Testing for Utility Maximization with Error and the Loss of Power

Jan Heufer*

Forthcoming in *German Economic Review*

Abstract

A procedure is suggested to decide whether or not to treat a consumer who violates the Generalized Axiom of Revealed Preference (GARP) as "close enough" to utility maximization. It is based on the reduction of the power the test has against random behaviour. It can also be used to compare different efficiency indices.

Keywords: Consumer Choice; Efficiency Index; GARP; Nonparametric Tests; Test Power; Revealed Preference

JEL classification: C14; C91; D12.

*Jan Heufer, TU Dortmund, Wirtschafts- und Sozialwissenschaftliche Fakultät, Volkswirtschaftslehre (Mikroökonomie), D-44221 Dortmund, Germany. Tel.: +49 231 755 7223; fax: +49 231 755 3027. E-mail: jan.heufer@tu-dortmund.de.

1 Introduction

Within a standard framework of decisions by consumers on competitive budgets, revealed preference methods offer an unambiguous way of testing whether a set of observations on consumption could have been generated by a single utility maximizing consumer. The test was originally developed by Afriat (1967). Varian (1982) showed that his Generalized Axiom of Revealed Preference (GARP) is equivalent to Afriat's condition of cyclic consistency. Consistency with GARP can easily be tested. Afriat's Theorem shows that consistency with GARP is equivalent to the existence of a non-satiated, continuous, monotonic utility which rationalized the data, i.e. the observed choices maximize a single well behaved utility functions given the budget constraints.

Examples of apparent violations of rational behaviour in controlled experiments are abundant (e.g. Allais 1953, Ellsberg 1961, Kahneman 1994, Kahneman and Thaler 2006). However, many experiments examine the behaviour of inexperienced subjects who play for rather low amounts of money. With experience and increased incentives, observed behaviour is more in line with rationality. For example, Rapoport et al. (2003) find that with high stakes and multiple iterations, players approach equilibrium play in a centipede game. Another example is the work of Menkhoff et al. (2010), who find that institutional investors behave more sophisticated than both laymen and investment advisors (where the latter usually have less access to information and earn lower salaries than institutional investors). The drop in consumer expenditure upon entry into retirement is an example of a non-experimental high-stake observation that seemingly violates the rationality hypothesis implicit in the assumption of consumption (or perhaps rather utility) smoothing over the life cycle (see for example Banks et al. 1998). Several explanations for this observations have been proposed in order to reconcile it with the utility smoothing hypothesis; most recently, Lührmann (2010) finds that the drop can be explained by home production as a

substitute for expenses.

Within the rich literature on testing variants of the utility maximization hypothesis, this article focuses on testing for GARP on standard competitive budget; we use data from so called induced budget experiments. If a consumer's decisions are inconsistent with GARP one would like to have a test for "almost optimizing" behaviour, or one might want to have an idea of how severe this violation of utility maximization is. One such measure is the Afriat Efficiency Index (AEI, Afriat 1972) or Critical Cost Efficiency Index, which is widely used. One practical problem of the AEI is that there is no natural critical value of the AEI – without further information it is difficult to decide whether an AEI of, say, 95% should still count as "close to utility maximization".

Bronars (1987) suggests a Monte Carlo approach to determine the power the test has against random behaviour. The approximate power of the test is the percentage of random choices which violate GARP.

Surprisingly it has not always been noted that accepting certain consumers who exhibit less than 100% efficiency as "close enough" to GARP decreases the power of the test. Sippel (1996) attributes the first notice of the problem to Famulari (1995). He shows that for data from three experiments (Battalio et al. 1973, Mattei 2000, Sippel 1997) the test for GARP loses most of its power when accepting most subjects as close enough to GARP. Fisman et al. (2007), Choi et al. (2007a) and Choi et al. (2007b) compute and compare the distribution of the AEI of their experimental data and of random choices based on Bronars' procedure and arrive at more optimistic results.

The aim of this paper is to establish a procedure for testing almost optimizing behaviour based on the loss of power if we accept some consumers with GARP violations as close enough to GARP. This reveals the tradeoff between accepting consumers with an AEI of less than 100% and the correspondingly higher rate of random choices which then pass the test. The researcher can then decide what level of power is still acceptable given the

number of consumers he can then treat as utility maximizers. A small loss in power may be appropriate if that allows to treat many more consumers as utility maximizers – assuming that this is an objective – whereas a researcher should refrain from losing power if that loss would lead to only a small number of additional consumers accepted as close enough to GARP.[1]

Whenever data is tested for consistency with GARP or similar axioms, the power of the test, i.e. the probability that a set of random choices is not accepted as utility maximizing, is an important indicator. If the power is too low, then the result that, say, experimental subjects are utility maximizers is insignificant, as even many random choices cannot be rejected as utility maximizing. A higher power increases the confidence in the test result. On the other hand, it is quite likely that experimental subjects sooner or later will make a "mistake" if asked for enough choices on different choice sets. For example, Fisman et al. (2007) report that only eight out of their 78 subjects in a dictator game had no violations of GARP, and those were the ones who always chose selfish allocations. Many economists might feel that such a result should not be used to completely dismiss the assumption of utility maximizing economic agents, even as justified simplification of economic modelling. In particular, many of the subjects had very high AEI and were thus "close" to utility maximization.

Accepting many subjects as "close enough" may be an objective on its own if the researcher uses the GARP test as a pre-screening mechanism before going on to use the data to estimate parametric utility functions. Rejecting subjects who are "close" to utility maximization may seem wasteful if the loss of power associated with the rejection is negligible. If on the other hand accepting many subjects as rational leads to a substantial loss in power, the results of the study may become meaningless. Thus, a researcher would be greatly misled if the tradeoff described here is ignored.

[1] Heufer (2008) uses a similar approach to compare the AEI with a new efficiency index.

We can also try to relate the tradeoff to a measure of success, such as Selten's (1991) measure of predictive success. It is shown how a reinterpretation of Selten's measure can be helpful in determining the optimal tradeoff.

There are also alternatives to the AEI. The method suggested in this paper can also be used to compare several efficiency indices. An index can dominate another index in the sense that basing the decision of which subjects to reject on the first index always leads to smaller decrease in power. In that case a researcher is well advised to use the first index. This does not require the assumption that choosing the largest acceptable portion of subject as "close enough" to utility maximization is an objective. This also allows a natural application of Selten's measure of success: If one efficiency index accepts more utility maximizing data with minor errors for any given efficiency level than an alternatives index, it is more successful.

The remainder of this article is organized as follows. Section 2 gives a short introduction to revealed preference theory and describes the suggested procedure. This section also discusses the interpretation of power in the context of this article, and the relationship with Type I and Type II errors. Section 3 applies the procedure to simulated utility maximizing data with stochastic error and to data from experimental dictator games. Section 4 concludes.

2 Theory

2.1 Preparations

A set of observed consumption choices consists of a set of chosen bundles of commodities and the prices and incomes at which these bundles were chosen. Let $X = \mathbb{R}^{\ell}_{+}$ be the

commodity space, where $\ell \geq 2$ denotes the number of different commodities.[2] The price space is $P = \mathbb{R}^{\ell}_{++}$, and the space of price-income vectors is $P \times \mathbb{R}_{++}$. Consumers choose bundles $x^i = (x^i_1, \ldots, x^i_\ell)' \in X$ when facing a price vector $p^i = (p^i_1, \ldots, p^i_\ell) \in P$ and an income $w^i \in \mathbb{R}_{++}$. A budget set is then defined by $B^i = B(p^i, w^i) = \{x \in X : p^i x^i \leq w^i\}$. The entire set of M observations on a consumer is denoted as $S = \{(x^i, B^i)\}_{i=1}^{M}$. When we assume that we observe choices of many different consumers (say, subjects in an experiment), we let N denote the number of consumers and let $S(n)$ denote the set of observations for the nth consumer.

A utility function $u(x)$ *rationalizes* a set of observations S if $u(x^i) \geq u(x)$ for all x such that $p^i x^i \geq p^i x$ for all $i = 1, \ldots, n$.

The following definitions are needed to recover consumer preferences that are implicit in a set of consumption choices: An observation x^i is *directly revealed preferred* to x, written $x^i R^0 x$, if $p^i x^i \geq p^i x$; *revealed preferred* to x, written $x^i R^* x$, if for some sequence of bundles (x^j, x^k, …, x^m) such that $x^i R^0 x^j$, $x^j R^0 x^k$, …, $x^m R^0 x$. In this case R^* is the *transitive closure* of the relation R^0; *strictly directly revealed preferred* to x, written $x^i P^0 x$, if $p^i x^i > p^i x$.

For consistency with the maximization of a piecewise linear utility function, Varian (1982) introduced the following condition: The set of observations S satisfies the *Generalized Axiom of Revealed Preference* (GARP) if $x^i R^* x^j$ implies [not $x^j P^0 x^i$]. It can then be shown (Afriat 1967, Varian 1982) that if the data satisfy GARP then there exists a concave, monotonic, continuous, non-satiated utility function that rationalizes the data.

Several goodness-of-fit measures have been proposed. Arguably the most popular measure for the severity of a violation is the Afriat efficiency index (AEI) due to Afriat (1972). Reporting the AEI has become a standard for experimental studies.[3] To obtain the

[2]The following notation is used: For all $x, y \in \mathbb{R}^{\ell}$ we write $x \geq y$ for $x_i \geq y_i$ for all i, $x > y$ for $x_i \geq y_i$ and $x \neq y$ for all i, and $x \gg y$ for $x_i > y_i$ for all i. We denote $\mathbb{R}^{\ell}_{+} = \{x \in \mathbb{R}^{\ell} : x \geq 0\}$ and $\mathbb{R}^{\ell}_{++} = \{x \in \mathbb{R}^{\ell} : x \gg 0\}$.

[3]See, for example, Sippel (1997), Mattei (2000), Harbaugh et al. (2001), Andreoni and Miller (2002),

AEI, budgets are shifted towards the origin until a set of observations is consistent with GARP. Let e be a number between 0 and 1. Define the relation $R^0(e)$ to be $x^i R^0(e) x^j$ if $e\, p^i x^i \geq p^i x$, and let $R^*(e)$ be the transitive closure of $R^0(e)$. We then say that the set of observations S satisfies GARP(e) if $x^i R^*(e) x^j$ implies [not $e\, p^j x^j > p^j x^i$]. Then the AEI is the largest number e such that GARP(e) is satisfied. Note that AEI can be interpreted as a measure of wasted income: If a consumer has an AEI of, say, .95, then he could have obtained the same level of utility by spending only 95% of what he actually spent to obtain this level.

See Gross (1995) for a survey of other measures.

2.2 Power against Random Behaviour

Depending on the characteristics of the budget sets, the chance of violating GARP can differ substantially. A completely rational consumer will always be consistent. However, even a consumer who makes purely random decisions has a chance to satisfy GARP. Bronars (1987) suggests a Monte Carlo approach to determine the power the test has against random behaviour. The approximate power of the test is the percentage of random choices which violated GARP. Bronars' first algorithm follows Becker's (1962) example by inducing a uniform distribution across the budget hyperplane. For Bronars' second algorithm, the random choices are generated by drawing ℓ i.i.d. uniform random variables, $z_1, \ldots, z_\ell$, for each price vector, and calculate budget shares $Sh_j = z_j / \sum_{k=1}^{\ell} z_k$. The random demand for commodity x_j is then calculated as $x_j = (Sh_j\, w)/p_j$.

Note that statistical power is usually defined as the probability that the test will reject a false null hypothesis, i.e. that there is no Type II error. This raises the question of whether the null hypothesis is indeed false whenever a set of random decisions happens to satisfy GARP; after all, the random decisions are indistinguishable from utility maximizing

Février and Visser (2004), Choi et al. (2007b), Fisman et al. (2007), Dickinson (2009).

behaviour. However, we know that the observed decisions are generated by a random process, and if we continue to use this random process to generate more and more decisions, then we will eventually observe a violation of GARP if a violation of GARP is possible (i.e. if at least two budgets intersect). In any case, the power of the test in this context should be understood as the probability that random decisions violate GARP, and it tells us how meaningful the test for GARP is. A test with low power will tell us little about whether or not consumers are utility maximizing, and a test with higher power is arguably preferable to examine the utility maximizing hypothesis.

2.3 The Procedure

Varian (1990) suggests a 95% AEI as the critical value for acceptance of GARP-violating sets of observations as utility maximizing, "for sentimental reasons". There is, however, no natural critical value. We therefore suggest to generate random choices on the budget sets and to recompute Bronars' power for all observed efficiency levels between 0 and 1. This will give us an idea of how much power the test loses if we accept GARP-violating observations as close enough to GARP. This procedure also allows one to compare different efficiency indices.

To approximate the power of a GARP test if we allow deviations from utility maximization, we need to generate random choices on the budget sets. To get a good approximation of power, we need to generate many sets of decisions:

A We want to generate N^{Sim} "random decision makers" with M choices on M budgets each, using Bronar's first algorithm:[4]

For all $n^{\text{Sim}} = 1, \ldots, N^{\text{Sim}}$,

- for all $m = 1, \ldots, M$,
  - draw a random point SP from the $(\ell - 1)$-simplex using a simplex point

[4]The results for Bronars' second algorithm are very similar.

picking algorithm;

– set $(x_1^m, \ldots, x_\ell^m) = \left(\frac{SP_1\, w^m}{p_1^m}, \ldots, \frac{SP_\ell\, w^m}{p_\ell^m}\right)$;

- set $T(n^{\text{Sim}}) = \{(x^i, B^i)\}_{i=1}^M$.

Procedure **A** can then be used to compute the approximate power of the test by testing each set $T(n^{\text{Sim}})$ for consistency with GARP; the fraction of sets which are inconsistent gives the power. Note that N^{Sim} should be large in order to provide a reliable estimate. In his example, Bronars (1987) used $N^{\text{Sim}} = 200$. Given modern computers, much larger N^{Sim} are feasible; we suggest to use at least $N^{\text{Sim}} = 5,000$.

The second step is to compute the AEI of the real consumers and of the random decision makers:

B1 We want to compute the AEI for each of the N consumers in the observed data sets:

For all $n = 1, \ldots, N$,

- compute the AEI for consumer n, i.e. for the set of observations $S(n) = \{(x^i, B^i)\}_{i=1}^M$;
- let $\text{AEI}(n)$ denote the AEI for this consumer.

B2 We want to compute the AEI for sets of random choices, with N^{Sim} sets overall:

- Execute Procedure **A**;
- for all $n^{\text{Sim}} = 1, \ldots, N^{\text{Sim}}$,
  - compute the AEI for consumer n^{Sim}, i.e. for the set of observations $T(n^{\text{Sim}}) = \{(x^i, B^i)\}_{i=1}^M$;
  - let $\text{AEI}^{\text{Sim}}(n^{\text{Sim}})$ denote the AEI for this consumer.

The third step is to compute the loss of power of the test for all possible AEI:

C We want to compute how much power we lose when we accept certain consumers as utility maximizing:

- Sort the set $\{\text{AEI}(i)\}_{i=1}^N$ by decreasing values of $\text{AEI}(i)$, which gives the sorted set $\{\text{SAEI}(i)\}_{i=1}^N$;

- for all $n = 1, \ldots, N$,

  set $f(n) = 1 - \frac{1}{N^{\text{Sim}}} \sum_{j=1}^{N^{\text{Sim}}} \delta_{j,n}$, where

$$\delta_{j,n} = \begin{cases} 1 & \text{if } \text{AEI}^{\text{Sim}}(j) \geq \text{SAEI}(n) \\ 0 & \text{otherwise.} \end{cases}$$

Given sets of observations of N consumers on M budgets each, $f(i)$ gives the approximate power of the test when we accept the i subjects with the highest AEI as "close enough" to GARP.

3 Application

3.1 Simulated Data

To illustrate the procedure, we take data from a known generating function and add stochastic error to simulate measurement error. Ideally, we would like to accept all of sets obtained in this way for reasonably low error terms as utility maximizing without thereby reducing the power of the applied test.

We use a similar procedure as applied in Fleissig and Whitney (2003, 2005). First, we generate data from a five commodity Cobb-Douglas utility function given by

$$U(x^*) = \prod_{i=1}^{5} x_i^{*\alpha_i}, \qquad \text{with } \sum_{i=1}^{5} \tilde{\alpha}_i = 1 \tag{1}$$

We use random parameters each time by drawing each $\tilde{\alpha}_i$ from a uniform distribution $\mathcal{A} \sim \mathcal{U}[.05, .95]$ and then normalizing it such that $\sum_{i=1}^{5} \alpha_i = 1$, i.e. $\alpha_i = \tilde{\alpha}_i / \left(\sum_{i=1}^{5} \alpha_i\right)$.

For the Monte-Carlo experiment we assume that we observe the demand according to the given utility function with some measurement error that fluctuates by $\kappa\%$ around the

true demand in a manner defined below; we use $\{\kappa_1, \kappa_2, \kappa_3, \kappa_4\} = \{.05, .1, .2, .25\}$.

The datasets have $M = 20$ observations each, with expenditures drawn from a uniform distribution $\mathcal{W} \sim \mathcal{U}[10000, 12000]$. Price vectors are drawn from a uniform distribution $\mathcal{P} \sim \mathcal{U}[95, 100]$. These expenditures and prices lead to many intersections of budget sets which can lead to many violations of GARP. As in Fleissig and Whitney (2003, 2005), we use a new set of random budgets for each consumer. This avoids the problem of finding an optimal or representative set of budgets for the analysis.[5]

The data are generated by the following steps:

A' We want to generate N "artificial consumers" with M choices each:

For all $n = 1, \ldots, N$,

- draw a random vector $\tilde{\alpha}$, with each $\tilde{\alpha}_i$, $i = 1, \ldots, 5$, drawn from $\mathcal{A}$, and set $\alpha_i = \tilde{\alpha}_i / \left(\sum_{i=1}^{5} \tilde{\alpha}_i\right)$;
- for all $m = 1, \ldots, M$,
  - draw a random budget B^m by drawing a random expenditure w^m from $\mathcal{W}$ and a random price vector p^m from $\mathcal{P}$;
  - generate demand $x^{*,m}$ on the budget B^m which maximizes Eq. (1) with parameter vector α;
  - generate demand with stochastic error by multiplying each $x_i^{*,m}$ by $(1 + \varepsilon_i)$ for $i = 1, \ldots, 5$, where $\varepsilon_i \sim \mathcal{U}[-\kappa, \kappa]$, i.e. set $\tilde{x}_i^m = x_i^{*,m}(1 + \varepsilon_i)$;
  - to keep expenditure constant, normalize $\tilde{x}^m$, i.e. set $x_i^m = \frac{\tilde{x}_i^m \, w^m}{p^m \, \tilde{x}^m}$;
- set $S(n) = \{(x^m, B^m)\}_{m=1}^{M}$.

For illustrative purposes we use procedure **A'** to generate four data sets with $N = 10,000$ each, for the four different values of κ. These are used as sets of observations. We use the same set of budgets for each of the values of κ. Then, again using the same set of budgets,

[5]The procedure was repeated several times, using a fixed set of budgets for each consumer each time, with somewhat less "smooth" but otherwise very similar results. A very similar procedure was used in Fisman et al. (2007) and Choi et al. (2007a), who drew a new set of random budgets for each of the subjects in their experiments.

we execute procedure **A** with $N^{\text{Sim}} = 50$ per set of budgets. Given the number of budgets and the fact that all sets of budgets are drawn from the same distributions, this will yield very robust results – it generates $500,000$ random decision makers overall.[6] Using the two data sets generated by **A'** and **A**, we can then proceed to execute procedures **B** and **C**.

Before we analyse our consumers with procedure **C**, it is helpful to look at the distribution of the AEI computed by procedure **B** in the sets of random choices generated by procedure **A** for the same distribution of budgets. Figure 1 (dashed line) shows the power of the test for given critical levels of the AEI. It illustrates that, while for the requirement of full rationality ($e = 1$) the test power is very high (.966), small absolute deviations from 1 cause the power to drop very quickly. For a critical level e of the AEI of .95 we would already lose all power. The result illustrates that Varian's (1990) suggestion of a critical value of 95%—"for sentimental reasons"—is to be understood as a tongue-in-cheek remark: There is no "natural" critical value, and the choice of an appropriate critical value depends on the nature of the budgets on which we observe choices.

Define for convenience a function $g : [0, 1] \to \{0, \ldots, N\}$ as

$$g(e) = \begin{cases} 0 & \text{if SAEI}(N) < e \\ \max_{i \in \{1,\ldots,N\}} i \text{ such that SAEI}(i) \geq e & \text{otherwise,} \end{cases}$$

Then, $g(e)/N$ will give the fraction of observations which are accepted as close enough to GARP given an acceptable AEI of e, and $f(g(e))$ gives the power the test maintains against random behaviour. The results for $g(e)/N$ are also depicted in Figure 1 (solid lines) for three values of κ.

The result is shown in Figure 2: Here, we plot the power of the test – $f(g(e))$ – on

[6] Generating the random choices and computing their AEI took about three hours on an Intel i5-2500k with 3.3 GHz using parallel computing with four cores. All results are practically the same for any number of random decision makers between $10,000$ and $500,000$.

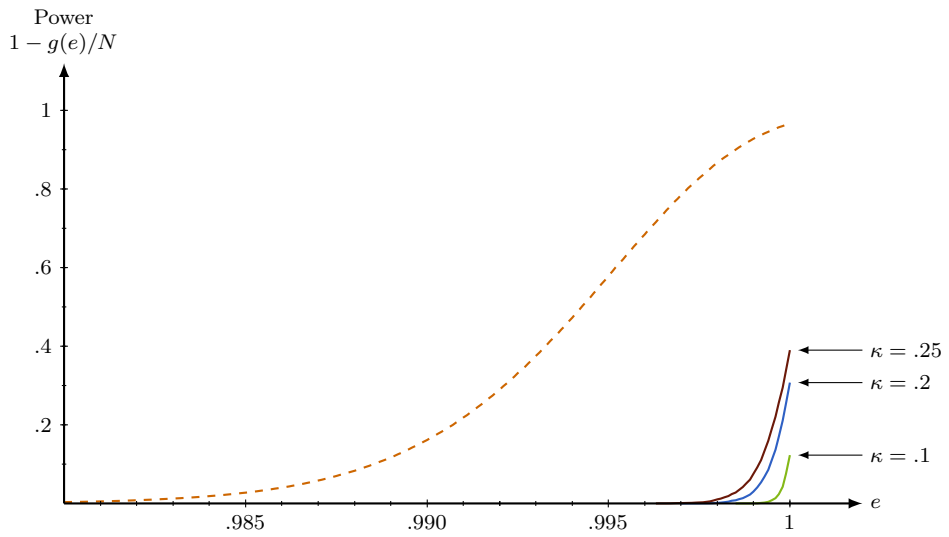

Figure 1: The dashed line shows the fraction of random decision makers who are not accepted as close enough to GARP (i.e., Bronars' power of the test) depending on the critical AEI e, for the budget sets used in procedure **A'**. The solid lines shows the fraction of consumers who are not accepted as close enough to GARP depending on e.

the horizontal axis and the fraction of observations accepted as close enough to utility maximization – $g(e)/N$ – on the vertical axis.

Note that, by the definition of statistical power, $1 - f(\cdot)$ gives the probability of making a Type II error. If we think about the data generated by procedure **A'** as data that, given that it comes from the maximization of a utility function, should without exception be accepted as close enough to GARP, than $1 - [g(f(\cdot))/N]$ gives the probability of making a Type I error. Note however that once we add stochastic error any statement about the "rationality" of such an observation is debatable. We may presume that for very low values of ε the observations should still be accepted, and that for very high values of ε the observations should not be accepted, but we cannot know what "low" and "high" values of ε are.

Figure 2 shows that for small stochastic errors we can indeed accept more than 98% of all sets of observations as close enough to GARP without losing any power at all, and that

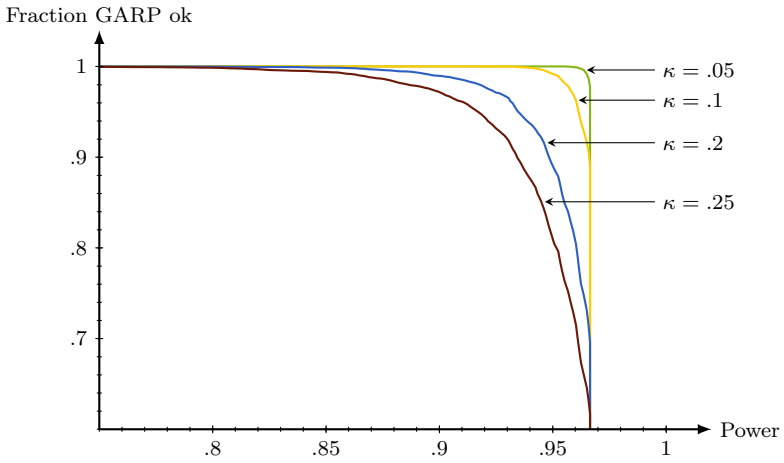

Figure 2: Simulated Data. The plots show the fraction of simulated consumers who are accepted as close enough to GARP depending on the maintained power.

we can accept 100% of all sets of observations with less than 2 percentage points of loss of power. However, for large stochastic errors, we can only accept about 60% of all sets of observations without any loss of power, while we would lose more than 16 percentage points of power if we accept all sets. This is not a weakness of the procedure; indeed, it reflects the fact that simulating random behaviour is a straightforward and easy way to get an idea about how to interpret the AEI, and that, as originally rational behaviour is distorted more and more, still accepting this behaviour as rational comes at an increasingly higher cost.

Selten (1991) argues in favour of a *measure of predictive success* for *area theories*, given by $m(r, a) = r - a$, where r is the *hit rate* and a is the *area*.[7] The hit rate is defined as "the relative frequency of correct predictions", and the area as "the relative size of the predicted subset compared with the set of all possible outcomes" (Selten 1991, p. 154). As discussed above, we do not know which of the observations we should accept as rational; thus, we cannot simply equate the hit rate with the fraction of accepted observations. However, we suggest to interpret the measure in the context of this article as follows: The prediction is that consumers who are close enough to utility maximization have at least an AEI of

[7] I would like to thank an anonymous referee for bringing this to my attention.

$e^* \in [0,1]$. Instead of interpreting $1 - e^*$ as the area of the prediction, we measure the area by the fraction of random choice sets which have at least an AEI of e^*. If we are then willing to call "accepting a consumer as close enough to utility maximization" a "success", then $m(r,a)$ serves as an objective function for the researcher. Note that this reinterpretation of $m(r,a)$ satisfies the same axioms as described by Selten (1991), in particular the axiom about cost-benefit evaluations (Axiom 4): Whether or not a prediction $m(r_1,a_1)$ is more successful than $m(r_2,a_2)$ depends only on the differences $r_1 - r_2$ and $a_1 - a_2$.

Figure 3 shows the measure for the data generated by procedure **A'**. Note that the "hit rate" for a prediction e is $g(e)/N$, and the area is given by $1 - f(g(e))$. As expected, the smaller the stochastic error, the greater is the maximum value obtainable for the objective function.

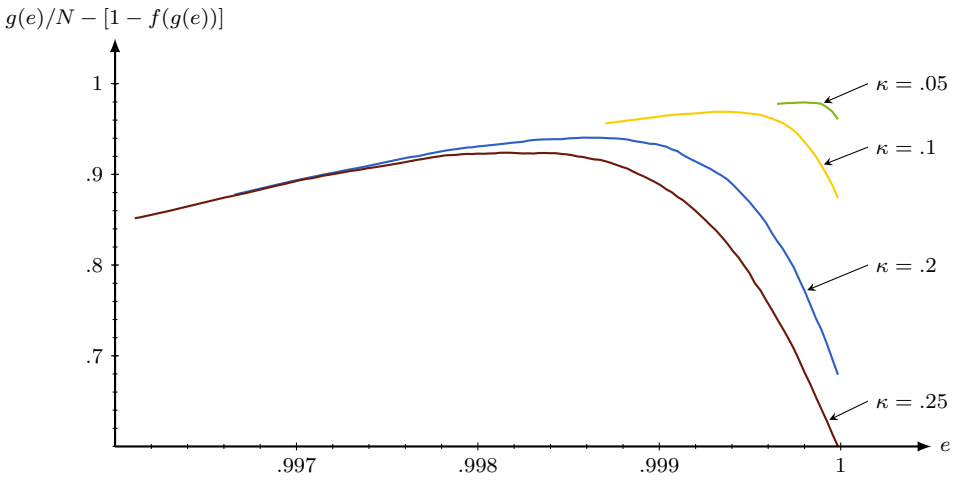

Figure 3: Simulated data. The difference between the fraction of consumers and the fraction of random choice sets accepted as close enough to utility maximization.

3.2 Experimental Data

To further illustrate the procedure, we use laboratory data from two experimental dictator games of Andreoni and Miller (2002) and Fisman et al. (2007). In the former experiment,

the same set of budgets was used for each subject. Bronars' power is based on the first algorithm with $50,000$ repetitions of procedure A. In the latter experiment, budgets were drawn randomly for each subject. Bronars' power is based on the first algorithm with 400 repetitions of procedure A for each subject's budget sets, resulting in $30,400$ repetitions.[8] Figure 4 replicates Figure 1 for the experimental data: It shows the fraction of random decision makers who are not accepted as close enough to GARP (i.e., Bronars' power of the test) depending on the critical AEI e.

The results are reported in Figure 5. Note that we can accept roughly 90% of the subjects of both experiments as "close enough" to utility maximization with a power of roughly 78%. However, this is the power based on the marginal subject; for all other accepted subjects, the subjects in the experiment of Fisman et al. (2007) "passed" a test which was, loosely speaking, more demanding (see also Figure 4). This is not surprising given that the subjects were asked to make choices on more budgets.[9]

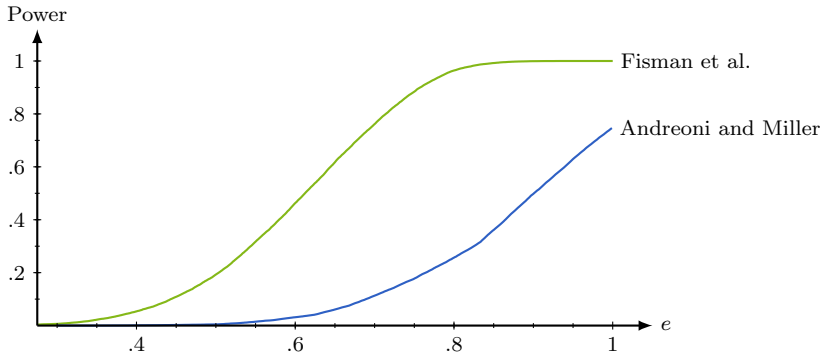

Figure 4: The plots show the fraction of random decision makers who are not accepted as close enough to GARP depending on the critical AEI e, for the budget sets used in Andreoni and Miller (2002) and Fisman et al. (2007).

Figure 6 shows the measure of success defined in Section 3.1 for both sets of experimental

[8]In Fisman et al. (2007), 50 budgets were drawn from identical distributions, hence all subjects faced very similar budgets.

[9]This is not to say that more budgets make a study necessarily more meaningful; here is another possible tradeoff between test power on the one hand and subject fatigue on the other hand.

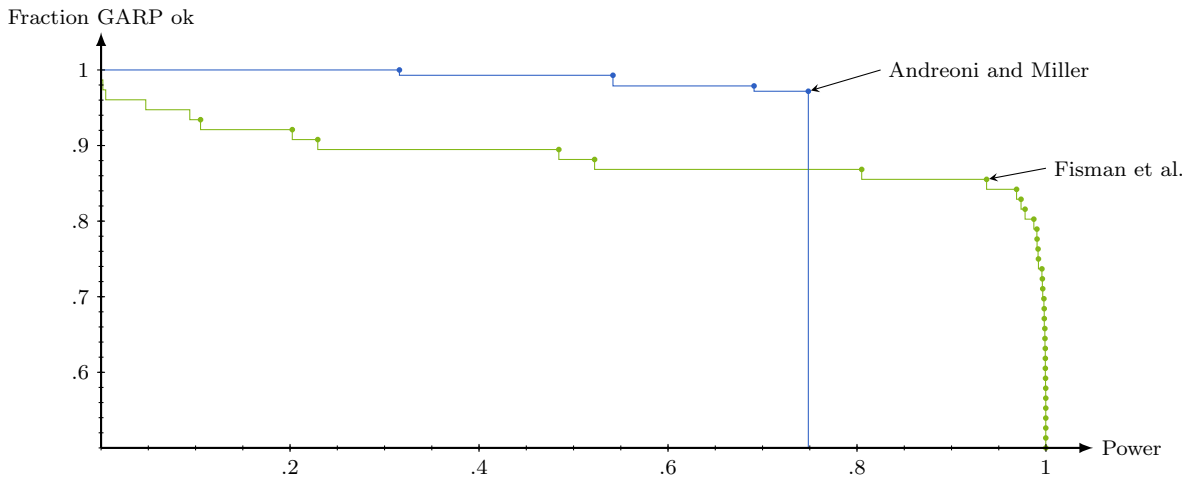

Figure 5: Data from Andreoni and Miller (2002) and Fisman et al. (2007). The plots show the fraction of subjects who are accepted as close enough to GARP depending on the maintained power.

data. The optimum for the Fisman et al. (2007) data is obtained by setting the critical value e of the AEI to .806, which allows to accept 81.1% of all subjects as close enough to utility maximization. Several subjects in the Andreoni and Miller (2002) experiments had an AEI of practically 1 resulting from GARP violations due to choices on the intersection of two budgets. These subjects therefore violated GARP, but passed the test for *any* $e < 1$, which explains the two points at $e = 1$. This problem does not occur for the random decision makers, thus the power of the test does not depend on whether or not these subjects with an AEI of $e - \varepsilon$ are accepted. The optimum for the Andreoni and Miller (2002) data is therefore obtained by setting the critical value to $e - \varepsilon$, which allows to accept 97.1% of all subjects.

4 Discussion and Conclusion

The application to simulated data shows that for reasonably small error terms one can accept almost all of the choices with stochastic error as close enough to utility maximizing

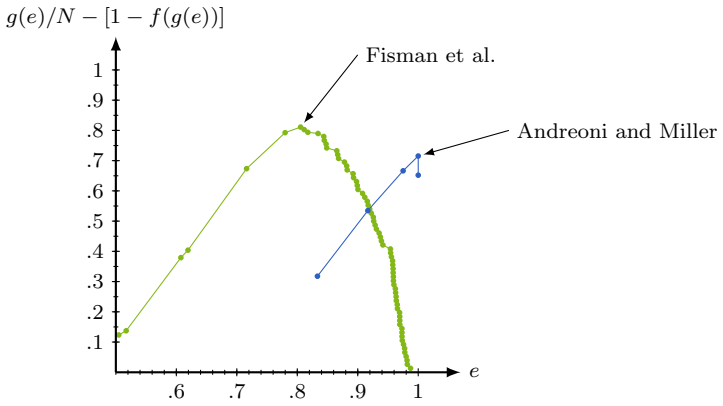

Figure 6: Data from Andreoni and Miller (2002) and Fisman et al. (2007). The difference between the fraction of consumers and the fraction of random choice sets accepted as close enough to utility maximization.

without losing much power. For larger error terms the loss of power is substantial. The application to experimental data shows that the experiments from Andreoni and Miller (2002) and Fisman et al. (2007) suffer from lack of power once we allow deviation from 100% efficiency for most subjects. With the introduction of the graphical presentation of budgets to subjects (Choi et al. 2007b) which allows to collect more data in experiments, we can expect to see more experiments of the kind of Choi et al. (2007a) and Fisman et al. (2007). Trading off power against accepting subjects as utility maximizers can be a delicate business. This article contributes to the analysis of such data.

The procedure described in this note can also be used to compare different methods of measuring the extent of violations of utility maximization. For this it is necessary to compute the different efficiency indices for simulated utility maximizing choices with stochastic errors and repeat this for Bronars' procedure. One can then compare the loss of power when basing the decision which observations to accept as close enough to utility maximizing on the different measures. If one of the two indices dominates the other one in the sense that for each marginal subject, the power from using the first index exceeds the power from using the second index, no assumption about the optimal tradeoff is necessary

if—ceteris paribus—a higher power is always desirable.

References

Afriat, S. N. (1967): The Construction of a Utility Function From Expenditure Data, *International Economic Review*, 8(1), 67-77.

——— (1972): Efficiency Estimation of Production Functions, *International Economic Review*, 13(3), 568–598.

Allais, M. (1953): Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine, *Econometrica, Economic Theory*, 21(4), 503–546.

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.

Banks, J., R. Blundell, and S. Tanner (1998): Is there a Retirement-Savings Puzzle?, *American Economic Review*, 88(4), 769–788.

Battalio, R. C., J. H. Kagel, R. C. Winkler, E. B. J. Fisher, R. L. Basmann, and L. Krasner (1973): A Test of Consumer Demand Theory Using Observations of Individual Consumer Purchases, *Western Economic Journal*, 11(4), 411-428.

Becker, G. S. (1962): Irrational Behavior and Economic Theory, *Journal of Political Economy*, 70(1), 1–13.

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

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

Choi, S., R. Fisman, D. M. Gale, and S. Kariv (2007b): Revealing Preferences Graphically: An Old Method Gets a New Tool Kit, *American Economic Review*, 97(2), 153-158.

Dickinson, D. L. (2009): Experiment Timing and Preferences for Fairness, *The Journal of Socio-Economics*, 38(1), 89-95.

Ellsberg, D. (1961): Risk, Ambiguity and the Savage Axioms, *Quarterly Journal of Economics*, 75(4), 643–669.

Famulari, M. (1995): A Household-Based, Nonparametric Test of Demand Theory, *Review of Economics and Statistics*, 2(2), 285–382.

Février, P. and M. Visser (2004): A Study of Consumer Behavior Using Laboratory Data, *Experimental Economics*, 7(1), 93–114.

Fisman, R., S. Kariv, and D. Markovits (2007): Individual Preferences for Giving, *American Economic Review*, 97(5), 1858-1876.

Fleissig, A. R. and G. A. Whitney (2003): A New PC-Based Test for Varian's Weak Separability conditions, *Journal of Business and Economic Statistics*, 21(1), 133–144.

——— (2005): Testing for the Significance of Violations of Afriat's Inequalities, *Journal of Business and Economic Statistics*, 23(3), 355–362.

Gross, J. (1995): Testing Data for Consistency with Revealed Preference, *Review of Economics and Statistics*, 77(4), 701–710.

Harbaugh, W. T., K. Krause, and T. R. Berry (2001): GARP for Kids: On the Development of Rational Choice Behavior, *American Economic Review*, 91(5), 1539-1545.

Heufer, J. (2008): A Geometric Measure for the Violation of Utility Maximization, *Ruhr Economic Papers*, #69, TU Dortmund University, Discussion Paper.

Kahneman, D. (1994): New Challenges to the Rationality Assumption, *Journal of Institutional and Theoretical Economics*, 150(1), 18–36.

Kahneman, D. and R. H. Thaler (2006): Utility Maximization and Experienced Utility, *Journal of Economic Perspectives*, 20(1), 221–234.

Lührmann, M. (2010): Consumer Expenditures and Home Production at Retirement – New Evidence from Germany, *German Economic Review*, 11(2), 225–245.

Mattei, A. (2000): Full-scale real tests of consumer behavior using experimental data, *Journal of Economic Behavior & Organization*, 43(4), 487-497.

Menkhoff, L., M. Schmeling, and U. Schmidt (2010): Are All Professional Investors Sophisticated?, *German Economic Review*, 11(4), 418–440.

Rapoport, A., W. E. Stein, J. E. Parco, and T. E. Nicholas (2003): Equilibrium Play and Adaptive Learning in a Three-Person Centipede Game, *Games and Economic Behavior*, 43(2), 239–265.

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

Sippel, R. (1996): A Note on the Power of Revealed Preference Tests with Afriat Inefficiency, sFB 303, Universität Bonn, Discussion Paper No. A-528.

——— (1997): An Experiment on the Pure Theory of Consumer's Behavior, *The Economic Journal*, 107(444), 1431-1444.

Varian, H. R. (1982): The Nonparametric Approach to Demand Analysis, *Econometrica*, 50(4), 945-972.

——— (1990): Goodness of Fit for Revealed Preference Tests, university of Michigan CREST Working Paper Number 13.

Acknowledgements

This paper is drawn from doctoral research done at Ruhr Graduate School in Economics at TU Dortmund University under the guidance of Wolfgang Leininger. I am grateful for his support and comments. Thanks to the editor and an anonymous referee for very helpful suggestions and comments which improved the quality substantially. Thanks to Yiquan Gu and Anthony la Grange for helpful comments. Thanks to James Andreoni and John Miller, and Raymond Fisman, Shachar Kariv, and Daniel Markovits for access to their data. The work was financially supported by the Paul Klemmer Scholarship of the RWI Essen, which is gratefully acknowledged.