

A Stochastic Test of the Number of Revealed Preference Violations*

Per Hjertstrand

Department of Economics, Lund University, Sweden and Statistics unit, Hanken School of Economics, Finland[†]

James L. Swofford

Department of Economics and Finance, University of South Alabama[‡]

March 2009

Abstract

Revealed preference tests are frequently used to check data on the behavior of agents for consistency with economic theory. Unfortunately these tests lack a stochastic element and thus one violation of revealed preference conditions causes a rejection of the behavior being tested. To remedy the lack of a stochastic element in revealed preference analysis, we suggest a simple and intuitive statistical procedure to test whether the observed number of violations is more consistent with rational behavior than uniform random behavior. This statistical test takes advantage of the fact that nonparametric revealed preference tests involve known prices and expenditures. Thus, the actual number of violations can be compared to the number of violations when behavior is uniform random. Simulation results show that the test has very good small sample properties. We implement the procedure using data from two well known economic experiments. First we test the data on altruism and second we test the choices of children. Our results suggest some different interpretations of the behavior of the experimental subjects than those implied by non-stochastic revealed preference results.

Keywords: Nonparametric; Revealed preference; Stochastic.

JEL Classification: C12; D12.

*We thank David Edgerton, Barry Jones and seminar participants at the 'Measurement Error: Econometrics and Practice' conference in Birmingham, UK in July 2007 for comments on an earlier version of the paper. We also thank William Harbaugh and James Andreoni for providing the data. Additionally, Hjertstrand thank the Department of Economics and Finance, University of South Alabama for hospitality during a visit, and the Jan Wallander and Tom Hedelius Foundation for research support.

[†]Address: Department of Economics, Lund University, P.O. Box 7082, S-220 07 Lund. email: per.hjertstrand@nek.lu.se.

[‡]Correspondence to: James L. Swofford, Department of Economics and Finance, University of South Alabama, Mobile, AL 36688, Phone: (251) 460-6705, email: jswoffor@usouthal.edu.

1 Introduction

Samuelson (1938) set forth the necessary condition under which data are revealed to be consistent with rational behavior. This condition became known as the Weak Axiom of Revealed Preference (WARP). Houthakker (1950), Afriat (1967) and Varian (1982) extended WARP and provided for implementation.

Empirical papers applying tests of revealed preference have appeared in various areas of research. Swofford and Whitney (1986) used these test procedures to examine if certain subsets of monetary assets are consistent with demand theory. More recent examples include Harbaugh et al. (2001) who used tests of revealed preference to study whether children in an experimental setting chose rationally and Andreoni and Miller (2002) who carried out experimental tests on the consistency of preferences for altruism.

The advantage of using revealed preference tests are that they are very easy to implement, do not require a large number of observations and do not assume any functional form for the utility function. The disadvantages associated with these procedures are that they are not tests in the usual statistical sense and as a consequence do not attach any stochastic element to the analysis. Consequently, revealed preference tests finding a single violation in a data set will reject rational behavior.

It is well known that there may be a stochastic element in the data due to various reasons. For example, agents may make optimizing errors, agents may be learning by doing or there may be measurement error in the data. Thus, a researcher finding violations in a particular data set would naturally ask if the number of violations is too many to be consistent with rational behavior.

Prior discussions of the violations from revealed preference tests can be divided into three main areas, discussions of the number of violations, discussions of the size of the violations and discussions of the power of revealed preference tests. Gross (1995) reviews the prior research and rejects the previous proposals for evaluating the number of violations of revealed preference in these papers as impractical or insufficiently informative.

Among the studies concerning the size of violations are Gross (1995), Fleissig and Whitney (2005), de Peretti (2005), Varian (1985) and Epstein and Yatchew (1985). All of these procedures are based on the idea of calculating the minimal adjustment that is sufficient for a data set to satisfy revealed preference and, in a second step, testing whether this adjustment is statistically significant.

Varian (1985) propose to compute the minimal perturbation to the observed data by minimizing the sum of the squared differences between the adjusted and observed quantities, under the restriction that the adjusted data satisfies revealed preference. Varian further suggest a test statistic of the null hypothesis that the data without measurement errors satisfy the revealed preference axiom, but the observed data violates it due to the fact that the data has been measured with errors. By assuming that the measurement errors follow a Gaussian distribution, the test statistic can be shown to be chi-squared distributed.

Recently, Fleissig and Whitney (2005) proposed two test procedures. For their lower bound test, they propose adding measurement errors to the quantity data while holding the expenditure and price data constant, and testing this perturbed data set for rationality. This is repeated a number of times to obtain an empirical distribution of the measurement errors. The other test, referred to as the upper bound test is based on adding slack terms to allow violations of rationality. A test statistic is construct by minimizing the maximum slack term required for the data to satisfy rationality, which can be evaluated by comparing its value to a simulated distribution.

Although these procedures have their advantages, they also suffer from disadvantages that may be serious. First, both Varian's and Fleissig and Whitney's procedures rely on the assumption that either the quantity or price data is measured with errors. That is, neither of the approaches are able to si-

multaneously model measurement errors in the quantity and price data. Second, they require that the researcher have an *a priori* measure of the measurement error variance, which can be difficult to obtain when analyzing data collected from different sources. The methods also require that the measurement errors are independent and identically-distributed (*i.i.d*) and follows a known distribution, which perhaps may be a too restrictive assumption for many applications. Finally, Varian’s method may require the solution to a computationally very burdensome nonlinear optimization problem and there can be difficulties obtaining an efficient solution, since there may exist many local minima and saddle points.

Bronars (1987) is an example of a few papers that discuss the power of revealed preference tests. This literature has generally found revealed preference tests to have reasonable power.¹

In this paper we address the question of how many violations of revealed preference that are enough to indicate behavior systematically inconsistent with economic theory. We propose a statistical test of the type of behavior that is being observed. This test allows us to say whether behavior is consistent with uniform random behavior or some type of systematic behavior such as rational behavior. Our approach is quite different from those mentioned above. In particular, we ask if there is a systematic way of determining when the number of violations are few enough that revealed preference cannot be said to be rejected by the data. For this purpose, we derive a test that relaxes the potentially restrictive assumptions about the data generating process that is required by Varian’s (1985) and Fleissig and Whitney’s (2005) procedures.

Since nonparametric tests of revealed preference involves known finite budget sets, all the information needed to conduct the proposed test is already in the hands of researchers conducting revealed preference analysis. One can use the data to test whether or not the actual number of violations of revealed preference is consistent with either uniform random behavior, rational behavior or economically perverse behavior.

Simulation results indicate that our test has very good small sample properties when applied to Cobb-Douglas utility functions. The test is shown to have very good size and power, and thus correctly detects rational and random behavior when data are generated under those hypotheses.

We apply the test to two experimental data sets. First, we use the data of Andreoni and Miller (2002) to investigate the systematic behavior of altruism. In contrast to their results, our results suggest that the altruistic choices in the experiments may not be as rational as previously thought. Second, we use data drawn from experiments conducted by Harbaugh et al. (2001) to test the systematic behavior of children. We find that children without violations of revealed preference seem to conform to rational behavior. Many of Harbaugh et. al. subjects were also found to choose randomly, and we even found a few subjects to exhibit what Becker (1962) termed perverse economic behavior.

Before discussing the test or any of the results in detail, we begin by reviewing some basics of revealed preference.

2 Revealed preference

Let there be n observations on the prices, $\mathbf{p}^i = (p_1^i, \dots, p_K^i)$; $i = 1, \dots, n$, of some K goods and assets and let $\mathbf{x}^i = (x_1^i, \dots, x_K^i)$; $i = 1, \dots, n$, denote the quantities of the K goods and assets. Samuelson (1938) and Houthakker (1950) set forth the following definition.

Definition 1 *Given an observation \mathbf{x}^i and bundle \mathbf{x} , we say that*

¹Power indices of revealed preference tests are discussed in detail in Andreoni and Harbaugh (2006), who also propose four new indices.

- (i) \mathbf{x}^i is directly revealed preferred to \mathbf{x} , written $\mathbf{x}^i D\mathbf{x}$, if $\mathbf{p}^i \mathbf{x}^i \geq \mathbf{p}^i \mathbf{x}$
- (ii) \mathbf{x}^i is strictly directly revealed preferred to \mathbf{x} , written $\mathbf{x}^i S\mathbf{x}$, if $\mathbf{p}^i \mathbf{x}^i > \mathbf{p}^i \mathbf{x}$.
- (iii) \mathbf{x}^i is revealed preferred to \mathbf{x} , written $\mathbf{x}^i R\mathbf{x}$, if $\mathbf{p}^i \mathbf{x}^i \geq \mathbf{p}^i \mathbf{x}^j, \mathbf{p}^j \mathbf{x}^j \geq \mathbf{p}^j \mathbf{x}^l, \dots, \mathbf{p}^m \mathbf{x}^m \geq \mathbf{p}^m \mathbf{x}$ for some sequence of observations $(\mathbf{x}^i, \mathbf{x}^j, \mathbf{x}^l, \dots, \mathbf{x}^m)$. In this case, the binary relation R is called the transitive closure of the relationship D .

In terms of this notation and definition, the Weak Axiom of Revealed Preference (WARP) is defined as follows.

Definition 2 A data set satisfies the Weak Axiom of Revealed Preference (WARP) if $\mathbf{x}^i D\mathbf{x}^j$ implies $\mathbf{p}^j \mathbf{x}^j < \mathbf{p}^j \mathbf{x}^i$, with $\mathbf{x}^i \neq \mathbf{x}^j$, for all $i, j = 1, \dots, n$.

A violation of WARP occurs when for some $\mathbf{x}^i D\mathbf{x}^j$, the condition $\mathbf{x}^j S\mathbf{x}^i$ is true or a violation of WARP happens if \mathbf{x}^i is shown to be directly revealed preferred to \mathbf{x}^j but \mathbf{x}^j is strictly directly revealed preferred to \mathbf{x}^i . Thus, in the two good example in Figure 1, the only way to have a violation of WARP is for one consumption bundle to fall on DE of budget constraint AE while the other consumption bundle falls on segment CD of budget set CF . In this case, the choice of the first bundle will violate the choice of the second bundle, and the choice of the second bundle will violate the choice of the first bundle, meaning there will be one pairwise violation. The total possible number of pairwise violations of WARP in a data set of n observations is $\frac{n(n-1)}{2}$.

Houthakker (1950) recognized that WARP is a necessary and sufficient condition for rationality only when K equals 2. In the multidimensional case WARP is a necessary, but not sufficient condition for rationality. Houthakker (1950) introduced the Strong Axiom of Revealed Preference (SARP), which is necessary and sufficient for consistency with neoclassical economic theory when $K > 2$.^{2,3}

Definition 3 A data set satisfies the Strong Axiom of Revealed Preference (SARP) if $\mathbf{x}^i R\mathbf{x}^j$ implies $\mathbf{p}^j \mathbf{x}^j < \mathbf{p}^j \mathbf{x}^i$, with $\mathbf{x}^i \neq \mathbf{x}^j$, for all $i, j = 1, \dots, n$.

WARP and SARP provides simple static testable conditions that have been extensively implemented in various applications. These test procedures, however, are quite restrictive. In particular, one might think that a data set that involves fewer violations of WARP or SARP was generated by behavior more consistent with economic theory than a data set that involves more violations. However, such a rule of thumb ignores how much opportunity the consumer has to violate revealed preference. If prices and total expenditure change in such a manner that no budget sets intersect in the positive quadrant, there can be no violations of revealed preference. In this case, no matter where the agent consumes on the budget, there can be no violation of the weak or strong axioms. Since it gives the researcher no information concerning consumer behavior, in such a case there is no information in finding zero violations of WARP or SARP. Thus, one violation may be either very important if it is the only one that could have occurred or it may be insignificant if only one violation occurs when many violations were possible. This discussion is closely related to the power of static revealed preference tests, see

²Varian (1982) introduced the Generalized Axiom of Revealed Preference (GARP). This condition is very similar to SARP, with the only difference being that GARP allows for multi-valued demand functions, while SARP only allows for single-valued demand functions.

³A violation of SARP in the case $K = 2$ implies indeed a violation of WARP. However, it should be pointed out that the number of violations of WARP and SARP in a data set may differ in the case when $K = 2$, since there may be violations through the transitive closure of the binary relation in SARP.

Andreoni and Harbaugh (2006). Blundell et al. (2004) suggested one possible solution to this problem by combining non-parametric revealed preference procedures with parametric methods over regions of the data where the non-parametric tests may lack power. Our test, although quite different, uses the stochastic properties contained in the revealed preference violations to increase the power of revealed preference methods.

Another related problem is that, if there is a single violation of WARP or SARP, then the hypothesis that the data is generated by a well-behaved utility function is rejected. This follows from the non-stochastic nature of revealed preference tests. In other words, these procedures does not attach any measure of uncertainty such as sampling variance or confidence interval to the analysis. This is very restrictive particularly for the researcher who may be using nonparametric analysis to study whether some observed data is consistent with economic theory. Such a researcher finding few violations may be tempted to attribute these violations to measurement error. This is particularly true of a researcher pre-testing data to narrow the scope of stochastic parameter estimation. To remedy these problems, we set forth in the next section a statistical test of the number of violations of revealed preference.

3 A statistical test of observed behavior

3.1 The stochastic approach to revealed preference

Let us first consider a general discussion of the stochastic approach to revealed preference. Recall that n denotes the number of observed bundles and that the total number of pairwise comparisons in a data set is equal to $N = \frac{n(n-1)}{2}$. A convenient way to record a violation of revealed preference is to define a variable I_j for $j = 1, \dots, N$ as

$$I_j = \begin{cases} 1 & \text{if comparison } j \text{ is a violation} \\ 0 & \text{otherwise.} \end{cases}$$

The number of violations of revealed preference in the standard (static) revealed preference procedures is simply the sum

$$V = \sum_{j=1}^N I_j.$$

Assume now that I_j is a random variable and define a new variable $\pi_j = P(I_j = 1)$, which is the probability that the j :th comparison is a violation. In this case, V is the sum of Bernoulli trials with varying probabilities of success, and will be distributed as a weighted Binomial distribution. The expected value, denoted EV , of V is

$$\begin{aligned} EV &= E(V) = \sum_{j=1}^N E(I_j) = \sum_{j=1}^N P(I_j = 1) \\ &= \sum_{j=1}^N \pi_j, \end{aligned}$$

and the variance is given by

$$\begin{aligned}
\sigma_V^2 &= \text{Var}(V) = \sum_{j=1}^N \text{Var}(I_j) + 2 \sum_{j=2}^N \sum_{i=1}^{j-1} \text{Cov}(I_j, I_i) \\
&= \sum_{j=1}^N \left(E(I_j^2) - E(I_j)^2 \right) + 2 \sum_{j=2}^N \sum_{i=1}^{j-1} \left(E(I_j I_i) - E(I_j) E(I_i) \right) \\
&= \sum_{j=1}^N \pi_j (1 - \pi_j) + 2 \sum_{j=2}^N \sum_{i=1}^{j-1} \left(E(I_j I_i) - \pi_j \pi_i \right), \tag{1}
\end{aligned}$$

where $E(I_j I_i) = P(\text{violation in comparison } j \cap \text{violation in comparison } i)$ is the joint probability of violations in comparisons j and i . As a special case, if all violations I_j and I_i are independent, then the variance collapses to

$$\sigma_V^2 = \sum_{j=1}^N \pi_j (1 - \pi_j).$$

A test of consumer behavior can now be formulated as follows. Let the variable V be the number of violations from a standard static revealed preference test, i.e. WARP, SARP or GARP. Under the assumption of some specific behavior, we can compare the observed number of violations given by V with the expected number of violations, EV from the assumed behavior. A test of the hypothesis that the observed data is generated under the assumed consumer behavior may therefore be written as $H_0 : V = EV$ against the alternative $H_1 : V \neq EV$. The test statistic is calculated as

$$Z = \frac{V - EV}{\sigma_V} \sim N(0, 1), \tag{2}$$

which is asymptotically normally distributed because of the asymptotic normality of Bernoulli trials. Finally, it is useful to note that σ_V in (2) can be replaced by a consistent estimate without changing the asymptotic distribution of Z .

3.2 Random consumer behavior

What remains from the previous discussion is to specify a suitable hypothesis of consumer behavior. One could, of course, model rationality as the null hypothesis against the alternative of irrational behavior. However, such a specification require that we know the variance of how agents chose bundles on the budget set. But this is impossible without specifying a particular model of rational behavior. Therefore, since we expect agents to be rational and since there are many possible models of rational behavior, it seems more logical to specify a given irrational type of consumer behavior against the notion of unspecified rational behavior.

A natural hypothesis of irrational behavior is uniform or random consumer behavior. Becker (1962) showed that random behavior is all that is required to get downward sloping demand curves.⁴ This choice of irrational behavior have frequently been employed as an alternative hypothesis to rationality; one notable example being Bronars (1987) who constructed a method for calculating the approximate power of revealed preference tests for the simple alternative hypothesis of random behavior. Varian (2006) remarked that there currently does not seem to exist other more suitable hypotheses besides Becker's, that is amenable as an alternative to rational behavior, and can be applied with the same sorts of data as for revealed preference tests.

⁴See also McCausland (2009) who present an alternative theory of random consumer demand.

Consider again Figure 1, and assume that consumer behavior is uniform random. Under the null hypothesis of random behavior, the probability of a violation of the weak axiom (WARP) is⁵

$$P(I = 1 | H_0 : \text{Random behavior}) = \pi^0 = \frac{DE}{AE} \times \frac{CD}{CF},$$

and the variance of this violation is

$$Var(\pi^0) = \pi^0(1 - \pi^0) = \frac{DE}{AE} \times \frac{CD}{CF} \times \left(1 - \frac{DE}{AE} \times \frac{CD}{CF}\right)$$

The expected number of violations in the data set is thus

$$EV^0 = \sum_{j=1}^N \pi_j^0,$$

and the variance $\sigma_{V^{obs}}^2$ is given by (1) with π_j replaced by π_j^0 .

A statistical test of the number of violations of revealed preference under random behavior can be formulated as

$$\begin{aligned} H_0 & : V^{obs} = EV^0, \text{ i.e. Observed behavior is uniform or random,} \\ H_1 & : V^{obs} < EV^0, \text{ i.e. Observed behavior is consistent with rational choice,} \\ H_2 & : V^{obs} > EV^0, \text{ i.e. Observed behavior is economically perverse,} \end{aligned}$$

where V^{obs} is the number of violations calculated from the static revealed preference test. The test statistic is

$$Z^{obs} = \frac{V^{obs} - EV^0}{\sigma_{V^{obs}}} \sim N(0, 1). \quad (3)$$

If the Z^{obs} -value is negative and significant, then the number of violations is more consistent with rational consumer theory than uniform random behavior on the part of consumers. In other words, the null hypothesis, H_0 , in this case is rejected in favor of the alternative hypothesis of rational behavior, H_1 . Second, if the Z^{obs} -value is not significant, then the null hypothesis that the observed behavior is consistent with uniform random behavior cannot be rejected. Finally, if the Z^{obs} -value is positive and significant, then the number of violations is consistent with perverse economic behavior on the part of consumers.

Note as mentioned above, our test does not require many of the restrictive assumptions that are required in the size of violation procedures of Varian (1985) and Fleissig and Whitney (2005). In particular, the test proposed here does not require any specification of how the observed data is generated or how the measurement errors affects the true data, nor does it require the researcher to specify the distribution or amount of the measurement errors. Yet another advantage of our test is that it can be applied to data with errors in both the prices and quantities.

Since we expect agents to be rational, it is important that our test has good power properties. Thus, we want our test to be able to reject random behavior when true behavior is rational. This is confirmed by the simulation results in Section 5.2.

There might be a concern about the test statistic (3), because of the fact that EV^0 has to be calculated. However, since EV^0 is a function only of the exogenously given prices and income, this value is indeed non-stochastic, and can therefore be compared to the outcome of the stochastic process V^{obs} .

⁵All other combinations of observed behavior lead to zero violations in Figure 1. Thus, the random probability of their occurrences do not affect the expected number of violations.

One may speculate that using a stronger condition for rationality than the weak axiom is more suitable in applications. This would be at the expense of increasing the computational burden. Andreoni and Harbaugh (2006) remarked on this issue for one of their methods for calculating the power of revealed preference, by saying that even if their procedure is also applicable to a stronger notion of rationality, it would be more difficult, with dubious net benefit. Of course, one may argue the same for the test proposed here. Additionally, even when applied to WARP, as we have pointed out, simulation evidence in Section 5.2 suggest the test to be very powerful.

Finally, a problem that arises when the variance in (1) is calculated concerns the covariance terms. As argued by Aizcorbe (1991), the choices under random behavior are indeed correlated since there are pairwise comparisons that share budget sets. For pairwise comparisons that do not involve the same budget sets, independence will hold because outcomes for one pairwise comparison have no implications for the other. While Aizcorbe (1991) did not provide explicit formulas for calculating the covariance we will show that they are easily derived, and although carried out in the two-dimensional case can be straightforwardly extended to the multi-dimensional case.

Consider once again Figure 1. In this case, we compare choices on budget lines AE and CF . Now, consider Figures 2 and 3 and compare the choice on budget line AE with a new budget GI . Since Figures 1-3 share the same budget line AE , the probability of a violation will be conditional upon the outcome of the other comparisons, and they are all therefore dependent.

Recall first that under random behavior, the unconditional probability of a violation in Figure 1 is

$$P(\text{violation in Figure 1}) = \frac{DE}{AE} \times \frac{CD}{CF}.$$

We have two cases depending on if $DE < HE$ or $DE > HE$. We begin with the first case $DE < HE$, depicted in Figure 2. The unconditional probability of violating rationality in Figure 2 is

$$P(\text{violation in Figure 2}) = \frac{GH}{GI} \times \frac{HE}{AE}.$$

Since $DE < HE$ and given that there were a violation in Figure 1, we know with certainty that the agent has chosen in the rejection region of budget AE in Figure 2. Hence, the conditional probability of a violation in this figure given that a violation has occurred in Figure 1 must equal

$$P(\text{violation in Figure 2} \mid \text{violation in Figure 1}) = \frac{GH}{GI}.$$

Now using Bayes theorem, we may derive the joint probability of violations in Figures 1 and 2 as

$$P(\text{violation in Figure 2} \cap \text{violation in Figure 1}) = P(\text{violation in Figure 2} \mid \text{violation in Figure 1}) \times P(\text{violation in Figure 1}),$$

so that

$$P(\text{violation in Figure 2} \cap \text{violation in Figure 1}) = \frac{GH}{GI} \times \frac{DE}{AE} \times \frac{CD}{CF}.$$

Since the covariance between violations equals $Cov(I_j, I_i) = E(I_j I_i) - E(I_j) E(I_i) = P(I_j \cap I_i) - \pi_j \pi_i$ from (1), we have

$$\begin{aligned} Cov(\text{violation in Figure 2}, \text{violation in Figure 1}) &= \frac{GH}{GI} \times \frac{DE}{AE} \times \frac{CD}{CF} - \frac{DE}{AE} \times \frac{CD}{CF} \times \frac{GH}{GI} \times \frac{HE}{AE} \\ &= \frac{GH}{GI} \times \frac{DE}{AE} \times \frac{CD}{CF} \times \left(1 - \frac{HE}{AE}\right) \\ &= \frac{GH}{GI} \times \frac{DE}{AE} \times \frac{CD}{CF} \times \frac{AH}{AE}. \end{aligned}$$

Now, consider the second case, $DE > HE$ depicted in Figure 3. The unconditional probability of a violation is

$$P(\text{violation in Figure 3}) = \frac{GH}{GI} \times \frac{HE}{AE}.$$

Given $DE > HE$ and the fact that there were a violation in Figure 1, we know that the agent would not necessarily chose a bundle in the rejection region HE of budget AE in Figure 3. In fact, the probability of having chosen a bundle in the rejection region is

$$\frac{HE}{DE}.$$

Thus the probability of violating rationality in Figure 3 given that a violation has occurred in Figure 1 is

$$P(\text{violation in Figure 3} \mid \text{violation in Figure 1}) = \frac{HE}{DE} \times \frac{GH}{GI}.$$

Applying Bayes theorem gives the joint probability of violating rationality in Figures 1 and 3 as

$$\begin{aligned} P(\text{violation in Figure 3} \cap \text{violation in Figure 1}) &= \frac{HE}{DE} \times \frac{GH}{GI} \times \frac{DE}{AE} \times \frac{CD}{CF} \\ &= \frac{HE}{AE} \times \frac{GH}{GI} \times \frac{CD}{CF}, \end{aligned}$$

and the covariance to be

$$\begin{aligned} Cov(\text{violation in Figure 3}, \text{violation in Figure 1}) &= \frac{HE}{AE} \times \frac{GH}{GI} \times \frac{CD}{CF} - \frac{DE}{AE} \times \frac{CD}{CF} \times \frac{GH}{GI} \times \frac{HE}{AE} \\ &= \frac{GH}{GI} \times \frac{HE}{AE} \times \frac{CD}{CF} \times \left(1 - \frac{DE}{AE}\right) \\ &= \frac{GH}{GI} \times \frac{HE}{AE} \times \frac{CD}{CF} \times \frac{AD}{AE}. \end{aligned}$$

This analysis shows that the covariance is always greater than zero, but cannot be larger than $0.5^4 = 0.0625$, and will differ across violations if $DE \neq HE$.⁶

4 Implementing the test

This section shows how to implement the test in a K -dimensional setting. All information that is needed to implement the test is contained in the ratio between the volume of the rejection region and the volume of entire budget set. Consequently, we need only to provide a procedure for calculating that ratio.

Assume, as before, that the agent faces at times $i, j = 1, \dots, n; i \neq j$, two different budget sets of K goods. It is well-known that the budget set in this case is a $K - 1$ dimensional simplex, for example, corresponding to a line segment (see Figures 1-3) in the bidimensional case, a triangle in the trivariate case, and a tetrahedron in the four-dimensional case.

Let us consider the trivariate case in Figure 4 in more detail. Since the two simplices intersect, each triangle is divided into two closed half-spaces that cuts the simplex. Define the area where a violation of revealed preference is possible in simplex A as V_A and the violation area for simplex B as V_B .⁷ The probability of violating rationality is thus

$$P(V_A \cup V_B) = \frac{Vol(V_A)}{Vol(A)} \times \frac{Vol(V_B)}{Vol(B)},$$

⁶Note that the covariance is equal across violations if and only if $DE = HE$ for all comparisons.

⁷Note that simplex, or triangle A in Figure 4 is projected in such a way that the triangle (simplex) B is only seen as a line in the three-dimensional space.

where Vol in this case is an area, and would correspond to a volume when $K = 4$ and the hypervolume when $K > 4$. Given that the intersecting simplex is a separating convex hyperplane defined on the positive orthant, the half-spaces above and below this intersection will be convex. More importantly, the segment in the budget set where a potential violation of revealed preference is possible will therefore be a $K - 1$ convex polytope, see Ziegler (1998). This suggests that calculating the ratio between the volume of a cut-off simplex and the volume of the simplex itself - even if simple in the two and three dimensional cases is rather difficult in higher dimensions. Fortunately, Leroux (2000) provided a solution to this problem.

Let there be a $K - 1$ simplex with vertices $a_h; h = 1, \dots, K$ defined on the K dimensional space. Denote the line supporting two vertices a_h and a_k by Φ_{hk} , and let $H(v) \in \mathbb{R}^K$ be a hyperplane with normal vector v that cuts the simplex. It is assumed that H is not parallel to any face of the simplex. That is, for all $h, k = 1, \dots, K; h \neq k, \langle \Phi_{hk}, v \rangle \neq 0$, where $\langle \cdot, \cdot \rangle$ denotes the inner product. Moreover, H cuts Φ_{hk} at one point, denoted indifferently b_{kh} and b_{hk} . Thus, the point $b_{kh} \equiv b_{hk}$ is a linear combination of the vertices connecting Φ_{hk} , such that $b_{kh} = b_{hk} = \lambda_{kh}a_k + \lambda_{hk}a_h$, with $\lambda_{kh} + \lambda_{hk} = 1$, where λ_{kh} are the weights. Leroux (2000) defines

$$P_k = \prod_{\substack{h=1 \\ h \neq k}}^K \lambda_{kh},$$

to be the 'power' of a_k relatively to H '. Now, index the vertices of the simplex such that $a_j; j = 1, \dots, J$ belongs to the closed half-space of \mathbb{R}^K bounded by the hyperplane H . As an example, a_j would be the vertex belonging to the polytope given by V_A of simplex A in Figure 4. Leroux (2000, Theorem 2-4) derives the ratio between the $K - 1$ volume of the cut-off simplex and the $K - 1$ volume of the simplex itself to be the sum of the 'powers' of the J vertices belonging to the bounded half-space, or more formally

$$\frac{Vol(\text{half-space})}{Vol(\text{simplex})} = \sum_{j=1}^J \prod_{\substack{k=1 \\ k \neq j}}^K \lambda_{jk}.$$

A proof of this result can be found in Leroux (2000).

5 Small sample properties

This section investigates the small sample properties of our test. The size of the test is investigated by generating data under the null hypothesis of random consumer behavior. In addition, we investigate the power of the test by generating data under the alternative hypothesis of rational behavior.

5.1 Simulation setups

The first simulation design is aimed to calculate the size of the test. For this purpose, we generate quantities for $K = 5$ goods and $n = 40$ observations using Algorithm 2 in Bronars (1987). This algorithm draws for each good and observation a random variable Z from a uniform distribution $U_{[0,1]}$, and calculates $S_k^i = Z_k^i / \sum_{k=1}^5 Z_k^i$ for $k = 1, \dots, 5$ and $i = 1, \dots, 40$.⁸ The random quantities are then given by

$$x_k^i = S_k^i \frac{m^i}{p_k^i},$$

⁸Hence, $\sum_{k=1}^5 S_k^i = 1$ for all $i = 1, \dots, n$.

where $m^i = \mathbf{p}^i \mathbf{x}^i$ is the total expenditure measured at time i , and p_k^i is the price for good k and observation i . Following Gross (1995) and Fleissig and Whitney (2003 and 2005), m^i is generated from $U_{[10000,12000]}$ and p_k^i ; $k = 1, \dots, 5$ and $i = 1, \dots, 40$ from either $U_{[90,100]}$ or $U_{[95,100]}$.⁹ Hence, we consider in the first design a total of two cases depending on the price distribution.

The second simulation design calculates the power of the test under the alternative hypothesis of rational behavior. This simulation setup as in Gross (1995) and Fleissig and Whitney (2003 and 2005) consists of random samples of $n = 40$ observations and $K = 5$ goods from standard Cobb-Douglas utility functions

$$U(\mathbf{x}) = x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3} x_4^{\beta_4} x_5^{\beta_5},$$

subject to $\sum_{k=1}^5 \beta_k = 1$. This yields Marshallian demand functions

$$x_k^i = \beta_k \frac{m^i}{p_k^i}, \quad (4)$$

where m^i as before is the total expenditure, and p_k^i the prices. As in the first design, m^i is generated from $U_{[10000,12000]}$ and p_k^i from $U_{[90,100]}$ or $U_{[95,100]}$.

We consider two sets for the preference parameters (or expenditure shares) β_k . The first, referred to in our simulations as β_A is given by: $\beta_{A,1} = 0.60$, $\beta_{A,2} = 0.25$, $\beta_{A,3} = 0.10$, $\beta_{A,4} = 0.04$, $\beta_{A,5} = 0.01$, and hence set so that the expenditure share for the first good is relatively large compared to the others. The other, referred to as β_B consists of expenditure shares set more similar to each other and is given by: $\beta_{B,1} = 0.40$, $\beta_{B,2} = 0.30$, $\beta_{B,3} = 0.15$, $\beta_{B,4} = 0.10$, $\beta_{B,5} = 0.05$.

To resemble what is usually observed in actual data, we let the quantities be measured with errors. Thus, after having generated 40 observations of the total expenditure m and prices \mathbf{p} , we calculate the quantities without errors, x_k^i from (4) using either β_A or β_B . In the next step, x_k^i is multiplied with errors drawn from a uniform distribution $\varepsilon_k^i \sim U_{[1-\kappa, 1+\kappa]}$, where $\kappa \in \{0.05, 0.10, 0.20\}$, such as $\tilde{x}_k^i = x_k^i \times \varepsilon_k^i$; \tilde{x}_k^i denoting the quantity data with errors. That is, 3 amounts of measurement errors, 5%, 10% and 20% are considered throughout the experiments. In order to keep expenditure and prices constant in the sample, we normalize upon one good and set $\tilde{x}_l^i = \left(m^i - \sum_{k=1, l \neq k}^5 p_k^i \tilde{x}_k^i \right) / p_l^i$ for some l .¹⁰ This gives us a total of 12 different experiments in the second design, since we consider 2 different sets of preference parameters, each with 2 sets of prices and 3 sets of measurement errors.

All simulations designs and replications are calibrated with a significance level set to 5%. The number of replications throughout all simulations are set to 1,000.

5.2 Simulation results

The results from simulation design 1 are reported in Table 1. As can be seen from the table, our test has very good size and correctly detects random consumer behavior in 99.1% of all replications when prices are generated from $U_{[90,100]}$ and 98.9% when prices are generated from $U_{[95,100]}$. In the very few cases (0.9% for $p \sim U_{[90,100]}$ and 1.1% for $p \sim U_{[95,100]}$), the test incorrectly rejects random behavior in favor of rational behavior. It is worth noting that the predetermined significance level of 5% is clearly larger than the calculated size, which is a very encouraging finding. Another important finding is that our results seem to be robust to varying degree of relative prices. In fact, unreported findings from additional simulations with larger varying relative prices suggests similar results to the ones obtained in Table 1.

⁹The relative prices vary between 10.8% and 23.5% under these prices distributions.

¹⁰We normalize upon the last good in the simulations. Results were not sensitive to the choice of good.

The results from simulation design 2 is reported in Table 2. The numbers in the table indicates how often the two (incorrect) hypotheses of random and perverse behavior is rejected in favor of rational behavior. These results shows that our test has very good power and correctly accepts rational behavior, when the data is generated from that hypothesis. The findings are very encouraging, since our test finds rationality not only in a majority of all cases, but in all replications for the total 12 sets of experiments considered. It should be noted that this holds even for cases in which the data was shocked with 20% measurement errors.

We also performed simulations under a wide variety of values for the preference parameters β . The results from these experiments are omitted in the tables, since they were much like the ones reported.¹¹ This suggests our test to be robust to different Cobb-Douglas specifications. In addition, we performed simulations under various significance levels ranging from 1% – 5%. As expected, this led to size improvements at lower significance levels. However, we generally obtained power distortions for significance levels below 5%; in other words that we rejected rational behavior when data was generated under that hypothesis. For example, at the 1% significance level, we found our test to accept the incorrect hypothesis of random behavior in 100% of all replications when data was generated under rationality, with preferences β_B , 5% measurement errors and $p \sim U_{[90,100]}$. Given that the gains in size are negligible in relation to the loss in power, we suggest that the significance level should be set no lower than 5%.

The findings in Tables 1 and 2 might appear surprising to some. The natural explanation for these results is that even if measurement errors cause revealed preference violations in many cases, the actual number of violations is generally low. This was supported in simulations performed by Fleissig and Whitney (2003). In fact, they found that the Cobb-Douglas specifications used above with 5% and 10% measurement errors resulted in 6 or fewer violations of revealed preference in 96.3% of the 2,000 replications they considered. They further found, even when shocking the data with 20% measurement error, at most 10 violations over 80% of the time.¹² By contrast, we would expect random consumer behavior to result in a rather large amount of violations if there are many budget intersections and if relative prices vary enough. Since this generally seem to be the case in our simulations, we should also expect our test to perform well.

6 Applications to experimental data

One important area in which revealed preference tests have been frequently applied is to controlled experiments of rationality.¹³ This section applies our test to data from two recent experimental studies. One study investigates the systematic behavior of altruism and the other the choices of children.

¹¹All simulations results may be obtained upon request.

¹²Fleissig and Whitney (2005) applied their test of the size of revealed preference violations using the same data generating process as we do. As we have pointed out above, however, unlike our test, their test requires that the researcher knows the distribution and variance of the measurement errors. Fleissig and Whitney (2005) found that their upper bound test correctly detected rational behavior in all cases but two when the assumed measurement error distribution and variance agreed with the true underlying distribution and variance. However, they also showed that the performance of their test deteriorates the further away from the truth the assumed variance is. For example, if the researcher assumes 5% measurement errors when the true amount actually is 20%, they found their test to accept (the true underlying) rational behavior in 88.3% – 46.1% of all 10,000 replications depending on the price distribution and preference parameters.

¹³See Harbaugh et al. (2001) for a short survey.

6.1 Are altruistic choices rational?

In our first application, we investigate whether altruistic choices can be ascribed to rational behavior or to some other type of systematic behavior. This application originates in the experiments conducted by Andreoni and Miller (2002). In particular, they test whether a model of convex altruistic preferences can explain the observed data. They let their subjects make a series of choices under varying incomes and costs of giving money to another subject. The subjects were given the choice of passing some tokens to another subject and holding on to the rest. As the subjects have to decide upon two choices in the experiments, it follows that WARP in this case is a necessary and sufficient condition for rationality. Andreoni and Miller’s (2002) main finding from applying static revealed preference tests was that up to 98% of the subjects showed evidence of rational altruistic behavior.

The data can be divided up into two sets of sub-experiments. In the first, 142 subjects face 8 budgets, while the second consists of 34 subjects facing 11 budgets. Because of the limited number of available observations for each subject in the two sets, which may lead to poor test power when applying decision rules based on the asymptotic theory (the standard normal distribution), we employ an additional confidence interval for the Z^{obs} -statistic based on a bootstrapped empirical distribution. This bootstrap algorithm is constructed along the lines of that originally proposed by Andreoni and Miller (2002) and is particularly useful in the presence of panel data. The algorithm consists of making one draw from the set of subjects on each budget to form a synthetic subject. In the second step, calculate the number of WARP violations and the Z^{obs} -statistic for the synthetic subject. Denote the calculated Z^{obs} -statistic by Z_b^* . This is repeated B times, which yields a vector of bootstrapped values $\Psi = [Z_1^*, \dots, Z_B^*]$. The confidence interval is constructed by taking the $\alpha/2$ and $1 - \alpha/2$ percentile of Ψ , where α denotes the significance level. The null hypothesis of random consumer behavior is rejected if the Z^{obs} -statistic lies outside the implied interval between the $\alpha/2$ and $1 - \alpha/2$ percentile of Ψ . The number of bootstrap replications is set to $B = 1,000,000$, and we set the significance level to 5% throughout.

Table 3 reports the results from applying our test to the data in the first sub-experiment. From the total 142 subjects, we found 13 with WARP violations, with the total number of violations ranging from 1 to 3. The expected number of violations under random behavior were found to be 1.78 and the variance to be 4.79. The null hypothesis of random behavior cannot be rejected for any of the 142 subjects when the decision rule based on the standard normal distribution is used. More precisely, the Z^{obs} -statistic for the subjects without any violations were -0.81 and the confidence interval under the null of random behavior for these subjects were calculated to be $[-4.29, 4.29]$. Hence, the Z^{obs} -statistic for subjects with zero violations is highly insignificant.

Let us consider the results from applying the bootstrap algorithm. We found that 76.6% of the 1,000,000 synthetically created subjects violated WARP, with the mean number of violations being 1.81. From this we found the bootstrapped confidence interval to be $[-0.81; 1.93]$. Now, it follows that subjects with 1 – 3 violations have a highly insignificant Z -statistic, and the null hypothesis of random behavior cannot be rejected for these subjects. Interestingly, we also found that the hypothesis of random consumer behavior for subjects without WARP violations cannot be rejected. In other words, the results obtained here suggest that altruistic choices may not be as rational as previously thought. However, one should note that the Z^{obs} -statistic for subjects without violations (-0.81) were found to be on the lower boundary of the bootstrap confidence interval. Nevertheless, our results stands in contrast to Andreoni and Miller (2002) who argued based on their non-stochastic revealed preference results that altruistic choices tend to be rational.

The results from the second sub-experiment are reported in Table 3. Out of the total 34 subjects in

this set of experiments, there were 5 violating WARP, with one violation each. The expected number of violations under random consumer behavior were 3.94 and the variance 16.83. When using the decision rule based on asymptotic theory, we cannot reject the null hypothesis of random consumer behavior for any of the 34 subjects.

We again calculated confidence bounds from the bootstrap distribution. We found that 85.65% of the 1,000,000 synthetically created subjects violated WARP with an average number of violations to be 2.69. Applying the bootstrapped confidence intervals, we found that the null hypothesis of random behavior cannot be rejected for any of the 34 subjects. We should however note, as before, that the Z^{obs} -statistic for the 29 subjects without WARP were found to be on the lower boundary of the confidence interval. Even so, these results stands in contrast to Andreoni and Miller's (2002) since they suggest that we cannot reject the possibility that the altruistic choices were due to subjects making random choices.

6.2 Are children rational?

Harbaugh, Krause and Berry (2001) tested the rationality of children by offering them 11 budgets of chips and juice boxes. This study found that second-graders showed clear evidence of rationality, though also many inconsistencies. Fewer six-grade subjects and college student subjects were found to have violations of revealed preference compared to the second-graders, but that the maximal number of violations for the subjects did not differ much between the three groups.

We apply our procedure to this data to test whether the choices made by the children may be ascribed to rational behavior or to some other type of systematic behavior. Our analysis differs somewhat from Harbaugh et al. (2001) in that they consider discrete choice sets, which requires the notion of revealed preference to be slightly redefined. Instead, we consider continuous choice sets which allows for the usual definition of revealed preference as described in Section 2. The effect from this modification were that the number of violations increased for some of the subjects already found by Harbaugh et al. (2001) to violate WARP. From a total of 384 subjects we found 210 with zero violations of WARP and 174 with at least one violation.¹⁴ The average number of violations were 1.51, ranging from 0 to 25, with the maximal number of violations in the second- and six-grade groups being 12, and the maximal number of violations in the college student group being 25.¹⁵ Since there are only two goods in the experiment, WARP will indeed be a necessary as well as sufficient condition for rationality.

Table 3 reports the results. The expected number of violations under random behavior were found to be 3.94 and the variance 15.11. Consider first the results from the decision rule based on the standard normal distribution. The null hypothesis of random consumer behavior cannot be rejected for subjects with 12 violations of WARP or less. Thus, we find that all second- and six-grade subjects and most of the college students conform to uniform random behavior. One should, however, note from this that we cannot reject random behavior for subjects without violations of revealed preference. The 5 college student subjects with 15 or more violations of WARP were found to reject random behavior in favor of perverse economic behavior.

Consider next the results from applying decision rules based bootstrap critical values. The number of bootstrap replications is set as in the previous application to $B = 1,000,000$. The confidence interval under random behavior were calculated to be $[-0.76; 2.59]$. Since the Z^{obs} -statistic for the 210 subjects

¹⁴It should be noted that Harbaugh et al. (2001) only included 128 children in their experiments. However, they let the same children do the experiments 3 times, and therefore making the subjects turn up on three occasions in the data.

¹⁵Out of the 5 college students with 12 or more violations, there were 3 with 25 violations and 2 with 15 violations. All other college students but these five had a maximum of 5 violations.

without WARP violations were -1.01 , we found these subjects to reject random consumer behavior in favor of rational behavior. The Z^{obs} -statistic for the 54 subjects with one violation were calculated to be -0.76 and thus lying on the boundary of the lower critical value. This result has, however, a simple explanation. Many synthetically created subjects in the bootstrap were found to have one WARP violation, resulting in a probability mass at -0.76 in the empirical distribution. The p-value for these subjects were found to be 0.12 . In other words, in the case of subjects with one violation, we cannot reject random behavior if the significance level is set lower than 12% . The 115 subjects with 2 – 12 violations showed clear evidence of random consumer behavior. Finally, we found that the 5 college students with 15 or more violations conforms to perverse economic behavior. As a final remark, it is interesting to note that while there were (relatively) more second-graders than six-graders and college subjects that rejected rationality, only college student subjects exhibit perverse economic behavior.

7 Summary

Becker (1962) argues that the only ultimate defense in favor of the rationality axioms in economics are based on empirical grounds. To date, this defense has been weaker than desirable due to the non-stochastic nature of revealed preference tests. In this paper we address this problem by suggesting a statistical test of the number of violations of revealed preference.

As a way to interpret the number of violations of revealed preference, we have suggested a statistical test of the type of behavior that is being observed. This test allows the researcher to draw inference as to the type of consumer behavior that is consistent with the data. Observed behavior may be found to be consistent with rational behavior, random behavior or perverse behavior on the part of the consumer. This test requires no additional data than that required for non-stochastic revealed preference tests.

We present results from applying the proposed test to two experimental data sets. Our results suggest some different interpretations of the behavior of the experimental subjects than those implied by non-stochastic revealed preference results. As an example, we find cases where despite zero violations of revealed preference, the null hypothesis of random behavior cannot be rejected.

References

- [1] Afriat, S.N. (1967). The construction of utility functions from expenditure data. *International Economic Review* **8**, pp.67-77.
- [2] Aizcorbe, A. (1991). A lower bound for the power of nonparametric tests. *Journal of Business and Economic Statistics* **9**, pp.463-467
- [3] Andreoni, J. and W.T. Harbaugh (2006). Power indices for revealed preference tests. University of Wisconsin-Madison, Department of Economics Working paper 2005-10.
- [4] Andreoni, J. and J. Miller (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* **70**, pp.737-753.
- [5] Becker, G.S. (1962). Irrational behavior and economic theory. *Journal of Political Economy* **70**, pp.1-13.
- [6] Blundell, R.W., M. Browning and I.A. Crawford (2004). Nonparametric Engel curves and revealed preference. *Econometrica* **71**, pp.205-240.
- [7] Bronars, S.G. (1987). The power of nonparametric tests of preference maximization. *Econometrica* **55**, pp.693-698.
- [8] de Perreti, P. (2004). Testing the significance of the departures from utility maximization. *Macroeconomic Dynamics* **9**, pp.372-397.
- [9] Epstein, L.G. and A.J. Yatchew (1985). Non-parametric hypothesis testing procedures and applications to demand analysis. *Journal of Econometrics* **30**, pp.149-169.
- [10] Fleissig, A.R. and G.A. Whitney (2003). A new PC-based test for Varian's weak separability inequalities. *Journal of Business and Economic Statistics* **21**, pp.133-144.
- [11] Fleissig, A.R. and G.A. Whitney (2005). Testing for the significance of violations of Afriat's inequalities. *Journal of Business and Economic Statistics* **23**, pp.355-362.
- [12] Gross, J. (1995). Testing data for consistency with revealed preference. *Review of Economics and Statistics* **78**, pp.701-710.
- [13] Harbaugh, W.T., K. Krause and T.R. Berry (2001). GARP for kids: On the development of rational choice behavior. *The American Economic Review* **91**, pp.1539-1545.
- [14] Houthakker, H.S. (1950). Revealed preference and the utility function. *Economica* **17**, pp.159-174.
- [15] Leroux, A. (2000). Cutting a simplex with a hyperplane: Question of volume. Université Aix-Marseille III, GREQAM Working paper Number 00A13.
- [16] McCausland, W.J. (2009). Random consumer demand. *Economica* **76**, pp.89-107.
- [17] Samuelson, P.A. (1938). A note on the pure theory of consumer's behaviour. *Economica* **5**, pp.61-71.
- [18] Swofford, J.L. and G.A. Whitney (1986). Flexible functional forms and the utility approach to the demand for money: A nonparametric analysis. *Journal of Money, Credit and Banking* **18**, pp.383-389

- [19] Varian, H.R. (1982). The nonparametric approach to demand analysis. *Econometrica* **50**, pp.945-973.
- [20] Varian, H.R. (1985). Nonparametric analysis of optimizing behavior with measurement error. *Journal of Econometrics* **30**, pp.445-458.
- [21] Varian, H.R. (2006). Revealed preference. In M. Szenburg, L. Ramrattan and A.A. Gottesman (Eds.) *Samuelsonian economics and the twenty-first century*. Oxford University Press.
- [22] Ziegler, G. (1998). *Lectures on Polytopes*. Springer Verlag.

Table 1: Simulation results from design 1.

Type of behavior	Random	Rational	Perverse
$p \sim U_{[90,100]}$	0.991	0.009	0.000
$p \sim U_{[95,100]}$	0.989	0.011	0.000

Notes: The numbers in the table refers to the proportion of replications for which the hypothesis of random behavior, rational behavior and perverse economic behavior is accepted. The true data generating process follows the null hypothesis of random consumer behavior. The number of replications is 1,000.

Table 2: Simulation results from design 2.

Preference Type	α_A			α_B		
	5%	10%	20%	5%	10%	20%
$p \sim U_{[90,100]}$	1.000	1.000	1.000	1.000	1.000	1.000
$p \sim U_{[95,100]}$	1.000	1.000	1.000	1.000	1.000	1.000

Notes: The numbers in the table refers to the proportion of replications for which the alternative hypothesis of rational behavior is accepted. The true data generating process follows the alternative hypothesis of rational consumer behavior. The number of replications is 1,000.

Table 3: Results from experimental data.

Hypothesis (Type of behavior)	H_0^{SN}	H_1^{SN}	H_2^{SN}	H_0^B	H_1^B	H_2^B
Harbaugh et al. (384 subjects)						
Observations with violations (174)	169	0	5	169	0	5
Observations without violations (210)	210	0	0	0	210	0
Andreoni and Miller (142 subjects)						
Observations with violations (13)	13	0	0	13	0	0
Observations without violations (129)	129	0	0	129	0	0
Andreoni and Miller (34 subjects)						
Observations with violations (5)	5	0	0	5	0	0
Observations without violations (29)	29	0	0	29	0	0

Notes: H_0^{SN}, H_1^{SN} and H_2^{SN} indicates a decision rule based on the standard normal distribution and H_0^B, H_1^B and H_2^B indicates a decision rule based on the bootstrapped distribution. H_0 refers to random behavior, H_1 is rational behavior and H_2 is perverse economic behavior. The significance level is set to 5%.

Figure 1: Pairwise comparison of intersecting budgets, $K = 2$.

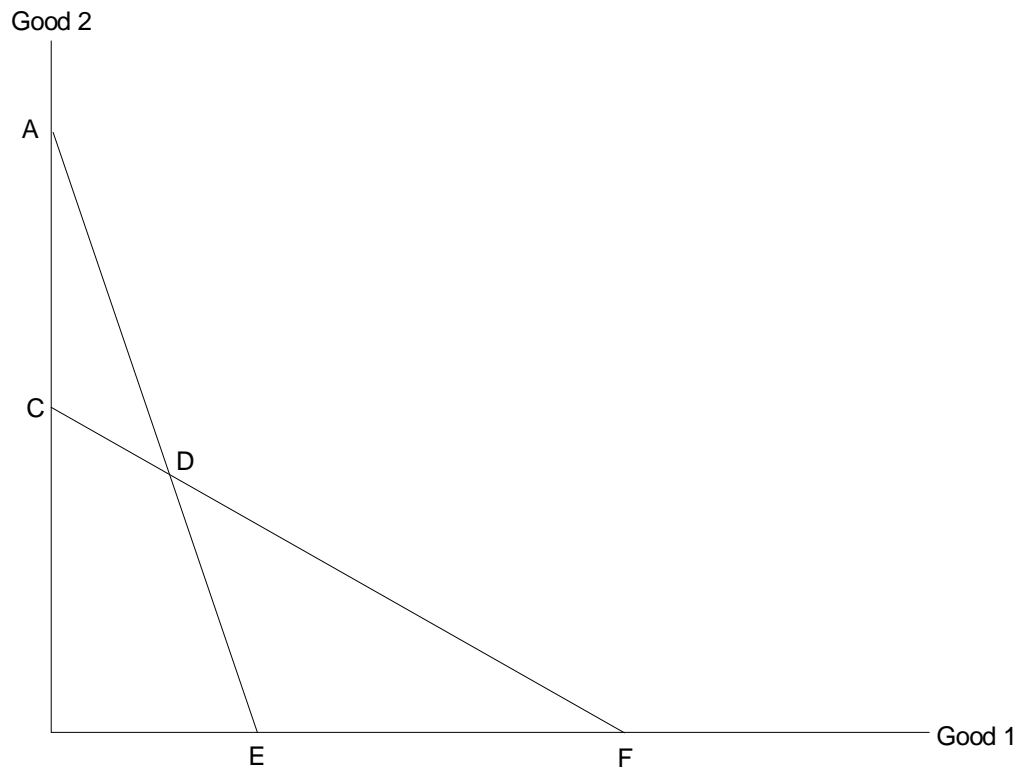


Figure 2: Pairwise comparison of intersecting budgets conditional on the outcome in Figure 1, $K = 2$.

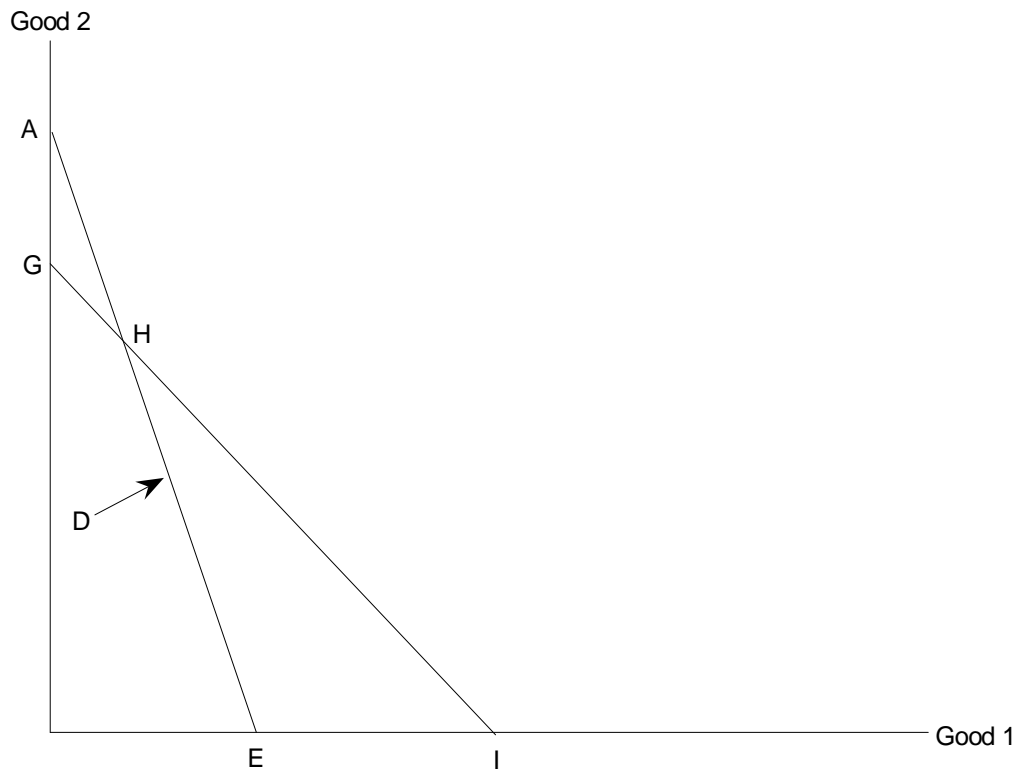


Figure 3: Pairwise comparison of intersecting budgets conditional on the outcome in Figure 1, $K = 2$.

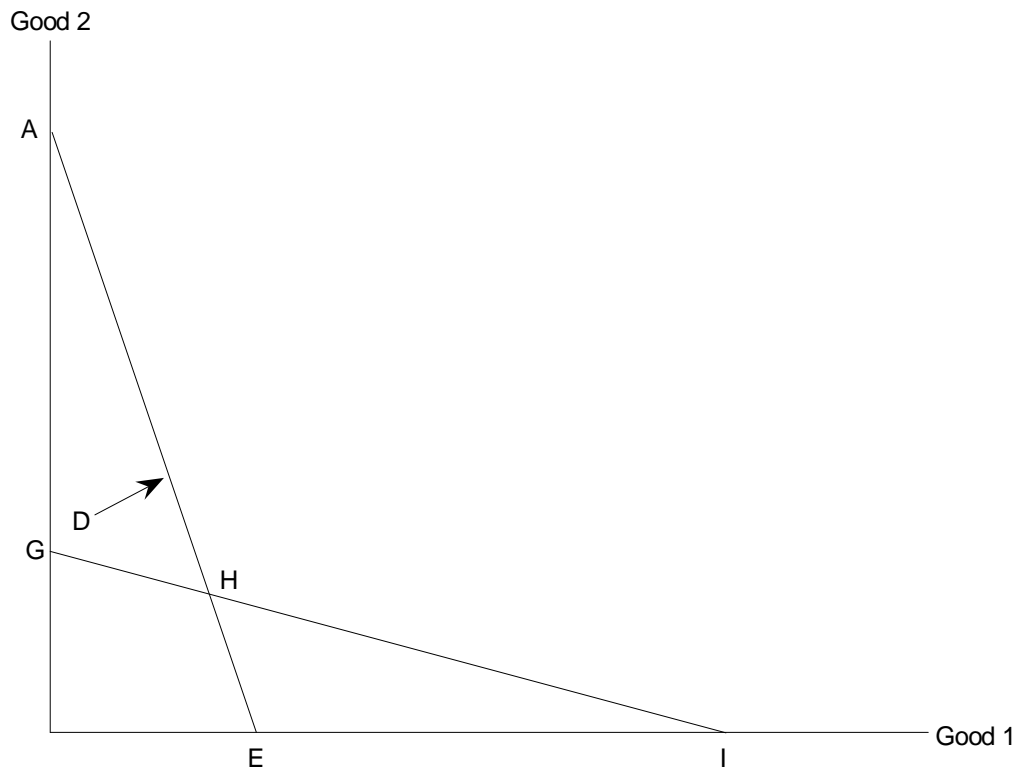


Figure 4: Pairwise intersecting budget sets, $K = 3$.

