	79 (25)
Törnqvist	977	69%	0.22% (1.02%)	−0.77% (1.05%)	85 (28)
Alternative benchmark	977	29%	0.08% (0.13%)	−0.25% (0.21%)	79 (24)
38 products + Alternative benchmark	977	24%	0.09% (0.15%)	−0.15% (0.19%)	73 (29)

Standard deviations are in parentheses.

data set that can generate violations of SARP are small (in the sense of Beatty & Crawford, 2011). So why do we not observe high levels of rationality in the actual household data? The answer appears to be that households do not buy all 38 products in each month. On average, households consumed products from just 8.5 categories in a month. Thus, they are often at a corner of the budget set. This suggests that uniform random simulations (which almost never hit the corner of the budget set) are not a suitable benchmark for a large number of product categories. We address this concern below when we use an alternative distribution to generate our comparison simulations.

As a second robustness check, we run our analysis on data temporally aggregated at the two-week level rather than the monthly level. The results are shown in line 3 of table 5. There is some evidence that data at this level contain more serious violations of rationality than does our baseline case. Almost no households are perfectly rational, and MCI is higher than in the baseline case (0.25% rather than 0.08%). However, notice that in this case, we observe 48 choices instead of 24, as in the baseline, so consumers have more opportunities to exhibit irrationality. The fact that the mean percentile is relatively similar to that of the baseline case suggests that the larger number of observations per household is driving much of the increase in absolute levels of rationality.

Next, we try a number a different price indices (Paasche, Laspeyres, and Törnqvist) and compare these to the baseline index.[18] These results are reported in lines 4 to 6 of table 5.

[18] First, we normalize the price using the first period price and again weight it by the amount purchased per month,

$$P^P_{Jt} := \left(\sum_{i\in J}(w_{it})\left(\frac{p_{i0}}{p_{it}}\right)\right)^{-1},$$

which is a Paasche index. Second, we weight the normalized price by the amount purchased in the first period (fixed basket),

$$P^L_{Jt} := \sum_{i\in J}(w_{i0})\left(\frac{p_{it}}{p_{i0}}\right),$$

which is a Laspeyres index. Third, we weight (the log of) the normalized price by an average of the weight in the first period and that period,

$$P^T_{Jt} := \sum_{i\in J}\frac{1}{2}(w_{it}+w_{i0})\ln\left(\frac{p_{it}}{p_{i0}}\right),$$

which is a Törnqvist index.

While the levels of MCI and Selten score differ across the indices, we find that for all four indices, the level of MCI is small and below the average MCI for uniform random choice. Also, the overall picture of household rationality, as measured by the average percentile rank, is similar for all four indices.

Alternative benchmark. As an additional robustness check, we use an alternative benchmark for the predictive power adjustment. Rather than using choices that are drawn from a uniform random distribution over budget shares, we draw budget shares for each category of goods from the observed distribution of shares across all households and budget sets. With this approach, implausible choices, such as spending all of a household's monthly grocery budget on ice cream, are heavily down-weighted in assessing the performance of a measure. Additionally, the mean and variance of budget shares are allowed to vary across good categories in a manner that mimics the actual data. This alternative benchmark is very similar to the bootstrapping approach of Andreoni et al. (2013) but differs in that we sample from all chosen budget shares, not just budget shares chosen from identical budget sets.[19]

The results for the baseline data set (in which products are grouped into three categories) are shown on the seventh line of table 5. They show that our consumers are somewhat less rational relative to this benchmark than to the uniformly random benchmark (on average, our households fall in the 79th percentile of the simulated distribution with the alternative benchmark, as compared to the 84th percentile with the baseline benchmark). This suggests that the empirical distribution of budget shares is concentrated in regions that are less likely to cause violations of rationality than is the uniform distribution. However, the effect is not dramatic.

On the other hand, the clustering of budget shares in the empirical distribution has a dramatic effect when considering all 38 product categories. With this benchmark, the Selten score becomes −0.15%, meaning that on average, actual index values are less than the benchmark values. Further, the mean percentile is similar to the baseline data set with the same benchmark.

[19] See Andreoni et al. (2013) for a thorough review of alternative benchmarks.

TABLE 6.—OLS REGRESSION OF MINIMUM COST INDEX ON DEMOGRAPHIC VARIABLES

	Minimum Cost Index	Selten Score (for MCI)	Selten Score (for AEI)
Age ≥35 & <65	0.001% (0.012%)	0.000% (0.021%)	0.186% (0.222%)
Age ≥65	0.002% (0.017%)	−0.051% (0.030%)*	0.435% (0.274%)
3−4 members	−0.006% (0.010%)	−0.061% (0.018%)***	0.405% (0.163%)**
>4 members	−0.004% (0.014%)	−0.072% (0.029%)**	0.386% (0.233%)*
2 heads	−0.005% (0.010%)	−0.063% (0.019%)***	0.571% (0.180%)***
≥ $20,000 and < $45,000	−0.019% (0.016%)	−0.018% (0.026%)	0.179% (0.243%)
≥ $45,000	−0.015% (0.016%)	−0.010% (0.027%)	0.100% (0.261%)
High school	−0.028% (0.030%)	−0.012% (0.048%)	0.248% (0.402%)
College	−0.008% (0.032%)	0.033% (0.050%)	−0.058% (0.422%)
N	977	977	977
F(9,967)	0.83	6.46	3.65
p-value	0.5875	<0.001	<0.001
R^2	0.0099	0.0523	0.0136

Robust standard errors in parentheses. Significant at *10%, **5%, ***1% for robust standard errors.

This result shows that classifying revealed preference axioms as demanding or undemanding according to Selten's measure can depend heavily on how the benchmark is determined. Which benchmark is more appropriate for our data set? The benchmark based on all possible choices has two potentially large disadvantages relative to the target area based on all realized choices: (a) it considers budget shares that are almost inconceivable (such as spending all income on just one category) and (b) it gives the same mean and variance to the budget shares of all good categories. There is strong evidence that the latter is not true for the households in our data set. By looking at actual budget shares, we learn that some good categories command a much larger share of a household's budget on average. For example, seventeen good categories have an average budget share of less than 1%, while three have an average budget share over 12%. In addition, the set of chosen budget shares has much higher variance than an equally sized set of random budget shares. For chosen budget shares, the average standard deviation across good categories is 5.4%, and the maximum standard deviation for any single good category is 16.0% (for milk). For an equal-sized set of random budget shares, the average standard deviation across good categories is 2.7%, and the maximum standard deviation for any single good category is just 2.8%.

It only makes sense to use realized choices as a benchmark when there are enough unique sets of budget shares to encompass a wide range of choices. Otherwise, the target is impossibly small (no likely choices consistent with revealed preference axioms) or impossibly large (all likely choices are consistent with revealed preference axioms). For our data set, out of the 23,448 chosen budget shares, there are 23,241 unique vectors of budget shares.

B. Are Some Households More Rational Than Others?

We next examine to what extent demographic variables can explain differences in the level of rationality between households. We do this by regressing MCI and the Selten score for each household on dummy variables for the demographic variables available in the data: the age and education of household heads, household income, household size, and the number of household heads. Table 6 reports the results of these OLS regressions.[20]

When the dependent variable is the raw MCI value, our results are similar to those found in Echenique et al. (2011). First, all coefficients point in same direction as those in Echenique et al. (2011). Households with heads over age 35 are less rational than those under 35, and households with more than two members are more rational than those with those with one. Households with at least a high school education are more rational than those without, and those that earn more than $20,000 are more rational than those that earn less than this amount.[21] Second, while none of our coefficients is significant, the lowest *p*-value is found on the coefficient that indicates the difference between the middle-income and low-income households (a *p*-value of 0.16), in line with Echenique et al. (2011).

The dependent variable that we are most interested in is the Selten score for each household rather than the raw MCI value. This is important because there might be systematic differences between groups in the predictive power of the rationality measures. For example, different groups might have different numbers of intersecting budget lines. This could lead to differences in the underlying raw index values that have nothing to do with differences in the rationality of these groups.

Using the Selten score, we find that households of retirement age are more rational than the youngest households, those with multiple members are more rational than those with one, and those with two heads are more rational than those with single heads. These results demonstrate the importance of using the Selten score, rather than the raw index values, to explore demographic differences. First, none of the coefficients is significant if we use raw index values. Second, the coefficient on the oldest households reverses direction, suggesting that the oldest households are more

[20] Excluded categories are the youngest and least educated household heads, the smallest and lowest-income households, and households with just one household head.

[21] Note that as in Echenique et al. (2011), these relationships are not monotonic.

rational than the youngest households and this coefficient is significant at the 10% level. This result is sensible if the additional time and effort that seniors put into shopping (as documented by Aguiar & Hurst, 2007) allows them to make more consistent choices.[22]

One interesting result from the regression analysis regards the rationality of households with a single head. A significant literature has developed to examine the conditions under which aggregation of preferences within the household can lead to irrational choices at the household level (see, e.g., Cherchye et al., 2008). Intuitively, in households with more than one head, different heads may have different preferences. Depending on how these preferences are aggregated, this may lead to cyclic choice behavior at the household level. Thus, we would expect households with a single head to be more rational than households with more than one head. In fact, looking at the third column of table 6, we find precisely the opposite: households with a single head have Selten scores that are *worse* than those of households with more than one head. This suggests either that bargaining between household heads is not an important cause of irrational choices or that some unobserved factor affects single-head households that makes them more prone to irrationality. One possibility is that households with more than one head have more time to devote to shopping and so make more rational choices.[23]

Finally, we perform the same regression analysis using AEI. As shown in last column of table 6, the broad message is similar across the measures, especially where the coefficients are significant.

IV. Comparison to Existing Empirical Literature

A small literature measures the degree of rationality in field settings.[24] Early papers made use of repeated cross-sectional data rather than panels. For example, Famulari (1995) uses repeated cross-sectional data aggregated by demographic characteristic to test the joint hypotheses of rationality and homogeneity.[25] Using the proportion of pairwise comparisons that violate GARP as a measure, Famulari finds that rationality does relatively well: only 0.7% of comparisons violate GARP, though, as the author points out, this is in part due to low power. Hoderlein (2011) takes a somewhat different approach, using techniques to control for unobserved heterogeneity to test integrability conditions using cross-sectional data from the British Family Expenditure Survey. He finds that the rationality assumption is acceptable for a large fraction of the population (see also Hoderlein & Stoye, 2014). Kitamura and Stoye (2013) extend this approach to stochastic choice.

Beatty and Crawford (2011) examine the predictive power of revealed preference tests using a panel of Spanish consumption data. For the data set they examine, the pass rate for GARP is high (95.7%), but the average target area is also large (91.2%). From this, they conclude that GARP is not very demanding.

More recently, Echenique et al. (2013) apply their money pump measure to scanner data and study the effect of demographic variables on the degree of irrationality. They find that most households violate GARP (80%), with an average money pump cost of about 6% of total expenditure. They also point out that the power of these tests is low, but come to the conclusion that this is because the random benchmark is unsuitable. They find that younger, richer, more educated, and larger households have higher rationality values. They do not consider power-adjusted measures, which we show can deliver different results.

Choi et al. (2014) collect data from field experiments on choices over lotteries from a panel of over 2,000 Dutch subjects. They find that rationality is significantly higher in subjects than under the random benchmark. They also find significant differences in rationality between demographic groups, with high-income, high-education, male, and younger subjects showing higher levels of consistency.

[22] However, the finding in Aguiar and Hurst (2007) that seniors search out lower prices challenges our assumption that all households face the same prices.

[23] When we consider the robustness checks of section IIIA, the regression results are little changed.

[24] Andreoni and Miller (2002) and Choi et al. (2007) test the degree of rationality in laboratory experiments.

[25] See also Blundell, Browning, and Crawford (2003) and Hoderlein and Stoye (2014), which make use of repeated cross sections.

REFERENCES

Afriat, Sydney, "The Construction of a Utility Function from Demand Data," *International Economic Review* 8 (1967), 67–77.

——— "On a System of Inequalities in Demand Analysis: An Extension of the Classical Method," *International Economic Review* 14 (1973), 460–472.

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

Andreoni, James, Benjamin Gillen, and William Harbaugh, "The Power of Revealed Preference Tests: Ex-Post Evaluation of Experimental Design," unpublished manuscript (2013).

Andreoni, James, and John H. Miller, "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism," *Econometrica* 70:2 (2002), 737–753.

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

Beatty, Timothy, and Ian Crawford, "How Demanding Is the Revealed Preference Approach to Demand?" *American Economic Review* 101:6 (2011), 2782–2795.

Blow, Laura, Martin Browning, and Ian Crawford, "Revealed Preference Methods for the Consumer Characteristics Model," *Review of Economic Studies* 75 (2008), 371–389.

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

Bronars, Stephen, "The Power of Nonparametric Tests of Preference Maximization," *Econometrica* 55:3 (1987), 693–698.

Caprara, Alberto, Paolo Toth, and Matteo Fischetti, "Algorithms for the Set Covering Problem," *Annals of Operations Research* 98 (2000), 352–371.

Cherchye, Laurens, Bram De Rock, Jeroen Sabbe, and Frederic Vermeulen, "Nonparametric Tests of Collectively Rational Consumption Behavior: An Integer Programming Procedure," *Journal of Econometrics* 147 (2008), 258–265.

Cherchye, Laurens, Bram De Rock, Bart Smeulders, and Frits C. R. Spieksma, "The Money Pump as a Measure of Revealed Preference Violations: A Comment," *Journal of Political Economy* 121:6 (2013), 1248–1258.

Cherchye, Laurens, Ian Crawford, Bram De Rock, and Frederic Vermeulen, "The Revealed Preference Approach to Demand," *Contributions to Economic Analysis* 288 (2009), 247–279.

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

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman, "Who Is (More) Rational?" *American Economic Review* 104:6 (2014), 1518–1550.

Crawford, Ian, and Krishna Pendakur, "How Many Types Are There? A Simple, Theory-Consistent Approach to Unobserved Heterogeneity," *Economic Journal* 123:567 (2013), 77–95.

Echenique, Federico, Sangmok Lee, and Matthew Shum, "The Money Pump as a Measure of Revealed Preference Violations," *Journal of Political Economy* 119:6 (2011), 1201–1223.

Erickson, Tim, and Ariel Pakes, "An Experimental Component Index for the CPI: From Annual Computer Data to Monthly Data on Other Goods," *American Economic Review* 101:5 (2011), 1707–1738.

Famulari, Melissa, "A Household-Based, Nonparametric Test of Demand Theory," this REVIEW 77:2 (1995), 371–382.

Garey, Michael R., and David. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (New York: Freeman, 1979).

Gross, John, "Testing Data for Consistency with Revealed Preference," this REVIEW 77:4 (1995), 701–710.

Hoderlein, Stefan, "How Many Consumers Are Rational?" *Journal of Econometrics* 164:2 (2011), 294–309.

Hoderlein, Stefan, and Jorg Stoye, "Revealed Preferences in a Heterogeneous Population," this REVIEW 96:2 (2014), 197–213.

Houthakker, Hendrick S., "Revealed Preference and the Utility Function," *Economica* 17 (1950), 159–174.

Houtman, Martijn, and Julian A. H. Maks, "Determining all Maximal Data Subsets Consistent with Revealed Preference," *Kwantitative Methoden* 19 (1985), 89–104.

Kitamura, Yuichi, and Jorg Stoye, "Nonparametric Analysis of Random Utility Models: Testing," CeMMAP working papers CWP36/13, Centre for Microdata Methods and Practice, Institute for Fiscal Studies (2013).

Koo, Anthony, "An Empirical Test of Revealed Preference Theory," *Econometrica* 31:4 (1963), 646–664.

Richter, Marcel, "Revealed Preference Theory," *Econometrica* 34:3 (1966), 635–645.

Samuelson, Paul, "A Note on the Pure Theory of Consumers' Behaviour," *Economica* 5 (1938), 61–71.

Selten, Reinhard, "Properties of a Measure of Predictive Success," *Mathematical Social Sciences* 21:2 (1991), 153–167.

Smeulders, Bart, Laurens Cherchye, Bram De Rock, and Frits Spieksma, "Goodness of Fit Measures for Revealed Preference Tests: Complexity Results and Algorithms," *ACM Transactions on Economics and Computation* 2:1 (2014), 3.

Varian, Hal R., "The Nonparametric Approach to Demand Analysis," *Econometrica* 5:4 (1982), 945–973.

——— "Goodness-of-Fit for Revealed Preference Tests," unpublished manuscript, University of Michigan (July 1993).

——— "Revealed Preference" in Michael Szenberg, Lall Ramrattan, and Aron A. Gottersman, eds., *Samuelsonian Economics and the Twenty-First Century* (New York: Oxford University Press, 2006).