Testing for Rationality with Consumption Data: Demographics and Heterogeneity*

Mark Dean† and Daniel Martin‡

June 19, 2011

Abstract

In this paper, we introduce a new measure of how close a set of choices are to satisfying the observable implications of rational choice, and apply it to a large balanced panel of household level consumption data. We use this method to answer three related questions: (i) "How close are individual consumption choices to satisfying the model of utility maximization?" (ii) "Are there differences in rationality between different demographic groups?" (iii) "Can choices be aggregated across individuals under the assumption of homogeneous preferences?" Crucially, in answering these questions, we take into account the power of budget sets faced by each household to expose failures of rationality. To summarize our results we find that: (i) while observed violations of rationality are small in absolute terms, our households are only moderately more rational than the benchmark of random choice; (ii) there are significant differences in the rationality of different groups, with multi-head households more rational than single head households, and the youngest households more rational than middle age households; (iii) the assumption of homogenous preferences is strongly rejected: choice data that is aggregated across households exhibits high levels of irrationality.

*We are grateful to Ian Crawford, Stefan Hoderlein, Hiroki Nishimura, Krishna Pendakur, Jesse Perla, Debraj Ray, Michael Richter, Joerg Stoye, Juan Carlos Suarez Serrato, Alistair Wilson, and participants in the NYU NRET and 2010 CIREQ/CEMMAP seminars for helpful comments; Shachar Kariv and Martijn Houtman for providing their computer programs; and Adam Sachs for research assistance.

†Department of Economics, Brown University. Email: mark_dean@brown.edu.

‡Center for Experimental Social Science and Department of Economics, New York University. Email: daniel.martin@nyu.edu.

1 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 work of Samuelson [1938] and Houthakker [1950], the necessary and sufficient conditions for a data set to be consistent with utility maximization have been well known: the preference relations revealed by choice must be acyclic. This condition is; however, a very demanding one. A single inconsistent choice is enough for an individual to be classified as irrational, even if most of their choices are consistent with rationality. In practice, individuals appear to exhibit some irrationality in almost every data set – from the laboratory or from the field.[1] This has lead to a concerted effort to find measures of 'goodness of fit' for the assumption of utility maximization.[2]

In this paper, we develop a novel goodness of fit measure for the assumption of rationality,[3] and apply it to a large balanced panel of household level consumption data. We use this method to answer three related questions: (i) "How close are individual consumption choices to satisfying the model of utility maximization?" (ii) "Are there differences in rationality between different demographic groups?" (iii) "Can choices be aggregated across individuals under the assumption of homogeneous preferences?" Crucially, in answering these questions, we take into account the power of budget sets faced by each household to expose failures of rationality. To summarize our results we find that: (i) while observed violations of rationality are small in absolute terms, our households are only moderately more rational than the benchmark of random choice; (ii) there are significant differences in the rationality of different groups, with multi-head households more rational than

[1] See for example Koo [1963], Varian [1982], Manser and McDonald [1988], Famulari [1995], Andreoni and Miller [2002], Choi, Gale, Fisman, and Kariv [2007], Crawford and Pendakur [2008], Beatty and Crawford [2010], and Echenique, Lee, and Shum [2010].

[2] Many of the proposed methods will be discussed briefly below, but see Koo [1963], Afriat [1967], Varian [1982], Houtman and Maks [1985], Famulari [1995], Gross [1995], Kalai, Rubinstein, and Spiegler [2002], Hoderlein and Stoye [2009], Echenique, Lee, and Shum [2010], and Apesteguia and Ballester [2011]. Varian [2006] provides an excellent review.

[3] 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 why a household may appear to violate acyclicality which are perfectly sensible – for example due to preference shocks. Moreover, the aggregation of goods into categories can introduce violations of acyclicality that were not in the original underlying data. Essentially, we are testing whether consumption choices can be modeled as resulting from utility maximization, not whether departures from this model are rational.

single head households, and the youngest households more rational than middle age households; (iii) the assumption of homogenous preferences is strongly rejected: choice data that is aggregated across households exhibits high levels of irrationality.

The rationality measure we introduce is based on one proposed by Houtman and Maks [1985]. Their measure identifies the minimum number of observed choices that need to be removed from a data set before the remaining data satisfies acyclicality. While the Houtman-Maks measure has a number of advantages,[4] it has been criticized on the basis that it counts only the *number* of violations of rationality, without taking into account the *seriousness* of these violations (Varian [1991]). We therefore modify the Houtman-Maks measure to allow for revealed preference relations to be of varying strengths, based on the cost difference between the chosen and unchosen item. Thus, 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. Our modified measure calculates the lowest cost way of removing revealed preference observations such that the remaining data satisfies acyclicality, where the cost of removing a relation is equal to the strength of that relation. We call this the HM-efficiency (or HM-e) measure of rationality.

As with all rationality measures, the raw numbers provided by our index are uninformative on their own, as they tell us nothing about the potential of a data set to demonstrate violations of rationality. Consider, for example, a data set consisting of choices from disjoint choice sets. In this case, any pattern of choice would be perfectly consistent with rationality, but the resulting 'perfect' index values would tell us nothing. We address this problem by including a novel adjustment for the power of a data set to uncover irrationality. For a given measure of rationality, we define the 'Selten-Bronars' score as the distance between the observed index values and the mean of simulated values generated under the assumption of random choice, following the approaches suggested by Bronars [1987], Selten [1991], and Beatty and Crawford [Forthcoming].

We apply our measure to a large balanced panel of consumption data. This 'Homescan' data set records the prices and quantities of all packaged food and beverage purchases made in any grocery store, convenience store, discount store, or drug store for a sample of 977 households over a period of 24 months.[5] Our results initially look promising for the hypothesis of utility maximization because

[4] Most notably its robustness to small numbers of errors, and its applicability to a wide range of data sets.

[5] Previously used by Aguiar and Hurst [2007]. Point-of-sale scanner data from a single store chain is used by Echenique et al. [2010].

deviations from full rationality seem small in absolute terms: we find that on average it is only necessary to remove revealed preference relations equal to 0.16% of a household's total expenditures to make the remaining data set consistent with acyclicality, and the maximum amount that needs to be removed for any household is just 2.68%. However, our comparison with the benchmark of random choice indicates that a seemingly small absolute index value does not necessarily imply a high level of rationality: on average, households facing the same budget sets as our households, but who chose at random from the budget line, would have an average HM-e index of 0.45%. Thus, while the households we observe are more rational that those that choose at random, the difference is not large.

We find significant differences in rationality between demographic groups. Single head households are significantly *less* rational than those with multiple household heads. This result seems to run contrary to the predictions of collective models of intra-household allocation (see Cherchye, De Rock, Sabbe, and Vermeulen [2008]). We also find some evidence that the youngest households (under 40) are more rational than middle age households (40 to 59 years old). This result is compatible with the finding Choi, Kariv, Müller, and Silverman [2011] that younger people tend to make more consistent choices.

Finally, we find strong evidence against the hypothesis that households have homogenous preferences. When we pool together the choices of 30 different households, we find that the level of rationality is reduced dramatically. Approximately 12% of expenditure has to be removed in order to make the resulting data set consistent with rationality. This is substantially *larger* than the 8.12% on average that must be removed from a pool of simulated random households. In contrast, when we apply our methods to the household consumption data used by Blundell, Browning, and Crawford [2008] (henceforth BBC)[6] we do not find strong evidence of heterogeneity. The BBC data is less finely aggregated than the Homescan data we use, which may explain the difference in results.

Our paper makes two further contributions to the literature on goodness of fit measures for rationality. First, we introduce an algorithm for calculating measures of the Houtman-Maks type. Like the Houtman-Maks measure, our new measure can be very computationally intensive.[7,8] In

[6] This data is generated from the British Family Expenditure Survey.

[7] In fact, like the Houtman-Maks measure itself, the calculation of our measure is NP-hard.

[8] See Choi et al. [2007] for a case in which this constraint is binding in a data set consisting of only 50 choices.

order to overcome this problem, we take advantage of the fact the calculation of the HM-e (and HM) index can be translated into a form that is soluble by existing numerical solvers, a method orders of magnitude more powerful than techniques currently used in economics.[9] This method can be applied to the HM-e index, the standard Houtman-Maks measure, and can be used to approximate the minimal multiple rationale measure of Kalai et al. [2002]. Second, we compare existing rationality measures to our new measure and to each other. We find a high degree of correlation between measures: in all pairwise comparisons, Spearman's rank correlation coefficient is never below 0.80. However, we find differences where we would expect. Our new measure, which shares features of both the Afriat and Houtman-Maks measures, is more highly correlated with the Afriat and Houtman-Maks measures (0.91 and 0.84 respectively), than they are with each other (0.80). The same pattern also holds true if we look at the Selten-Bronars scores for each measure. This result suggests that our new measure may provide a useful compromise between measures that only count the number of violations of rationality and those that look only at the severity of those violations.

Section 2 describes the HM-e index in detail. Section 3 applies our measure to the Homescan data. Section 4 places our work in the context of the existing literature. Section 5 concludes.

2 A Measure of Rationality

Since the pioneering work of Samuelson [1938] and Houthakker [1950], the observable implications of utility maximization have been well known: the (strict) preference relations revealed[10] by choice must be acyclic.[11] This condition provides a simple, elegant, and nonparametric way of testing

[9] Our comparators are the methods used in Choi et al. [2007] and Houtman [1995].

[10] There are different approaches to exactly what choice behavior reveals about preference. A standard assumption that x is revealed strictly preferred to y when x is chosen and y is available. When considering choices from budget sets, a weaker assumption is that x is revealed strictly preferred to y if x is chosen when y is available at a strictly lower cost.

[11] If we observe choices from subsets of some grand set Z, and the binary relation $\succ$ on Z represents the (strict) preferences revealed by those choices, acyclicality means that there exists no sequence z_1, z_2, ..., z_n in Z such that

$$z_1 \succ z_2 \succ ... \succ z_n \succ z_1.$$

If the data set is finite, this acyclicality condition guarantees the existence of a utility function on Z that represents the revealed preference relation, and so choice.

whether a data set is in line with utility maximization.[12] Unfortunately, it is of only limited use in practice, as it provides no information as to whether a data set that contains revealed preference cycles is 'close' to satisfying rationality. 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, almost all data sets contain some revealed preference cycles.[13] It is therefore important to have some measure of the extent of these violations, allowing us to compare degrees of irrationality across individuals or decision making circumstances.

We introduce a new measure of rationality which is related to one proposed by Houtman and Maks [1985]. The Houtman-Maks (henceforth HM) index is based on the smallest number of observations that need to be removed for the remaining data to satisfy acyclicality. The number reported as the 'HM index' is the largest number of observations that are jointly acyclic, divided by the total number of observations. It provides a natural measure of rationality that has two big advantages. First, it can be applied to any form of choice data (and not just choices from budget sets). Second, it is robust in the face of a small number of irrational choices (an issue we return to below). However, it also has a weakness as a measure of rationality: it looks only at the *number* of violations that need to be removed, not the *severity* of these violations. If a consumer's preference cycles only involve objects choices that are very close to indifference, or involve only small cost differences, then we may not find those violations very damning to the concept of utility maximization.

This shortcoming is easiest to illustrate in the case in which the observed choices are over bundles of commodities from different budget sets. Consider the following choice behavior for hypothetical consumers A and B from budget sets in a commodity space that contains two goods (x and y):

- Budget set 1 : income is 10, price of good x is 2, price of good y is 2
  - A buys 1 unit of good x and 4 units of good y
  - B buys 2 units of good x and 3 units of good y
- Budget set 2 : income is 10, price of good x is 3, price of good y is 1

[12] It is also easy to implement using the Floyd-Warshall algorithm.

[13] For example, Choi et al. [2007] report that 83% of subjects violated GARP, and Cherchye et al. [2008] find that 55% of households fail the (unitary) test of GARP.

- A buys 3 unit of good x and 1 unit of good y
- B buys 3 unit of good x and 1 units of good y

Figure 1 illustrates the choices of these two consumers.

FIGURE 1 ABOUT HERE

Both of these consumers violate acyclicality, as in both cases the bundle bought in budget set 2 was available in budget set 1, and vice versa. However, the 'cost'[14] of the acyclicality violation for consumer A is higher than for consumer B. For A, the bundle chosen from budget set 2 was available at a cost of \$8 from budget set 1, while the bundle chosen in budget set 1 was available for \$7. For consumer B, the bundle chosen from set 1 was available at a cost of \$9 in set 2, while the bundle chosen in set 2 was available for \$8 in set 1. One could therefore think of the minimum 'cost' the acyclicality violation for A is \$2, while for B it is only \$1. Yet both consumers would have the same HM index.

In order to address the HM index's inability to account for the severity of violations, we generalize the measure to allow for the varying 'costs' of removing different revealed preference relations, depending on an external metric for the strength of each preference. These costs are represented by a weighting function, which carries information on the strength of different revealed preference relations. In general, any arbitrary weighting function could be used – in other words the weighting function is an input to rather than an output of the index. However, as we illustrate with the above example, in the case of choice of budget sets, one natural measure would be the cost difference between the chosen and unchosen bundle. So, if bundle x was chosen when bundle y was available for \$5 less, then the weight applied to the revealed preference of x over y would be 5.

The HM-e index is defined as the lowest cost way of removing all the cycles from the data according to this weighting function, divided by total expenditure.[15] Thus, the HM-e index bridges the gap between rationality measures that only count the *number* of violations of rationality and those that look only at the *cost, severity, or seriousness* of such violations (Afriat [1972] and Varian [1991]).

[14] Here cost can be thought of as a potential 'money pump' or as 'wasted' income.

[15] See the online appendix for a formal definition of the HM-e index.

One issue with both the HM and the HM-e index is computational complexity. Both problems are NP-hard,[16] and the difficulty of calculating indices of this type have often been binding in practice.[17] In the online appendix to this paper, we introduce an algorithm for finding the size of the largest subset of a choice data set that is consistent with acyclicality.[18] We call this problem the *maximal acyclical set problem*, or MASP. The key to our approach is to take advantage the fact that MASP is equivalent to the *minimum set covering problem* (MSCP),[19] which is well studied in the computer sciences and operations research literature. While the MSCP is also NP-hard, there are a wide variety of methods that are extremely efficient in solving it for practical cases and are included in standard 'solver' software packages (see Caprara, Toth, and Fischetti [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.[20] In tests on simulated data, our algorithm can handle data sets about an order of magnitude larger than methods currently used in economics.[21] Furthermore, because these solvers allow for the type of weighting function used in the HM-e index, they can also be used to calculate this index as well. As a result, our algorithm allows us to apply the HM-e index to the Homescan data we use in section 3 in ways that would be impossible using existing methods.

One natural question is how the HM-e index relates to other measures, both theoretically and practically. In order to discuss both these issues at the same time, we will delay this discussion until after presenting our results.

2.1 Power

One issue with rationality measures is that it can be hard to interpret what a particular value tells us about the underlying data. For example, consider a data set in which we observe choices from two disjoint choice sets. In this case the HM-e index will be 0 for *any* observed pattern of choice.

[16] 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.

[17] See Choi et al. [2007] for an example in which the constraint is binding with only 50 choices.

[18] A usable version of our algorithm is available at https://files.nyu.edu/djm431/public/software.

[19] As shown in Garey and Johnson [1979].

[20] 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).

[21] Our two comparators are the methodology applied by Choi et al. [2007] (henceforth CGFK) and Houtman [1995].

In other words, such a data set offers no meaningful test of rationality. One way to address this shortcoming is to compare the values of the HM-e index in the data to the distribution of values we would see under some alternative 'null hypothesis' for behavior. Such a comparison allows one to determine whether observed behavior shows more, less, or similar levels of rationality than the null hypothesis.

One popular benchmark is to compare index values to those that one would expect to see under uniformly random choice, where in each choice set individuals have an equal chance of choosing any object in the choice set.[22] Although random choice is a relatively weak null hypothesis, it is applicable to almost any choice setting.[23] The role of random choice in determining the statistical power of rationality measures is discussed by Bronars [1987] and is applied to Selten's measure of predictive success by Beatty and Crawford [Forthcoming]. While we base most of our analyses on this 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. Thus, with this alternative benchmark, the simulated household has an equal chance of choosing any *observed* budget shares, rather than any *feasible* budget shares.

Once we have selected and generated a benchmark, the next question is how to compare the actual data to this benchmark. For a joint test of all consumers, one can compare the *distribution* of the index values in the data with the *distribution* of index values generated under the null hypothesis using some nonparametric measure of the difference between distributions (such as the Kolmogorov-Smirnov test). In the case of a single observation, one can simply read off the percentile of the simulated data in which that observation falls. Another intuitive approach is to subtract the average simulated value from the actual value – in the style of Selten's measure of predictive success [1991]. We refer to this adjustment, which combines elements of predictive power and statistical power tests, as the 'Selten-Bronars' score.

[22] Or, in the case of budget sets, an equal chance of choosing any bundle on the budget line.

[23] Alternatively, we could generate a distribution of possible index values for a given choice environment using a more plausible error model or decision rule. For example, see Choi et al. [2007] and Andreoni and Harbaugh [2006].

3 Testing for Rationality in Scanner Data

We now apply our measure to a set of consumption data collected by the marketing firm ACNeilson. 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 to 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 transactions that occur across retail outlets. In addition, households maintain detailed demographic information that is updated annually (see the appendix of Aguiar and Hurst [2007] for a more complete description of the data).

From the initial data set, which included purchases from 2,100 households, we kept those households that participated for the entire 24 month period and had at least one purchase in every month. For the remaining 977 households, we have an average of 20.5 purchases and 7 store trips per month. Table 1 summarizes the demographics of our sample households.

TABLE 1 ABOUT HERE

In our baseline data set, we aggregate purchases at the monthly level, to alleviate concerns about the fact that some items are storable. Because the data set includes packaged food and beverage transactions for 11,517 UPC codes, it is necessary to define product categories and create price indices for each category. We place products into one of three categories: beverages, meals, and snacks. The full data set contains price information on 384,964 beverage purchases, 307,391 meal purchases, and 132,499 snack purchases. We examine the effect of both of these aggregation assumptions in section 3.1.1 below.

It should be noted that for utility maximization to imply acyclicality of revealed preference relations in this data set requires further assumptions. First, food and beverages must be additively separable in the utility function from utility for other goods and services. This assumption is strong, but standard in the literature (see for example Echenique et al. [2010]).

We also implicitly assume that, in any given period, all households face the same price for each product category as we use the same price index for all households. It is necessary to make this assumption 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.[24] Thus, if we had chosen to use a household specific price index, it would have restricted our attention to only those households with complete price information for the entire period, resulting in a loss almost 85% of households.[25]

Because prices do vary within the period and among stores, the assumption of a single price per good per period could introduce error. However, this limitation is also present when using most standard price indices, including the CPI, so it is common to papers that conduct revealed preference tests using standard price indices (such as Blundell, Browning, Crawford [2003]). We use a Stone price index in our primary analyses, but later compare our results to those obtained using other methods: Torvist, Laspeyres, and Paasche (see online appendix for further details).

3.1 Are People Rational?

Our first task is to apply our the HM-e index to find out how close the households in our baseline data set are to satisfying acyclicality. Table 2 summarizes the results of our tests.[26]

TABLE 2 ABOUT HERE

The HM-e score reported in table 2 suggests that, in absolute terms, the behavior of households in our baseline sample is close to that of the paradigmatic rational agent. While only 31% 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: about 0.16% of total expenditure. There is significant variation across households in their absolute degree of rationality: the maximum HM-e index we find is 2.68% of total expenditure. The top panel of figure 2 shows the distribution of HM-e index values in our sample population.

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

[25] 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.

[26] Note that we are treating the unit of analysis here as an individual household: we calculate the index value for each household, then average across these values.

FIGURE 2 ABOUT HERE

These raw values; however, tell us little about whether these results should be considered as providing strong support for the model of individual rationality – so far we know nothing of the power of this data set to identify irrational choice. In order to answer this question, we employ the Selten-Bronars score (power adjustment) introduced in section 2.1 to the HM-e index. 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' as a random bundle on the budget line. We then calculate the HM-e index for these 24 simulated choices. This procedure is repeated 50 times to create a distribution of index values generated under random choice for that particular household. For a given index and household, we describe as the Selten-Bronars score the index value recorded by the household minus the mean index value of the simulated data.

The 'Selten Bronars' column of table 2 reports the average Selten-Bronars score across households for the HM-e index. This 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 around 0.45% of total expenditure to achieve rationality, giving a Selten-Bronars score of -0.29%.

Thus, 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 [2010] and Echenique et al. [2010]. 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 that the actual household.

The 'percentile' column of table 2 reports the average of these percentiles across households. This measure supports the hypothesis that our observed households are on average more rational than the random benchmark, but not dramatically so: the average percentile for our households is 73. Further evidence that the behavior of random choosers is similar to that of our baseline population can be seen in figure 2. The bottom panel of the graph shows the distribution of index

values from the simulated data for all households. While this distribution is statistically different to those generated by the households in our sample (at the 1% level using the Kolmogorov-Smirnov test), the differences are not stark.

Note that this data set *does* have the power to identify irrational choice: on average only about 10% of our simulated households exhibit perfect acyclicality. Rather, what this result tells us is that a HM-e index of (for example) 0.5% would in fact reveal quite a lot about the underlying rationality for most households: even someone who was choosing at random along the budget line would not expect to have an HM-e index so high.

3.1.1 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 3 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 Homescan data. These results are shown in the second line of table 2. The value of the HM-e index for our households is little changed (0.18%, as compared to 0.16% in the baseline data). However, the amount of rationality observed in the simulations changes dramatically: at the 38 product level, there are extremely few violations of rationality in the random choice data. This means that the household data is on average *less* rational that the simulated data at this level of disaggregation, as indicated by a positive Selten-Bronars score: the average cost of removing irrationality from the random data is only 0.01%, giving a Selten-Bronars score of 0.17%.

The lack of irrationality in the random choice data reflects the fact that, with 38 product categories, the regions of the budget lines that can generate violations of GARP are small (in the sense of Beatty and Crawford [2010]). 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 8.2 categories in a month. Thus, they are often at a 'corner' of the budget set. This suggests that random simulations (which almost never hit the corner of the budget set) are not a suitable benchmark in this case. 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 2 week level, rather than the monthly level. The results are shown in line 3 of table 2. There is some evidence that data at this level contains more serious violations of rationality than does our baseline case. Almost no households are perfectly rational, and the HM-e index is higher than in the baseline case (0.48% rather than 0.16%). However, notice that in this case we observing 48 choices instead of 24 as in the baseline case, so consumers have more opportunities to exhibit irrationality. The fact that the Selten-Bronars score and mean percentile are relatively similar to that of the baseline case suggests that the larger sample size is driving much of the increase in measured rationality.

Next we try a number a different price indices (Torvist, Laspeyres, and Paasche) and compare these to the baseline case in which we use the Stone price index. These results are reported in lines 4-6 of table 2. In terms of raw value of the HM-e index, the data based on different price indices splits into two different groups: the Stone and Paasche indices give very similar results, as do the Torvist and the Laspeyres, with the former giving raw HM-e scores that are roughly double those of the latter. This grouping makes sense because the Stone and Paasche indices use a basket of goods that changes in each period based on expenditure in that period, while the Torvist and the Laspeyres indices use a more stable basket of goods based in part on expenditure in the first period.

Line 7 of table 2 reports the results if we use data collected only from purchases made from the largest chain store in the sample. This makes little difference to our results.

As a final robustness check, we use an alternative benchmark. Rather than using choices that are drawn from a uniform random distribution on the budget line, we draw budget shares for each category of goods from the observed distribution of shares across all households and budget sets. The results for the baseline data set are shown on the 8th line of table 2. They show that our consumers are somewhat more rational relative to this benchmark than to the uniformly random benchmark (on average our households fall in the 79th percentile of the simulated distribution, compared to the 73rd percentile with the baseline benchmark). This suggests that the empirical distribution of budget shares is concentrated in regions that are more 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-Bronars score becomes negative, meaning that on average, actual index values are lower than the benchmark scores. Further, the mean percentile is similar to the baseline data set with the same benchmark. Thus, this alternative benchmark appears to be more appropriate when considering a large number of product categories.

3.2 Are Some People 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 the HM-e index value and Selten-Bronars score for each household on demographic variables available in the Homescan data: whether there is a child in the household, the number of household heads, the number of regular shoppers in the household, the age of household heads, the income bracket of the household, and whether a household head graduated from college.[27] Table 3 reports the results of this regression.

TABLE 3 ABOUT HERE

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

The most 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 for example Cherchye et al. [2008]). Intuitively, in multi-person households, different household members 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 single person households to be more rational than multi-person households. In fact, we find precisely the opposite: households with a single head have Selten-Bronars scores that are *higher* than those of multi-person

[27] When the Selten-Bronars measure is the dependent variable, we use standard OLS regression. When the raw HM-e index is the dependent variable, we use a Tobit regression due to censoring at 0.

households. These differences are significant both statistically (at the 0.1% level) and economically – single households waste 0.11% more of their income relative to the random benchmark than do multi-head households. This suggests either that intra-head bargaining is not an important cause of irrational choices, or that there is some unobserved factor affecting single head households that makes them more prone to irrationality. We also find no evidence that households in which multiple people do the shopping show any more irrationality than those in which a single person does the shopping.

We also find significant effects of age on rationality. Interestingly, the relationship appears not to be linear. Households with 'young' (under 40) household heads are more rational than those with heads in the middle range (significant at the 5% level). This is in line with other results in the literature (e.g. Choi et al. 2011). However, older (over 65) households also seem to be somewhat more rational than middle age households, though this is only significant at the 12% level. This result is more puzzling, though it is compatible with the result of Aguiar and Hurst [2007] that seniors invest more time and effort in shopping.

The results on the relationship between income and rationality show the importance of using the Selten-Bronars score, rather than the raw index values. While we find a significant relationship between income group and raw value, this relationship disappears when the dependent variable is the Selten-Bronars score. This suggests that the differences in the raw values are being driven by the ability of the different data sets to demonstrate irrational behavior, rather than any difference in irrationality by the different households.

Finally, we find no relationship between rationality and either education or the presence of a child in the household.

3.3 Aggregation and Homogeneous Preferences

We next turn to the question of whether individual households can be aggregated together under the assumption of homogeneous preferences. Note that there are two related, but different issues here. First, one could ask whether, if we sum up all the demands coming from the individual households, the resulting *aggregate* demand satisfies acyclicality.[28] Second, we could ask whether

[28] In line with the work of Varian [1982], Bronars [1987], and Houtman and Maks [1985], we find that aggregated data appears extremely rational in absolute terms, even when the underlying data is not so, and that the power of

it is appropriate to assume that all households are maximizing the same set of preferences, in order to treat (for example) 24 monthly observations from 30 households as if they were 720 observations from the *same* household. In this section, we concentrate on the latter question, which is addressed in the work of Hoderlein and Stoye [2010], and is of importance (for example) in the estimating of cost-of-living changes (Blundell, Browning and Crawford [2003, 2008]).

In order to test the assumption of homogeneity, we examine the 30 individuals (single person households) who are under the age of 40 and are employed.[29] Treated as separate households, this group exhibits average levels of rationality similar to those of our baseline sample.[30]

TABLE 4 ABOUT HERE

Table 4 shows the rationality measures applied to the pooled data. The results unambiguously show that these individuals cannot be aggregated under the assumption of homogeneous preferences. The HM-e index is almost 12%, which is higher than all of the values produced by 50 random choice simulations, resulting in a positive Selten-Bronars score.

One natural question is whether it is possible to aggregate consumers together along demographic lines. In the preceding sample, we have already controlled for household composition (no children, single head, 1 primary shopper), age, and to some degree, income level. Even with these controls, the pooled data still exhibits substantial irrationality. Table 4 also shows our rationality measures when the consumers are further restricted to just female consumers (15 consumers) or male consumers (15 consumers). For female consumers in this sample, the assumption of homogeneity appears to be a reasonable one (they are in the 90th percentile of random choices), but for male consumers the assumption is still strongly rejected.

---

such rationality tests is very low.

[29] Treating these 30 individuals as a single entity gives 30 times 24, or 720 choices. Given the density of revealed preference information in our data set, this is approximately the largest number of choices for which our algorithm can work in reasonable time. In contrast, competing algorithms such as those of Houtman [1995], can handle choice sets of fewer than 100 observations.

[30] On average, an HM-e index of 0.11%, a Selten-Bronars score of -0.25, and an HM-e index in the 73rd percentile of the random distribution.

4 Comparison to Previous Work

4.1 Other Measures of Rationality - Theory

The existing literature includes many other measures of how close a data set is to satisfying rationality. In this section, we begin by discussing how these measures relate to the HM-e index in theory. We then compare these measures in practice, by comparing the results they give for our data set.

One of the earliest measures of rationality was provided by Afriat [1967], which uses 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. Afriat's measure is the largest e^* such that there are no preference cycles revealed at that efficiency level. Apesteguia and Ballester [2010] point out one problem with the Afriat measure – that it looks only at the worst violation of rationality, ignoring all others. Thus, a single bad choice can make the Afriat index arbitrarily small. Varian [1991] suggested modifying the Afriat index to allow for different efficiency levels for different budget sets. The Varian index is therefore a vector of efficiency levels, one for each observed choice, that removes all preference cycles related to that choice. Varian suggests finding such a vector that maximizes the smallest e^t, and then using this value as a summary statistic. While this value is relatively easy to compute, it is also not robust to a single bad choice.[31]

A second class of rationality measures are exemplified by the counting measure suggested by Famulari [1995]. This measure counts the number of times that GARP is violated (i.e. the number x's and y's such that x is indirectly revealed preferred to y, but y is revealed preferred to x). As with the Afriat index, this measure is not robust, in the sense that a single choice can lead to a large change in the value of the index because it can create many new cycles. Furthermore, it contains no metric of how 'severe' is a violation of GARP. This final issue is corrected by the 'money pump' measure of Echenique et al. [2010], which counts up the total cost of all violations of GARP in a data set. This measure is closest in spirit to our HM-e index and shares its computational complexity. However, this measure can also be susceptible to big efficiency losses due to a single bad choice.[32]

[31] Alternative summary statistics based on the Varian vector of efficiencies are robust – for example the minimal absolute distance of this from the unit vector. However, this measure is very complicated to compute.

[32] The difference between the money pump and the HM-e measures can be illustrated in the following example.

A third approach to constructing measures of rationality is taken by Apesteguia and Ballester [2010]. They provide axiomatic foundations for a class of rationality indices they call 'weighted loss indices'. Though similar in spirit, the HM-e index does not fall into the class of weight loss indices as it violates the composition axiom.[33]

A different window into rationality is provided by the 'rationalization by multiple rationales' (RMR) measure, which is based on the idea that agents can have different rationales for different states. For example, the relative ranking of an umbrella and a bicycle may differ depending on whether it is raining or not. If we do not observe these different states, then the resulting choices may appear irrational. This notion was captured by Kalai, Rubinstein, and Spiegler [2002], who introduced the concept of *rationalization by multiple rationales*. A rationale is a preference ordering, and a data set is rationalized by a collection of rationales if all observations are explicable as the maximization of one of the rationales. Thus, choice data that can be rationalized by n rationales can be thought of as being generated by an individual who at any time is in one of n different 'states' and in each state has a different set of preferences. Such an approach has also been applied to the analysis of household level data, to determine if household choices can be rationalized as preference maximization by one of the members of the household (Deb [2008], Nobibon et al. [2011]) and to determine if households are heterogeneous (Crawford and Pendakur [2010]).

4.2 Other Measures of Rationality – Practice

Clearly, the various measures described above *can* provide different answers as to how rational are individuals based on the same set of choices. However, it remains an open question as to whether these differences are important in practice. In this section, we redo the analysis of section 3.1 with our baseline data set, but using four alternative measures of rationality: the HM, RMR, Afriat, and Famulari measures.[34] Table 5 repeats our analysis of the baseline data set using these measures.

TABLE 5 ABOUT HERE

Say that bundle x was chosen when y was available for \$10 less, y was chosen when z was available for \$10 less, and z was chosen when x was available for 5¢ less. The money pump measure of Echenique et al. [2010] would say the cost of this cycle was \$20.05, whereas the HM-e measure would say the cost was 5¢. Thus, a single observation can lead to very large costs according to the money pump metric.

[33] Though the HM-e index satisfies the other axioms in the paper.

[34] We do not use the money pump measure of Echenique et al. [2010] due to computational complexity.

Broadly speaking, these alternative measures give the same impression of the rationality of our agents as does the HM-e index: violations from rationality exhibited by our households are small in absolute terms, but are not far from those exhibited by households that choose randomly. The average (normalized) HM index for our households is about 95%, meaning that on average the maximal acyclical set contains about 95% of all observed choices. This is about 4% larger than the size of the set for random choosers. The average RMR for our households is 1.73,[35] meaning that they generally need only a small number of rationales to explain their choices (the maximal number of rationales we find for our households is 3). The average Afriat index is around 99%, meaning that there is around a 1% loss of efficiency on average, with a 2% average efficiency loss for random choosers. Finally, the Famulari index is 0.70, meaning that the percentage of revealed preferences that are involved in cycles is just 0.7%. This compares to a violation rate of 1.65% for random choosers.

We also perform the regression analysis of table 3 using these 4 alternative rationality measures. Again, the broad message is similar across the measures. In all cases, single person households are less rational according to both the raw measure and the Selten-Bronars score. In most cases, income is significant in the raw score, but not in the Selten-Bronars score (the one exception being the HM index Selten-Bronars score, in which higher income is related to less rationality). The relationship between age and rationality is not present for all measures: the 'hump' shape relationship between age and rationality is also observed in the Famulari measure, but not in the HM, RMR, or Afriat measures.

TABLES 6 ABOUT HERE

Table 6 shows Spearman's rank correlations between the various measures across households.[36] All measures provide very similar rankings over households: none of the various measures we look at have rank correlations below 0.80. However, the degree of correlation does vary between the measures – ranging between 0.80 (the Afriat and the HM measures) and 0.91 (the Afriat and HM-e measures). The fact that the HM-e index is more correlated with each of these measures than they

[35] Our algorithm for calculating the RMR is an approximation that provides only an upper bound on the number of rationales needed.

[36] For this comparison, these measures have been normalized so that a higher value of the measure is always less rational.

are with each other confirms that the HM-e index is 'in between' the pure counting measures (HM index and Famulari) and the 'severity' measures (Afriat).[37]

4.3 Previous Tests of Rationality

There is a small existing literature that tests the degree of rationality in laboratory and field settings. Crawford and Pendakur [2010] estimate the RMR index for a cross section of data on milk demand. They find that the 500 households in their survey can be explained by either 4 or 5 types. This is compatible with our finding on homogeneity – even though one type may be a bad assumption, a small number of types fully rationalizes the data. Echenique et al. [2010] examine rationality in a data set similar to ours. They conclude that, while violations of GARP are common, the cost of these violations is relatively low. 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 also find that younger, richer, better educated, and larger households have higher rationality values. Our study differs from theirs primarily in that (a) we focus on power-adjusted measures rather than raw values and (b) we also examine the question of whether consumers can be pooled together under the assumption of homogeneous preferences. Choi et al. [2011] collected experimental data on choices over lotteries from a panel of 2000 Dutch subjects. They found that rationality was significantly higher in their subjects than under the random benchmark. They also found significant differences in rationality between demographic groups: with high income, high education, male, and younger subjects showing higher levels of consistency. Hoderlein [2010] 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.

4.4 Homogeneity in the BBC Data

The substantial preference heterogeneity found using our Homescan data suggests that preference heterogeneity may appear with other consumption data sets as well. In order to examine this possibility, we repeat our analysis using a subset of the data from BBC.[38] Table 7 shows the

[37] The same pattern emerges if we use Pearson's correlation coefficient.

[38] The authors use 25 years of data from the (annual) British Family Expenditure Survey. For each household, consumption is aggregated into three categories: food, other nondurables, and services. Prices for each of these

HM-e index and associated Selten-Bronars score (based on 50 simulations) if we pool together all households in the greater London area from the BBC consumption data for the years 1993 to 1999. Unlike the Homescan data, we do not find evidence for severe preference heterogeneity across households: the HM-e index of 0.922 is lower than the mean of the index values for simulated data, resulting in a negative Selten-Bronars score. In fact, it is lower than all index values for simulated data, placing the actual index value in the 100th percentile. Thus, using this metric, the pooled households from the BBC data set are more rational than most of the individual households from the Homescan data. This suggests that the degree of preference heterogeneity may not be particularly large in the BBC data set.

TABLE 7 ABOUT HERE

We speculate that the reason for the lower level of preference heterogeneity in the BBC is due to a higher level of product aggregation. BBC aggregate to three categories: food and beverages, other nondurable goods, and services. All goods in the Homescan data fall into the first of these categories. Our results are consistent with the hypothesis that preferences are more homogeneous at higher levels of aggregation. Testing this assumption is clearly an avenue for further work.

5 Conclusion

We believe that the contribution of this paper is twofold. First, we have applied the HM-e index and importantly, the Selten Bronars measure, to a rich data set that tracks most packaged meal, beverage, and snack purchases of a group of 977 households in the Denver area over a period of two years. Our results show that while deviations from the predictions of utility maximization are not costly in absolute terms, this is not necessarily an indication of high levels of rationality: households that chose at random from the budget line would also not exhibit particularly costly variations of rationality. This accounting for power is important for understanding 'true' differences in rationality. We do; however, find strong evidence against the assumption that households can be treated as homogenous entities with the same preferences – an important point for welfare analysis.

Second, we have improved on the tools available to researchers who are interested in how close

categories are then calculated using the Retail Price Index, which is the same for all households in a given year.

a data set is to satisfying rationality. The HM-e index offers a new measure of rationality that combines elements of previous 'counting' measures (Famulari [1985], Houtman and Maks [1985]) and 'severity' measures (Afriat [1967]). When combined with the power adjustment we use, it provides an robust picture of rationality for a set of choices. Moreover, the algorithm we use to calculate the HM-e index is a significant improvement on those currently in use in economics, and helps to overcome the issue of computational complexity that has plagued many rationality measures (such as the HM and RMR indices).

An important avenue for future research would be to apply measures such as the HM-e index to a broader class of data sets in order to determine more thoroughly (a) how far people's consumption choices are from utility maximization and (b) the source of these discrepancies. Our results show that, in this data set, the costs of departures from rationality are small, but the Selten Bronars score tells us we would not *expect* them to be very big. An important question is therefore whether, in a data set which offer a lot of scope for irrationality, we observe very costly deviations. Furthermore, given that there are many factors that could lead to cyclic choices in consumption data – 'true' irrationality, preference shocks, changes in household composition, aggregation of prices and quantities – it is clearly important to uncover what is leading to observed irrationality. A promising avenue in this regard is the use of laboratory-style experiments in field settings, as demonstrated by Choi et al. [2011].

References

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

Afriat, S. (1972) "Efficiency Estimates of Production Functions," *International Economic Review*, 8, 568-598.

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

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

Andreoni, J. and W. Harbaugh (2006) "Power Indices for Revealed Preference Tests," Unpublished.

Apesteguia, J., and M. Ballester (2010) "A Measure of Rationality and Welfare," Unpublished.

Beatty, T. and I. Crawford (Forthcoming) "How Demanding is the Revealed Preference Approach to Demand?," *American Economic Review*.

Blundell, R., M. Browning, and I. Crawford (2003) "Nonparametric Engel Curves and Revealed Preference," *Econometrica*, 71, 205-240.

Blundell, R., M. Browning, and I. Crawford (2008) "Best Nonparametric Bounds on Demand Responses," *Econometrica*, 76, 1227-1262.

Bronars, S. (1987) "The Power of Nonparametric Tests of Preference Maximization," *Econometrica*, 55, 693-698.

Caprara, A., Toth, P., and M. Fischetti (2000) "Algorithms for the Set Covering Problem," *Annals of Operations Research*, 98, 352-371.

Cherchye, L., De Rock, B., Sabbe, J., and F. Vermeulen (2008) "Nonparametric Tests of Collectively Rational Consumption Behavior: An Integer Programming Procedure," *Journal of Econometrics*, 147, 258-265.

Choi, S., Gale, D., Fisman, R., and S. Kariv (2007) "Consistency and Heterogeneity of Individual Behavior under Uncertainty," *American Economic Review*, 97(5), 1921-1938.

Choi, S., S. Kariv, W. Müller, and D. Silverman (2011) "Who Is (More) Rational?," NBER Working Papers 16791, National Bureau of Economic Research, Inc.

Crawford, I. and K. Pendakur (2010) "How Many Types Are There? A Simple, Theory-Consistent Approach to Unobserved Heterogeneity." Unpublished.

Deb, R. (2008) "An Efficient Nonparametric Test of the Collective Household Model," Unpublished.

Echenique, F., Lee, S., and M. Shum (2010) "Revealed Preference Tests using Supermarket Data: the Money Pump." Social Science Working Paper, 1328. California Institute of Technology

Famulari, M. (1995), "A Household-Based, Nonparametric Test of Demand Theory," *Review*

of Economics and Statistics, 2(2), 371-382.

Fisman, R., Kariv, S., and D. Markovits (2007) "Individual Preferences for Giving," *American Economic Review*, 97(2), 153-158.

Garey, M. R. and D. S. Johnson (1979) "Computers and Intractability: A Guide to the Theory of NP-Completeness," New York: W.H. Freeman.

Gross (1995) "Testing Data for Consistency with Revealed Preference," *The Review of Economics and Statistics*, 701-710.

Hoderlein, S. (2010) "How Many Consumers Are Rational?," Unpublished.

Hoderlein, S. and J. Stoye (2009) "Revealed Preferences in a Heterogeneous Population," Boston College Working Papers in Economics 745, Boston College Department of Economics.

Houthakker, H. S. (1950) "Revealed preference and the utility function." *Economica*, 17, 159-174.

Houtman, M. (1995) "Nonparametric Consumer and Producer Analysis," Thesis Rijksuniversiteit Limburg Maastricht.

Houtman, M., and J. A. H. Maks (1985) "Determining all Maximal Data Subsets Consistent with Revealed Preference," *Kwantitatieve Methoden*, 19, 89-104.

Kalai, G., A. Rubinstein, and Ran Spiegler (2002) "Rationalizing Choice Functions By Multiple Rationales," *Econometrica*, 70(6), 2481-2488.

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

Manser, M. E. and R. J. McDonald (1988) "An analysis of substitution bias in measuring inflation, 1959-85" *Econometrica*, 56, 909-93

Nobibon, T., Cherchye, B., De Rock, J., Sabbe, J., and F. Spieksma (2011) "Heuristics for Deciding Collectively Rational Consumption Behavior," *Computational Economics*, 38(2): 173-204.

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

Varian, H. (1982), "The Nonparametric Approach to Demand Analysis," *Econometrica*, 5(4).

Varian, H. (1991) "Goodness-of-Fit for Revealed Preference Tests," University of Michigan CREST Working Paper # 13.

Varian, H. (2006) "Revealed Preference," In *Samuelsonian Economics and the Twenty-First Century.*

Household composition	
Child present	38%
Two household heads	72%
Number of primary shoppers	
1	75%
2	25%
3	<1%
Education	
College degree	39%
Age of household head(s)	
<40	24%
40-64	52%
>=65	24%
Household income	
<30k	12%
30k-49k	56%
>=50k	32%

Table 1: Summary of demographic characteristics of baseline sample

'Child present' means that a child under 18 was reported to be living in the house.
A shopper is 'primary' if he or she makes purchases equal to at least 25% of expenditure.
'College degree' means that all household heads reported having a college degree.
'Age of household head(s)' is the maximum reported age among household heads.

Description	Number of households	Perfectly rational	HM-e	Selten-Bronars	Percentile
Baseline	977	31%	0.16 (0.26)	-0.29 (0.35)	73 (29)
38 Products	977	25%	0.18 (0.30)	0.17 (0.30)	30 (42)
2 Week	397	1%	0.48 (0.52)	-0.32 (0.61)	69 (29)
Torvist	977	31%	0.08 (0.13)	-0.08 (0.15)	68 (31)
Laspayres	977	31%	0.08 (0.14)	-0.08 (0.16)	67 (31)
Paasche	977	31%	0.16 (0.26)	-0.29 (0.35)	73 (29)
Largest Chain	332	27%	0.24 (0.43)	-0.28 (0.51)	70 (30)
Alternative Benchmark	977	31%	0.16 (0.26)	-0.52 (0.43)	79 (24)
38 Products + Alternative Benchmark	977	25%	0.18 (0.30)	-0.27 (0.38)	73 (29)

Table 2: Summary of HM-e index results
Standard deviations in brackets.
'Perfectly rational' reports the proportion of households whose data generates no preference cycles.
'HM-e' reports the average across population households of the raw HM-e index value for each household (x 100).
'Selten-Bronars' reports the average across population households of the difference between the HM-e index for each household and the average of simulated values from a population who choose at random from the same budget sets as faced by that household (x 100).
'Percentile' shows the average across population households of the percentile rank of the simulated values that are equal to that household's actual value.
Baseline: baseline sample; 38 Product: products aggregated into 38 categories, rather than 3 categories; 2 week: purchases aggregated at the 2 week level, rather than the monthly level; Torvist, Layspayres, Passche: alternatives to the Stone price index; Largest chain: uses data only from the largest chain store, rather than all purchases; Alternative benchmark: uses random draws from observed distribution of budget shares to generate Selten Bronars benchmark, rather than uniform distribution over the budget line.

	Selten-Bronars	HM-e
Child	-0.037	0.033
	(0.027)	(0.027)
Single head	0.114***	0.075**
	(0.028)	(0.029)
2 Shoppers	0.021	-0.009
	(0.027)	(0.027)
3 Shoppers	-0.094	-0.138
	(0.131)	(0.137)
College degree	0.023	0.003
	(0.024)	(0.027)
Age <40	-0.056 **	-0.029
	(0.028)	(0.029)
Age >=65	-0.047	0.027
	(0.030)	(0.031)
Income <30k	0.047	0.068*
	(0.035)	(0.037)
Income >=50k	0.013	0.058**
	(0.027)	(0.027)
N	977	977
R^2	0.04	

Table 3: Regression of rationality measures on demographic variables
*First column shows the results of regressing the Selten-Bronars score on a series of demographic variables (using OLS). Each cell reports the parameter values and standard errors in parentheses. * indicates significance at 10%, ** at 5%, and *** 1%. In the second column, the dependent variable is the raw HM-e index value (using Tobit).*
***Child**: Takes value 1 if there is a child in the household. **Single head**: Takes value 1 if household has only one household head. **Shoppers x**: Takes value 1 if there are x number of people in the household that account for more than 25% of shopping expenditure. **College degree**: Takes value 1 if all heads in the household are educated to the college level. **Age x**: Takes value 1 if household head is in the age range x. **Income x**: Takes value 1 if household income is in range x.*

Description	Number of households	Number of observations	HM-e	Selten-Bronars	Percentile
Pooled Data	1	720	11.98	3.86	0
Females	1	360	3.00	-1.18	90
Males	1	360	7.37	3.34	0

Table 4: Summary of HM-e index results for the choices of employed individuals under 40 years old pooled together
Standard deviations in brackets.
'HM-e' reports the raw HM-e index value (x 100).
'Selten-Bronars' reports the difference between the HM-e index and the average of simulated values from a population who choose at random from the same budget sets (x 100).
'Percentile' shows the largest percentile rank of the simulated values that are equal to the actual value.

Description	Value	Selten-Bronars	Percentile
HM-e	0.16 (0.26)	-0.29 (0.35)	73 (29)
HM	0.95 (0.04)	0.04 (0.05)	80 (25)
RMR	1.73 (0.52)	-0.29 (0.54)	89 (17)
Afriat	0.99 (0.02)	0.01 (0.02)	73 (30)
Famulari	0.70 (0.84)	-0.95 (1.31)	73 (29)

Table 5: Summary of results for alternative rationality measures
'Value' reports the average across population households of the raw index value for each household (x 100 for HM-e and Famulari).
'Selten-Bronars' reports the average across population households of the difference between the index value for each household and the average of simulated values from a population who choose at random from the same budget sets as faced by that household (x 100 for HM-e and Famulari).
'Percentile' shows the average across population households of the largest percentile rank of the simulated values that are equal to that household's actual value.

	HM-e	HM	RMR	Afriat
HM	0.84			
RMR	0.82	0.85		
Afriat	0.91	0.80	0.81	
Famulari	0.86	0.90	0.83	0.81

Table 6: Spearman's rank correlation among different rationality measures. The sample is the 977 households in the baseline sample.

Description	Number of households	Number of observations	HM-e	Selten-Bronars	Percentile
Pooled Data	1	893	0.922	-0.939	100

Table 7: Summary of HM-e index results for the choices of BBC households in the greater London area from the BBC consumption data for the years 1993 to 1999 pooled together
'HM-e' reports the raw HM-e index value (x 100).
'Selten-Bronars' reports the difference between the HM-e index and the average of simulated values from a population who choose at random from the same budget sets (x 100).
'Percentile' shows the largest percentile rank of the simulated values that are equal to the actual value.

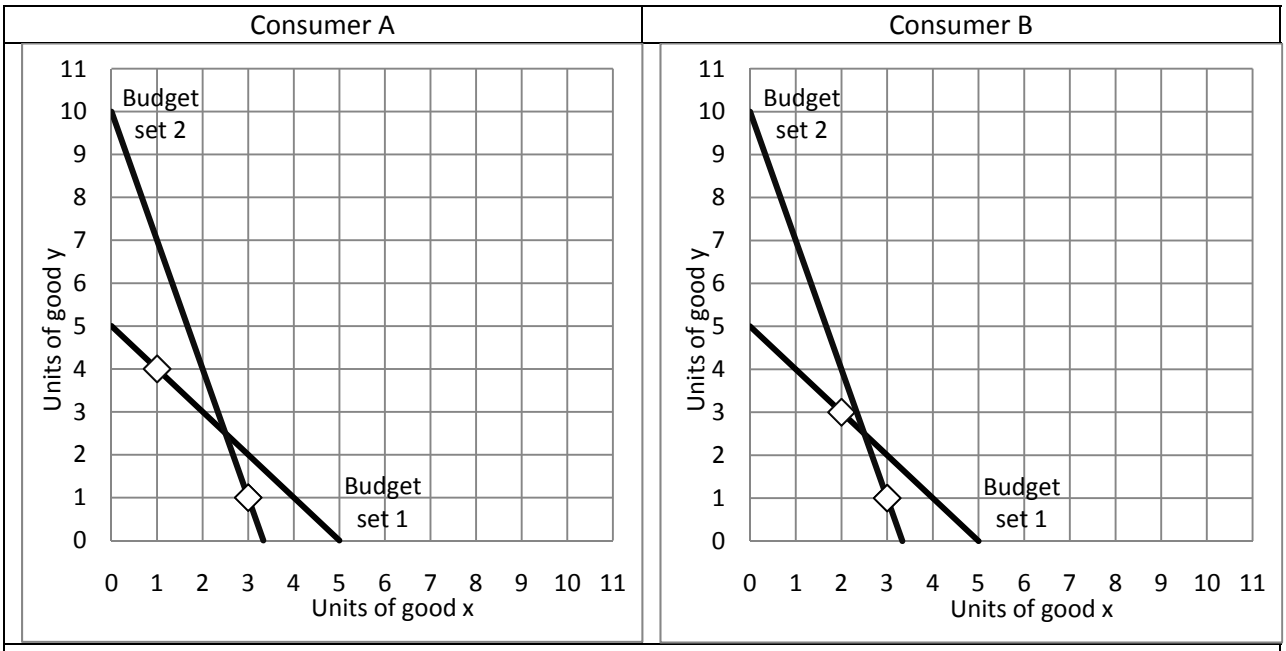

Figure 1: Hypothetical consumption choices

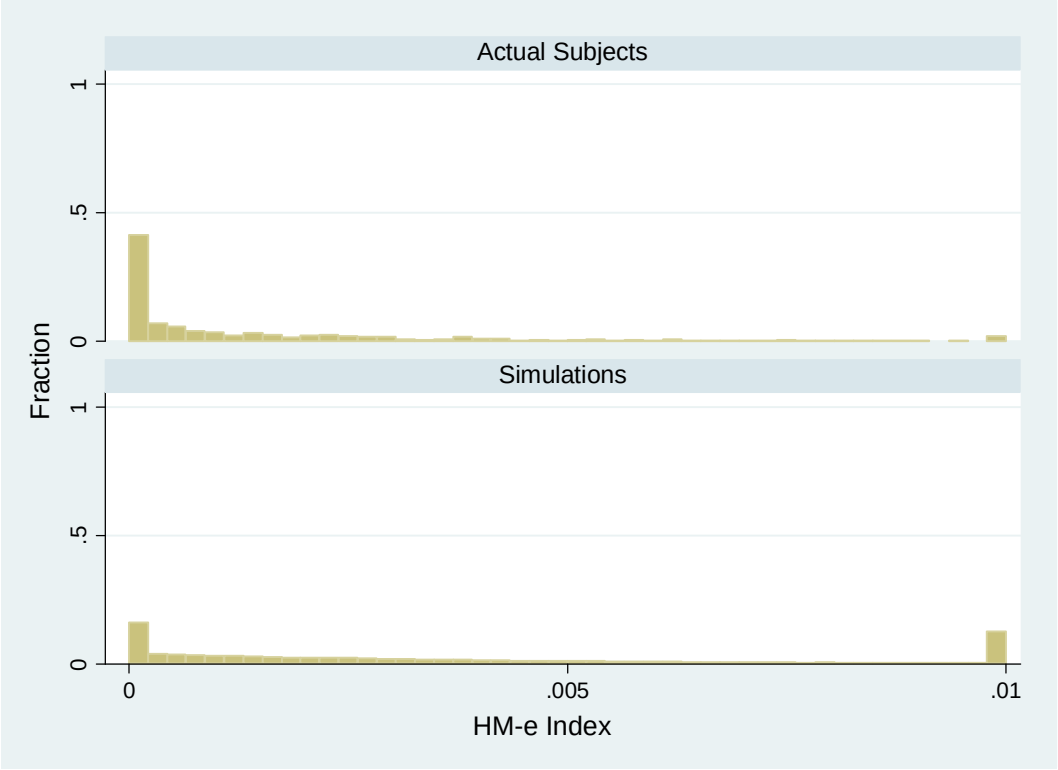

Figure 2: Distribution of HM-e index values (truncated at an HM-e index of 1%) in the baseline population (top panel) and in a simulated population of random choosers (bottom panel)