Macroeconomic Dynamics, **9**, 2005, 612–629. Printed in the United States of America.
DOI: 10.1017.S1365100505040381

A COMPARISON OF TWO METHODS FOR TESTING THE UTILITY MAXIMIZATION HYPOTHESIS WHEN QUANTITY DATA ARE MEASURED WITH ERROR

BARRY E. JONES
Binghamton University

PHILIPPE DE PERETTI
Université Paris 1 Panthéon-Sorbonne

The Generalized Axiom of Revealed Preference (GARP) can be violated because of random measurement errors in the observed quantity data. We study two tests proposed by Varian (1985) and de Peretti (2004), which test GARP within an explicit stochastic framework. Both tests compute adjusted quantity data that are compliant with GARP. We compare and contrast the two tests in theoretical terms and in an empirical application. The empirical application is based on testing a large group of monetary assets for the United States over multiple sample periods spanning 1960–1992. We found that both tests provided reasonable results and were largely consistent with each other.

Keywords: GARP, Weak Separability, Nonparametric Tests, Measurement Errors

1. INTRODUCTION

Varian (1982, 1983) described nonparametric methods that can be used to test data for consistency with optimizing models. The most widely applied test is the test of the Generalized Axiom of Revealed Preference (GARP). It has been used to test for utility maximization, as well as to test for weak separability. A defect of all such tests is that they do not incorporate measurement error into the analysis (Varian, 1985). Simulation studies have shown that the problem can be significant. Fleissig and Whitney (2003) showed that data generated from a single utility maximizing agent violated GARP in up to 20% of their trials when the observed quantity data contained uniformly distributed measurement errors in the range of $\pm 5\%$.[1] de Peretti (2004) obtained very similar results. The most commonly used nonparametric weak separability test is based on a sequence of GARP tests (see Varian, 1983).[2] Consequently, nonparametric weak separability tests may be even more adversely affected by measurement error.[3]

Address correspondence to: Barry E. Jones, Department of Economics, P.O. Box 6000, Binghamton University, Binghamton, NY 13902-6000, USA; e-mail: bjones@binghamton.edu.

© 2005 Cambridge University Press 1365-1005/05 $12.00

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

Varian (1985) introduced a test of optimizing behavior within an explicit stochastic framework that can account for measurement error. His empirical application was based on testing the Weak Axiom of Cost Minimization, but he described how to test GARP. Jones, Dutkowsky, and Elger (2005) applied the methodology to test GARP in an empirical study. De Peretti (2004) derived an alternative test of optimizing behavior that is specific to GARP. For GARP, the two test procedures follow the same logical structure and test the same null hypothesis, albeit in different ways. The null hypothesis of the tests is that the true data satisfy GARP, but the observed data may violate GARP as the result of *i.i.d.* measurement errors in the quantity data. The tests both follow a two-step logical structure if the observed data violate GARP. First, they use a numerical procedure to adjust the quantity data in order to satisfy utility maximization (referred to, hereafter, as the *adjustment procedure*). Second, they test the computed adjustments for statistical significance (referred to, hereafter, as the *test procedure*).[4]

In spite of the similarity in structure, the two tests differ on several key points. First, the adjustment procedures used in the tests are different. In Varian (1985), the adjustment procedure is based on the Afriat inequalities and it adjusts all of the bundles simultaneously. In de Peretti (2004), the bundles are adjusted iteratively in order to satisfy GARP, and not all bundles are adjusted. In addition, expenditure on the adjusted bundles is constrained to equal observed expenditure for all observations in de Peretti (2004) but not in Varian (1985). Second, the test procedures are also different. In Varian (1985), the minimized sum of squared errors computed by the adjustment procedure are less than the true sum of squared measurement errors under the null. If the true measurement errors are assumed to be normally distributed, this result can be used to produce a chi-square test statistic.[5] The test requires, however, that the tester knows the standard deviation of measurement errors. If it is unknown, the minimized sum of squared errors can, nevertheless, be used to calculate a bound on the unknown standard deviation such that if the standard deviation is greater than or equal to this bound then the null hypothesis cannot be rejected, see Varian (1985, p. 450). In de Peretti (2004), it is assumed that, under the null, the errors computed by the adjustment procedure inherit the *i.i.d.* property of the true measurement errors. GARP is rejected if the computed errors are not statistically independent or identically distributed.

In this paper, we describe the two tests in theoretical terms, highlighting the similarities and differences between them, and then illustrate the two tests in an empirical application. Aggregate monetary asset and user cost data are used in the application, because they often fail to satisfy GARP in empirical studies in contrast to aggregate consumption data. We tested monetary asset stocks and their associated user cost prices for consistency with utility maximization (GARP) using data for the United States from Thornton and Yue (1992). We found GARP violations in seven sample periods, and we implemented the de Peretti and Varian tests in those cases. The sample size and number of assets in each of our sample periods is

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

similar to those in many empirical studies that test monetary asset groupings for weak separability.

Fleissig and Whitney (2003, p. 137) found that measurement errors tended to produce relatively few GARP violations in their simulation study. Consequently, we might expect these two tests to reject the null hypothesis of utility maximization in samples that have large numbers of GARP violations. We were able to directly compare the two tests in six of the seven sample periods. The number of GARP violations were low in five of these six samples and very high in the other sample.[6]

We found that the de Peretti and Varian tests both provided reasonable results and were largely consistent with each other in our application. The tests both supported the null hypothesis of utility maximization in the majority of the five samples that had low numbers of GARP violations. The de Peretti test rejected the null hypothesis in only one of these five samples. We implemented the Varian test using a range of values for the standard deviation of measurement error, because the true value is unknown. We found that we would not be able to reject the null hypothesis in three of the five samples unless we used extremely low values for the standard deviation of measurement errors. In the other two samples, the results were only slightly less supportive.

The de Peretti test strongly rejected the null in the sample that had a large number of GARP violations, which conformed to our expectations. The results for the Varian test indicated that the standard deviation would have to be much higher for this sample than it would have to be for any of the others in order to be unable to reject the null. Thus, the Varian test is much less supportive of the null for this sample than for any of the other samples, which again conformed to our expectations.

The remainder of the article is organized as follows. In Section 2, we describe the two test procedures. In Section 3, we present our empirical application. In Section 4, we conclude.

2. TEST PROCEDURES

We begin by providing some basic notation and definitions. Let $\mathbf{x}_i^* = (x_{i,1}^*, \ldots, x_{i,K}^*)$ denote the true unobserved $(K \times 1)$ vector of quantities of goods consumed in period $i \in \{1, \ldots, I\}$. The corresponding $(K \times 1)$ price vector is denoted by $\mathbf{p}_i = (p_{i,1}, \ldots, p_{i,K})$. Let $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,K})$ denote the observed vector of quantities in period $i \in \{1, \ldots, I\}$, which possibly contains measurement error.

GARP is defined using standard revealed preference relations. Let P^0 denote *strictly directly revealed preferred to*, R^0 denote *directly revealed preferred to*, and R denote *revealed preferred to*, see Varian (1982) or de Peretti (2004). The observed dataset violates GARP if $\mathbf{x}_i R \mathbf{x}_j$ and $\mathbf{x}_j P^0 \mathbf{x}_i$ for any pair of observations $i, j \in \{1, \ldots, I\}$. A dataset is consistent with utility maximization if and only if it satisfies GARP; see Varian (1982).

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

2.1. Null Hypothesis

The null hypothesis for both the Varian (1985) and de Peretti (2004) tests is that the true unobserved dataset $\{\mathbf{x}_i^*, \mathbf{p}_i : i = 1, \ldots, I\}$ is consistent with GARP. The observed dataset $\{\mathbf{x}_i, \mathbf{p}_i : i = 1, \ldots, I\}$ may or may not satisfy GARP. $\mathbf{x}_i$ and $\mathbf{x}_i^*$ are assumed to be related to each other through a random multiplicative term representing measurement error:

$$x_{i,k}^* = x_{i,k}(1 + \varepsilon_{i,k}), \tag{1}$$

where the $\varepsilon_{i,k}$ are assumed to be zero-mean *i.i.d.* random variables with variance σ^2.

If the observed dataset violates GARP, then the adjustment procedures each calculate a perturbation of the observed quantity data that is consistent with GARP. We use the notation $\widehat{\boldsymbol{\zeta}}_i = (\widehat{\zeta}_{i,1}, \ldots, \widehat{\zeta}_{i,K})$ to denote the $(K \times 1)$ vector of adjusted quantities in period $i \in \{1, \ldots, I\}$ computed by either procedure, which satisfy GARP. Let $\widehat{\varepsilon}_{i,k}$ denote the *computed error* of good k in period i defined by (2).

$$\widehat{\varepsilon}_{i,k} = \frac{\widehat{\zeta}_{i,k}}{x_{i,k}} - 1. \tag{2}$$

2.2. The Varian (1985) Test

In Varian (1985), the adjustment procedure is based on minimizing objective function (3) subject to (4), in $\boldsymbol{\zeta}_i = (\zeta_{i,1}, \ldots, \zeta_{i,K})$, V_i, and μ_i for $i \in \{1, \ldots, I\}$.

$$F(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_I) = \sum_{i=1}^{I} \sum_{k=1}^{K} \left(\frac{\zeta_{i,k}}{x_{i,k}} - 1 \right)^2 \tag{3}$$

$$V_i \leq V_j + \mu_j(\mathbf{p}_j \cdot \boldsymbol{\zeta}_i - \mathbf{p}_j \cdot \boldsymbol{\zeta}_j) \ \forall i, j \in \{1, \ldots, I\}. \tag{4}$$

The *Afriat indexes*, V_i and μ_i, are constrained to be strictly positive for all $i \in \{1, \ldots, I\}$. The solution for the $\boldsymbol{\zeta}_i$ arguments is called the *minimal perturbation*, which we denote by $\widehat{\boldsymbol{\zeta}}_i$ for $i \in \{1, \ldots, I\}$. Varian (1982) proved that a dataset satisfies GARP if and only if there exist indexes satisfying the *Afriat inequalities* (4). Thus, the adjusted dataset $\{\widehat{\boldsymbol{\zeta}}_i, \mathbf{p}_i : i = 1, \ldots, I\}$ satisfies GARP, by construction. Under the null hypothesis, the true unobserved dataset $\{\mathbf{x}_i^*, \mathbf{p}_i : i = 1, \ldots, I\}$ also satisfies GARP.

Therefore, under the null, the following property must hold by definition:

$$\sum_{i=1}^{I} \sum_{k=1}^{K} (\widehat{\varepsilon}_{i,k})^2 = F(\hat{\boldsymbol{\zeta}}_1, \ldots, \hat{\boldsymbol{\zeta}}_I) \leq F(\mathbf{x}_1^*, \ldots, \mathbf{x}_I^*) = \sum_{i=1}^{I} \sum_{k=1}^{K} (\varepsilon_{i,k})^2. \tag{5}$$

If we assume that the true measurement errors, $\varepsilon_{i,k}$, are *normally distributed i.i.d.* random variables with zero mean and variance σ^2, then $\frac{F(\mathbf{x}_1^*, \ldots, \mathbf{x}_I^*)}{\sigma^2}$ is distributed as

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

a χ^2_{IK}. Let C_α be the critical value for a χ^2_{IK} at the α significance level. A standard hypothesis test would be to reject the null if $\frac{F(\mathbf{x}_1^*,\ldots,\mathbf{x}_I^*)}{\sigma^2} > C_\alpha$. The test statistic cannot be computed, however, because the true data are unobserved. To resolve this problem, Varian (1985) suggested rejecting the null hypothesis if

$$T \equiv \frac{F(\hat{\zeta}_1, \ldots, \hat{\zeta}_I)}{\sigma^2} > C_\alpha. \tag{6}$$

In applications, the test could be run using an estimate of the true standard deviation; see Varian (1985, p. 449).[7] Alternatively, the test could be run over a range of values to determine the effect the standard deviation on the results. In that regard, (6) can be used to calculate a bound on the standard deviation of measurement error, such that if σ is greater than or equal to this bound, then we cannot reject the null hypothesis. The results of this approach are inherently more subjective than that of a statistical test, but it can, nevertheless, be very informative. For example, Varian (1985, p. 450) argues that if the bound is "... much smaller than our prior opinions concerning the precision with which these data have been measured, we may well want to accept the maximization hypothesis." Varian (1985, pp. 452–455) interpreted his empirical results for the Weak Axiom of Cost Minimization in these terms (see also Jones, Dutkowsky, and Elger, 2005).

2.3. The de Peretti (2004) Test

In de Peretti (2004), the adjustment procedure is based on minimizing the objective function (3) iteratively using the information contained in the transitive closure matrix. A detailed description of the iterative adjustment procedure is provided in de Peretti (2004). We will focus on highlighting the differences between the de Peretti (2004) and Varian (1985) adjustment procedures. The main differences are that (i) the de Peretti adjustment procedure is iterative, and (ii) the constraints are different.

The de Peretti (2004) adjustment procedure builds on a method for adjusting a particular bundle, $\mathbf{x}_i$, which is involved in a GARP violation. The method is to minimize (7) subject to (8), in $\boldsymbol{\zeta}_i$.

$$\sum_{k=1}^{K} \left(\frac{\zeta_{ik}}{x_{ik}} - 1 \right)^2 \tag{7}$$

$$\boldsymbol{\zeta}_i R \mathbf{x}_j \text{ implies } \textit{not } \mathbf{x}_j P^0 \boldsymbol{\zeta}_i, \forall j \in \{1, \ldots, I\}. \tag{8}$$

The constraint (8) can be implemented through (9) and (10):

$$\boldsymbol{\zeta}_i \cdot \mathbf{p}_i = \mathbf{x}_i \cdot \mathbf{p}_i \tag{9}$$

$$\boldsymbol{\zeta}_i \cdot \mathbf{p}_j \geqslant \mathbf{x}_j \cdot \mathbf{p}_j, \text{ for all } j \text{ such that } \mathbf{x}_i R \mathbf{x}_j. \tag{10}$$

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

The adjustment procedure is an iterative algorithm. The iterative algorithm starts at the highest rupture of the preference chain and works downward, adjusting one bundle at each iteration until all violations have been eliminated. The bundle to be adjusted at each iteration is chosen from amongst the subset of bundles, B, involved in the highest remaining rupture of the preference chain. Specifically, (7) is minimized subject to (9) and (10) for all $i \in B$. The bundle that is adjusted is the one, $\hat{\imath}$, with the smallest minimized value of (7) and the corresponding adjusted quantity, $\hat{\boldsymbol{\zeta}}_{\hat{\imath}}$, is set equal to the optimal solution. The expenditure constraint (9) insures that the adjustment does not produce additional ruptures at higher levels of the preference chain. Each bundle is adjusted at most once. $\hat{\boldsymbol{\zeta}}_i$ is set to equal $\mathbf{x}_i$ for any bundles that were not otherwise adjusted by the procedure. The adjusted dataset satisfies GARP by construction.

The test procedure consists of testing the computed errors for independence and identical distribution. In this paper, we test the time series, s_t, consisting of the pooled nontrivial computed errors, see de Peretti (2004) for details. We estimate two auxiliary regressions:

$$s_t = c_1 + \alpha \cdot trend + \sum_{j=1}^{\tau 1} \gamma_j s_{t-j} \tag{11}$$

$$\mathbf{s}_t^2 = c_2 + \delta \cdot trend + \sum_{j=1}^{\tau 2} \sum_{k=1}^{\tau 2} d\beta_{jk} s_{t-j} s_{t-k}, \tag{12}$$

where

$$d = \begin{cases} 1 : \text{if } k \geqslant j, \\ 0 : \text{otherwise}, \end{cases}$$

and test the estimated coefficients of the regressions for joint significance, see Spanos (1999).[8] We refer to (11) as first-order dependence and trend heterogeneity and (12) as second-order dependence and trend heterogeneity.

2.4. Discussion

To avoid potential misinterpretation, we begin our discussion by noting that we do not interpret either adjustment procedure as estimating the true data or the true measurement errors.[9] The purpose of the Varian (1985) adjustment procedure is to produce computed errors that are no larger than the true errors, in the least squares sense, under the null. The purpose of the de Peretti (2004) adjustment procedure is to produce *i.i.d.* computed errors, under the null. It is not possible to compute adjusted data so as to provide as good of a fit as possible to the true data, because the true data are (by definition) unobserved.

We can now proceed to compare and contrast some particular aspects of the two adjustment procedures. First, the Varian (1985) adjustment procedure should typically result in smaller adjustments, in the least squares sense, than the de

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

Peretti (2004) adjustment procedure. The main reasons for this are that the de Peretti (2004) procedure is iterative and it adjusts the quantities within a particular bundle holding total expenditure fixed. As a result, it eliminates GARP violations by adjusting bundles along the observed budget lines. The quantities of some goods within an adjusted bundle will be adjusted upward and the quantities of others will necessarily be adjusted downward to offset the effect on total expenditure. In Varian (1985), the minimal perturbation is not constrained in terms of expenditure and all bundles are adjusted simultaneously. Therefore, it eliminates GARP violations, in part, by adjusting total expenditure. In this way, expenditure acts as a buffer variable, which allows for smaller adjustments relative to de Peretti's procedure.

The preceding argument is subject to an important caveat, however. The minimal perturbation cannot be computed exactly but must instead be approximated numerically using a nonlinear programming solver. The complexity of the nonlinear programming problem leads to the usual issue of convergence to a local minimum. If the numerical procedure used to approximate the minimal perturbation were to converge to a local minimum, but not to the global minimum, then the de Peretti (2004) procedure could produce a lower total sum of squared adjustments.

Second, the Varian (1985) adjustment procedure is much more time consuming to run than the de Peretti (2004) adjustment procedure. In Varian (1985), the adjustments are computed simultaneously subject to a system of nonlinear inequalities (4). The number of nonlinear inequality constraints is $I(I-1)$, which implies that the computational burden increases substantially as the sample size, I, increases. In addition, the number of adjusted quantities is IK, which implies that the computational burden of the procedure also increases with the number of goods, K. In contrast, the de Peretti (2004) procedure iteratively adjusts the data. The procedure involves repeated minimizations of (7) subject to (9) and (10) for different bundles. The burden of these minimizations does not increase significantly with either sample size or number of goods. Thus, in practice, the de Peretti (2004) procedure runs very quickly, whereas the Varian (1985) procedure can be very time consuming to run (especially, on a large sample).

We now turn to the test procedures. The two tests convey different information because of the different approaches used to test the null hypothesis. Varian (1985) suggests testing if the computed errors (needed to render the data consistent with GARP) are large, in the least-squares sense, given the standard deviation of measurement errors and assuming normality. De Peretti (2004) suggests testing if the computed errors are *i.i.d.*, assuming that they inherit the *i.i.d.* property from the true errors, under the null. In effect, Varian suggests looking at the magnitude of the computed errors, whereas de Peretti looks at their distribution. Large computed errors can be *i.i.d.* and, likewise, small computed errors may not be independent or identically distributed. Thus, the two tests provide different information and are, therefore, likely to be complementary in an empirical study. We now proceed to examine them in an empirical application.

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

TABLE 1. Description of monetary assets[1]

CUR	Sum of currency and traveler's checks
HDD, BDD	Demand deposits (for households and businesses)
OCD	Other checkable deposits
SNOWC, SNOWT[2]	Super NOW accounts (at commercial banks and thrifts)
ONRP	Overnight repurchase agreements
ONED	Overnight eurodollars
MMMF	Money market mutual funds
MMDAC, MMDAT[3]	Money market deposit accounts (at commercial banks and thrifts)
SDCB, SDSL	Savings deposits (at commercial banks and thrifts)
STDCB, STDTH	Small time deposits (at commercial banks and thrifts)
LTDCB, LTDTH	Large time deposits (at commercial banks and thrifts)
MMMFI	Institution only money market mutual funds
TRP	Term repurchase agreements
TED	Term eurodollars
SB	Savings bonds
STTS	Short-term Treasury securities
BA	Bankers acceptances
CP	Commercial paper

[1]See Thornton and Yue (1992, pp. 46–52) for complete description.
[2]SNOWC and SNOWT begin in 1983:1. After 1986:3, there is no distinction with NOWs.
[3]MMDAC and MMDAT are included in SDCB and SDSL after 1991:8.

3. EMPIRICAL APPLICATION

The application is based on testing the underlying data used by Thornton and Yue (1992) to calculate a set of Divisia monetary aggregates for the United States from 1960 to 1992. The data consists of monthly observations of nominal asset stocks and real user costs for the assets in the monetary aggregate L (*Liquid Assets*). See Barnett (1978, 1980) for derivations of the user cost formula. We converted these data to real per-capita stocks and nominal user costs using the CPI and a measure of population. For reference, the set of monetary assets is detailed in Table 1.

We cannot test the dataset for GARP over the entire sample, 1960–1992, because of inconsistencies in the data. This is mainly because of the introduction of new monetary assets, such as Super NOW accounts, money market deposit accounts, and money market mutual funds.[10] There are eight sample periods, denoted by $S1$–$S8$, spanning almost all of 1960–1992 that can be tested. See Table 2 for details.

Fisher and Fleissig (1997, pp. 461–464) divided their dataset (also from Thornton and Yue, 1992) into 21 subsamples of 1960:1–1993:5 and they tested groups of monetary assets (corresponding to M1A, M1, M2, M3, and L) for weak separability over each of them. Weak separability is the key property required for the existence of an economic monetary aggregate, see Barnett (1982), Swofford and Whitney (1994), and Barnett and Serletis (2000). A necessary condition for a group of assets to be weakly separable is that they satisfy GARP. In each of

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

TABLE 2. Sample periods

	Dates	# of Assets (K)	# of Obs. (I)
S1	1960:1–1962:12	13	36
S2	1963:1–1969:9	14	81
S3	1970:2–1973:10	17	45
S4	1974:4–1977:1	19	34
S5	1977:2–1982:11	20	70
S6	1983:1–1986:3	24	39
S7	1986:4–1991:8	22	65
S8	1991:9–1992:12	20	16

these 21 subsamples, they found that "... there were no violations of GARP for the entire set of data provided by the Federal Reserve."

3.1. Nonstochastic Tests

We tested the L monetary assets for GARP in our eight sample periods. We report the nonstochastic GARP test results in Table 3. The L assets violate GARP in all of our sample periods, except for $S8$. We report the number of GARP violations (*nvio*), the number of nontrivial binary comparisons $I(I\text{-}1)$, the violation rate $nvio/I(I\text{-}1)$, and inefficiency indexes for $S1$–$S7$.[11] The inefficiency index measures the severity of the violations, with low values indicating less severe violations. The results indicate that the number, rate, and severity of the GARP violations are all much higher for $S2$ and $S5$ than for any other sample. $S1$ has the least severe and least numerous violations.

$S2$ and $S5$ have extremely high numbers of GARP violations, 1479 and 442 respectively, with corresponding violation rates of 23% and 9%. We also computed the number of GARP violations recursively from observation 1 to i, for $i = 2, \ldots, I$. For $S2$ ($I = 81$), we found no violations up to $i = 45$, but rapid increases in violations thereafter. For $S5$ ($I = 70$), we found no violations up to $i = 41$, and only a moderate number of violations up until $i = 64$. The high number

TABLE 3. Nonstochastic GARP test by sample period

	S1	S2	S3	S4	S5	S6	S7	S8
Test Result	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Pass
Number of violations (*Nvio*)	2	1479	6	8	442	6	18	0
Number of comparisons	1260	6480	1980	1122	4830	1482	4160	240
Violation rate (%)	0.16	22.82	0.30	0.71	9.15	0.40	0.43	0
Inefficiency index (%)	0.01	2.60	0.12	0.10	1.75	0.03	0.16	0

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

of violations suggests that the rejection of GARP is probably not attributable to measurement errors in either $S2$ or $S5$, see Fleissig and Whitney (2003, p. 137).

We ran the two tests outlined in Section 2 over the L assets for $S1$–$S7$. We will refer to $S1$, $S3$, $S4$, $S6$, and $S7$ collectively as Group 1, and we will refer to $S2$ and $S5$ as Group 2. In the next two sections, we present results for the two tests. The two tests were run separately, one by each author. de Peretti ran his test and Jones ran the Varian test.[12]

3.2. The Varian (1985) Test

The minimal perturbation was calculated using the computer code from Jones, Dutkowsky, and Elger (2005), which makes use of the commercial solver FFSQP.[13] We were unable to compute the minimal perturbation for $S2$, but we were successful in all other samples.[14] The adjustment procedure was very time consuming to run for the two longest samples, $S5$ and $S7$.

In Table 4, we report results for the minimal perturbation. We report the sum of squared adjustments (SSR) given by (4) multiplied by 100^2, $RMSE = \sqrt{SSR/(KI)}$ multiplied by 100, and descriptive statistics for the computed errors multiplied by 100. The descriptive statistics include the mean, standard deviation, and the maximal absolute value of the computed errors (max).[15] $S2$ is omitted from the table, because of lack of results.

In Group 1, the computed errors are small. For $S1$, they are bounded by $\pm 0.11\%$. In $S4$ and $S6$, they are bounded by $\pm 0.48\%$ and $\pm 0.43\%$, respectively. The computed errors are slightly larger in $S3$ and $S7$, but $S7$ is the only sample in Group 1 with any computed errors over 1% in absolute value. In contrast, the computed errors are much larger for $S5$, which had some errors that were over 4% in absolute value.

We now turn to the test procedure. The standard deviation of measurement errors, σ, is required in order to implement the test, see (6). Because the true standard deviation is unknown, we ran the test using a range of different values for σ to show how it affects the results. In particular, we ran the test using $\sigma = 0.00025$ (0.025%), 0.0005 (0.05%), 0.00075 (0.075%), and 0.001–0.005 (0.1–0.5%). We present the results of the test in Table 5. We provide the value of the test statistic,

TABLE 4. Minimal perturbations by sample period

	S1	S3	S4	S5	S6	S7
$SSR * 100^2$	0.03706	6.34461	1.45911	301.48735	1.66616	5.49292
$RMSE * 100$	0.00890	0.09107	0.04753	0.46406	0.04219	0.06198
Mean $* 100$	0.00009	−0.00067	0.00014	0.00070	0.00012	0.00072
Std. dev. $* 100$	0.00891	0.09113	0.04756	0.46422	0.04221	0.06199
$max * 100$	0.10267	0.80984	0.47948	4.14875	0.42752	1.05478

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

TABLE 5. Varian (1985) test procedure by sample period

	S1	S3	S4	S5	S6	S7
T for $\sigma = 0.00025$	59.3	10151.4*	2334.6*	482379.8*	2665.9*	8788.7*
T for $\sigma = 0.00050$	14.8	2537.8*	583.6	120594.9*	666.5	2197.2*
T for $\sigma = 0.00075$	6.6	1127.9*	259.4	53597.8*	296.2	976.5
T for $\sigma = 0.001$	3.7	634.5	145.9	30148.7*	166.6	549.3
T for $\sigma = 0.002$	0.9	158.6	36.5	7537.2*	41.7	137.3
T for $\sigma = 0.003$	0.4	70.5	16.2	3349.9*	18.5	61.0
T for $\sigma = 0.004$	0.2	39.7	9.1	1884.3*	10.4	34.3
T for $\sigma = 0.005$	0.1	25.4	5.8	1205.9	6.7	22.0
95% Chi square	519.4	830.5	706.2	1488.2	1008.3	1519.1

*denotes rejection of the null at the 5% level.

T, defined by (6) as well as the values of the appropriate chi-square distribution at the 95% level.[16]

We first consider the test results for Group 1. If the standard deviation of measurement errors in the data is at least 0.1% (i.e., if $\sigma \geq 0.001$), we would be unable to reject the null hypothesis in any sample in Group 1. If it is 0.075% (i.e., if $\sigma = 0.00075$), we would only reject for $S3$. If it is 0.05%, we would only reject for $S3$ and $S7$. Although, we do not know the true standard deviation of measurement errors in the data, the test results indicate that our confidence in the observed data would have to be quite high to consider rejecting the null hypothesis in $S1$, $S4$, or $S6$.[17] If measurement errors were mean-zero and normally distributed with a standard deviation of 0.05%, then two standard deviations about the mean would correspond to measurement errors in the range of ±0.1% of the measured asset stocks, which does not seem unreasonable. The results are only slightly less supportive for the other two samples in Group 1.

The results are much less supportive of the null hypothesis in $S5$ than for any sample in Group 1. Specifically, if the standard deviation of measurement error in the data is 0.5% we would be unable to reject the null hypothesis. We would, however, reject the null if the standard deviation is 0.4% or lower.[18] Thus, the standard deviation of measurement errors would have to be much higher for this sample than it would have to be for any of the samples in Group 1 in order to be unable to reject the null. We would be reluctant to assume that measurement errors are as large as would be implied by a standard deviation of 0.5% unless we had corroborating empirical evidence to that effect.

3.3. The de Peretti (2004) Test

In Table 6, we report results for the computed errors from the de Peretti (2004) adjustment procedure.[19] We report the number of adjusted bundles (num), the adjustment rate (num/I), the sum of squared adjustments (SSR), $RMSE = \sqrt{SSR/(num \cdot K)}$, and descriptive statistics for the nontrivial adjustments.

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

TABLE 6. De Peretti (2004) adjustment procedure by sample period

	S1	S2	S3	S4	S5	S6	S7
Number adjusted (*num*)	1	41	3	3	23	3	4
Adjustment rate (%)	2.78	50.62	6.67	8.82	32.86	7.69	6.15
$SSR * 100^2$	0.10215	25130.16	6.99767	16.07538	1559.48	1.08930	18.66119
$RMSE * 100$	0.08864	6.61670	0.37042	0.53106	1.84124	0.12300	0.46050
Mean $* 100$	0.01651	0.18121	−0.04283	−0.05841	0.27567	−0.02243	−0.02232
Std. dev. $* 100$	0.09065	6.61999	0.37159	0.53253	1.82247	0.12179	0.46259
$max * 100$	0.24410	26.92493	1.58602	2.01969	7.94466	0.42873	2.31305

TABLE 7. De Peretti (2004) test procedure by sample period

	S1	S2	S3	S4	S5	S6	S7
Test Result	Pass	Fail	Pass	Pass	Fail	Fail	Pass
First order							
Auxiliary regression	0.26 [1]	336.71 [2]	0.32 [1]	2.05 [1]	91.56 [3]	2.97 [2]	1.46 [1]
P-value	0.77	0	0.73	0.14	0	0.04	0.23
Q-stat (1–4)	4.50	982.49	0.95	4.50	443.51	22.06	6.06
P-value	0.34	0	0.92	0.34	0	0	0.20
Second order							
Auxiliary regression	2.14 [1]	107.71 [3]	0.33 [1]	1.58 [1]	24.87 [4]	5.45 [3]	0.53 [2]
P-value	0.77	0	0.72	0.71	0	0	0.71
ML-stat (1–4)	0.27	628.17	1.69	1.58	206.26	37.75	1.30
P-value	0.99	0	0.79	0.81	0	0	0.86

Note: Lag for auxiliary regression is in brackets.

The adjustment procedure found a solution in all sample periods and was very fast.

Given the findings in Table 3, we might expect relatively small computed errors for the samples in Group 1 and relatively large errors for the samples in Group 2, which is exactly what we found. For Group 1, the smallest *RMSE* is for *S*1, consistent with much of our previous discussion. The largest *RMSE* in Group 1 is for *S*4. Only a few bundles are adjusted by the procedure in each of these samples. For *S*1, only one bundle is adjusted, whereas for *S*7, all 18 GARP violations are eliminated by adjusting only 4 bundles (adjustment rate of 6.15%). The results are very different for Group 2. The *RMSE* for *S*2 and *S*5 are much larger than for any Group 1 sample. The number of adjusted bundles was also much larger with adjustment rates of 50.62% for *S*2 and 32.86% for *S*5.

The descriptive statistics for each group show a similar pattern. For Group 1, the maximal absolute errors do not exceed 2.32% for any sample. The computed errors in Group 1 are very small relative to those in Group 2. For Group 2, the errors are bounded by ±26.93% for *S*2 and ±7.95% for *S*5, indicating that highly nontrivial adjustments in the observed data are needed in order to satisfy utility maximization.

In Table 7, we present the results of the *i.i.d.* tests. We provide F-tests for the hypothesis that all of the coefficients of the auxiliary regressions (11) and (12) are jointly zero. In addition, we provide the Ljung-Box Q-statistic for first-order independence and the McLeod and Li ML statistic for second-order independence. The tests are performed using the pooled nontrivial computed errors.

For the samples *S*1, *S*3, *S*4, and *S*7, the null hypothesis cannot be rejected, meaning that the computed adjustments are independent and identically

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

distributed, and the GARP violations should not be considered significant. Utility maximization is, therefore, accepted in these samples. For $S2$ and $S5$, the *i.i.d.* property is strongly rejected by all tests. Maximization is, therefore, rejected for both samples in Group 2. Interestingly, the utility maximization hypothesis is also rejected for $S6$.

We note that the number of GARP violations is systematically related to the number of adjusted bundles, with larger numbers of violations being associated with a higher adjustment rate. A consequence of this is that the auxiliary regressions are estimated with many more observations for $S2$ and $S5$ than for the other samples. The issue is particularly relevant to $S1$, where only one bundle was adjusted.

3.4. Comparison

We begin by comparing the results of the adjustment procedures. The computed errors from the procedures differ in two key respects. First, the de Peretti adjustment procedure adjusts a subset of the bundles, whereas the Varian adjustment procedure can (and typically does) adjust all bundles. The percentage of bundles adjusted by the de Peretti procedure was less than 10% for all samples in Group 1, but was larger for the samples in Group 2. Second, in theory, the de Peretti adjustments should have larger sum of squares as previously discussed. The *SSR* from the de Peretti procedure is higher than for the Varian procedure in all samples, except for $S6$. The *SSR* for the two procedures are approximately equal for $S3$. These results largely conform to our theoretical prediction.

The results from the two test procedures can be compared for six of the seven samples. For Group 1, the null hypothesis cannot be rejected for $S1$, $S3$, $S4$ or $S7$ using the de Peretti test, but is rejected for $S6$. The Varian test results are more subjective, because (6) depends on the standard deviation of measurement errors. If we assume that the standard deviation of measurement errors in the data is at least 0.075% then we would be unable to reject the null hypothesis for $S1$, $S4$, $S6$ or $S7$. For the remaining sample in Group 1, $S3$, we would reject the null if the standard deviation of measurement error were 0.075%, but we would be unable to reject if it were 0.1%. The Varian test is slightly less supportive for $S3$ than for the other samples, but our level of confidence in the observed data would have to be fairly high to consider the rejections significant in any of these five samples. We conclude that the results of the two tests are fairly consistent in Group 1. The main discrepancy between the results (in Group 1) is in regard to $S6$, where the de Peretti test rejected the null.

The results for $S6$ are interesting, because the number of GARP violations is low, the SSR obtained by both adjustment procedures is low, and the computed errors produced by both procedures are small. Nevertheless, the de Peretti test rejects in this sample, because the computed errors from the corresponding adjustment procedure are found to violate independence. The results for this sample underscore the differences in the decision rules of the two tests.

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

For Group 2, we were unable to compare results for $S2$, because we could not compute the minimal perturbation for this sample.[20] The de Peretti adjustment procedure eliminated the GARP violations in $S2$ without difficulty, although it produced highly nontrivial adjustments to the data. The de Peretti test strongly rejects the null for both $S2$ and $S5$. Correspondingly, the Varian test provided much less support for the null hypothesis in $S5$ than for any sample in Group 1.

4. CONCLUSIONS

We compared the Varian and de Peretti measurement error tests in theoretical terms and in an empirical application. In the empirical application, we used the two tests to determine if GARP violations were statistically significant for a large set of U.S. monetary assets over multiple sample periods spanning 1960–1992. The tests both supported the null hypothesis of utility maximization in the majority of the samples that had low numbers of GARP violations, although there were some discrepancies between them. The tests also could be directly compared in one sample that had a large number of GARP violations. The de Peretti test rejected the null in that sample and the Varian test was much less supportive of the null than it was in any of the other samples. The results from the two tests were, therefore, largely consistent with each other and conformed to our expectations, which were based on available Monte Carlo evidence. We suggest that researchers should consider using both tests in empirical studies until more extensive comparative results become available. These tests can both be run on datasets with reasonable sample sizes and numbers of goods, although the Varian test is much more time consuming to run.

Barnett and Choi (1989) evaluated both parametric and nonparametric weak separability tests in an extensive simulation study. They found that "[r]elative to the separability testing criterion, all models usually did poorly" and that ". . . the problem is not solved by conducting the test at a point, rather than globally, or by the use of Varian's nonparametric [weak separability] test." A promising direction for future research may be to incorporate measurement errors into nonparametric tests of weak separability. See Swofford and Whitney (1994, p. 248) for some supportive comments along these lines.

GARP is often tested in connection with testing for weak separability. In particular, Varian (1983) proposed a nonparametric test for weak separability based on a sequence of GARP tests, which has been widely used. The tests in this article can be used to determine if violations of GARP could have been caused by measurement errors in the observed data. The adjusted data produced by either of the two tests satisfies GARP and could, therefore, be used in subsequent tests or estimations that require GARP (see Varian, 1985, p. 449, for more general comments along the same lines). For example, Jones, Dutkowsky, and Elger (2005) calculated adjusted data using the Varian (1985) procedure, which they used in subsequent weak separability tests. A more sophisticated approach would be to generalize the measurement error approaches described in this article to test violations of

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

weak separability for significance. See, for example, de Peretti (2005a) for a weak separability test within an explicit stochastic framework.

Another direction for future research is to compare the two test procedures in a formal simulation study. De Peretti (2005) provides simulation results for his test, but there are no comparable results available (that we are aware of) for the Varian (1985) test. The computational burden of the Varian (1985) adjustment procedure is, however, a limiting factor at present. In addition, the results of the Varian test depend on the measurement error variance specified by the tester.

ACKNOWLEDGMENTS

The content of this joint article is based on a comment and reply to de Peretti (2004). We thank William Barnett for suggesting that we write a joint article, or "exchange," synthesizing our views. We thank Dan Thornton for a helpful discussion regarding the dataset used in this paper. We also thank Bob Rasche for providing us with historical monetary data, which we used in an earlier draft. We thank Don Dutkowsky, David Edgerton, Thomas Elger, and Chris Hanes for helpful comments and discussions.

Barry Jones was a visitor at Syracuse University when this paper was written. He thanks the Department of Economics at Syracuse University and the Department of Economics at Lund University for helpful comments in seminars. He also thanks AEM Design for providing FFSQP and the Jan Wallander and Tom Hedelius Foundation (Grant J03/19) for financial support.

All interpretations are the responsibility of the authors.

NOTES

1. They considered several different specifications of expenditure shares and price distributions. Looking across all their specifications, with 5% measurement errors, GARP was satisfied between 79.4% and 94% of the time (see Fleissig and Whitney, 2003, p. 137).

2. Weak separability implies that the marginal rates of substitution between goods in the separable group do not depend on the quantities consumed of goods outside the group. Consequently, demand for goods in the separable group depends only on group expenditure and group prices.

3. Fleissig and Whitney (2003, p. 138) also calculate the proportion of trials that are consistent with a necessary condition for a subgroup of goods to be weakly separable assuming that the dataset passed GARP. The necessary condition is that the prices and quantities of the subgroup are consistent with GARP. The data satisfied this necessary condition between 74% and 88.7% of the time, with measurement errors in the range $\pm 5\%$. Nonparametric weak separability tests will, therefore, reject the null much of the time because of measurement errors, given the violation rates for both GARP and the necessary condition.

4. Various nonstochastic methods have been proposed for assessing the severity of GARP violations. Chalfant and Alston (1988) proposed a measure for two observations that violate the Weak Axiom of Revealed Preference (WARP). Afriat (1967) and Varian (1990) developed an *inefficiency index*, which measures the "margin of error" in the consumer's optimization needed to eliminate all violations. Gross (1995) describes several additional methods, including the maximum possible violations and the violation rate.

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

5. See Epstein and Yatchew (1985) for an alternative test.

6. There were a very high number of GARP violations in the remaining sample, but we were unable to successfully implement the Varian test in that case. We did run the de Peretti test, however, and the results were similar for the two samples with high numbers of GARP violations.

7. Note that if the observed data actually satisfy GARP, then $F(\hat{\zeta}_1, \ldots, \hat{\zeta}_I) = 0$ by definition.

8. We also could have formulated multivariate tests for independence that explicitly account for the dependence pattern in specific goods.

9. See Varian (1985, p. 450–451) for related discussion.

10. The dataset is also inconsistent in some respects with the newer dataset described in Anderson, Jones, and Nesmith (1997).

11. Gross (1995) discusses more sophisticated definitions of the violation rate. Varian (1990) defines the Afriat efficiency index. The efficiency index is defined through the relations R_e^0 and P_e^0. The relations are defined as follows: $\mathbf{x}_i R_e^0 \mathbf{x}_j$ if $e\mathbf{p}_i \cdot \mathbf{x}_i \geq \mathbf{p}_i \cdot \mathbf{x}_j$ and $\mathbf{x}_i P_e^0 \mathbf{x}_j$ if $e\mathbf{p}_i \cdot \mathbf{x}_i > \mathbf{p}_i \cdot \mathbf{x}_j$. GARP_e is defined (using these relations) analogously to GARP. The efficiency index is the largest number, e^*, between 0 and 1, such that the data is consistent with GARP_{e^*}. We report the *inefficiency index* as a percentage, $100(1 - e^*)$, in Table 3.

12. The test results were analyzed separately and then reported to the other author. The interpretations of the results were, however, written jointly.

13. The programs are written in FORTRAN. See Zhou, Tits, and Lawrence (1997) for detailed description of FFSQP.

14. Specifically, the solver was unable to find a feasible point for the constraints in $S2$.

15. That is, $max = max_{i=1,\ldots,I:\, k=1,\ldots,K}\{|\hat{\varepsilon}_{i,k}|\}$.

16. We did not compute p-values, because the computed errors are unlikely to be normally distributed.

17. The results for $S1$ are particularly strong. Even if the standard deviation of measurement errors is as low as 0.025%, we would still not be able to reject the null for $S1$.

18. It is easy to determine that $\sigma = 0.0045$ (or 0.45%) is the approximate value where the accept/reject result changes.

19. All programs were written using SAS/IML language.

20. The problem does not appear to be sample size. We speculate that the problem is instead related to the extremely high number of GARP violations. As previously noted, Fleissig and Whitney (2003) found that measurement error did not tend to produce high numbers of GARP violations, which suggests that the problem may not be very important in practice.

REFERENCES

Afriat, S.N. (1967) The construction of utility functions from expenditure data. *International Economic Review* 8, 67–77.

Anderson, R.G., B.E. Jones, and T.D. Nesmith (1997) Building new monetary services indexes: concepts, data, and methods. *Federal Reserve Bank of St. Louis Review* 79(1), 53–82.

Chalfant, J.A. and J.M. Alston (1988) Accounting for changes in tastes. *Journal of Political Economy* 96, 391–410.

Barnett, W.A. (1978) The user cost of money. *Economics Letters* 1, 145–149. Reprinted in W.A. Barnett and A. Serletis (eds.), *The Theory of Monetary Aggregation*, pp. 6–10. Amsterdam: North-Holland.

Barnett, W.A. (1980) Economic monetary aggregates: an application of index numbers and aggregation theory. *Journal of Econometrics* 14, 11–48. Reprinted in W.A. Barnett and A. Serletis (eds.), *The Theory of Monetary Aggregation*, pp. 11–48. Amsterdam: North-Holland.

Barnett, W.A. (1982) The optimal level of monetary aggregation. *Journal of Money, Credit, and Banking* 14, 687–710. Reprinted in W.A. Barnett and A. Serletis (eds.), *The Theory of Monetary Aggregation*, pp. 125–149. Amsterdam: North-Holland.

Barnett, W.A. and S. Choi (1989) A Monte Carlo study of tests of blockwise weak separability. *Journal of Business and Economic Statistics* 7, 363–377. Reprinted in W. Barnett and J. Binner

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press

(eds.) (2004), *Functional Structure and Approximation in Econometrics*, pp. 257–287. Amsterdam: North Holland.

Barnett, W.A. and A. Serletis (2000) *The Theory of Monetary Aggregation*. Amsterdam: North Holland.

de Peretti, P. (2005a) Testing the significance of the departures from utility maximization. *Macroeconomic Dynamics* 9, 372–397.

de Peretti, P. (2005) A Multi-Step and Iterative Stochastic Nonparametric Test for Weak Separability. Unpublished manuscript, Université Paris 1 Panthéon-Sorbonne.

Fisher, D. and A.R. Fleissig (1997) Monetary aggregation and the demand for assets. *Journal of Money, Credit, and Banking* 29, 458–475.

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, 133–144.

Gross, J. (1995) Testing data for consistency with revealed preference. *Review of Economics and Statistics* 77(4), 701–710.

Jones, B.E., D.H. Dutkowsky, and T. Elger (2005) Sweep programs and optimal monetary aggregation. *Journal of Banking and Finance* 29, 483–508.

Spanos, A. (1999) *Probability Theory and Statistical Inference*. Cambridge: Cambridge University Press.

Swofford, J.L. and G. Whitney (1994) A revealed preference test for weakly separable utility maximization with incomplete adjustment. *Journal of Econometrics* 60, 235–249.

Thornton, D. and P. Yue (1992) An extended series of divisia monetary aggregates. *Federal Reserve Bank of St. Louis Review* 74(Nov./Dec.), 35–46.

Varian, H.R. (1982) The nonparametric approach to demand analysis. *Econometrica* 50(4), 945–973.

Varian, H.R. (1983) Non-parametric tests of consumer behavior. *Review of Economic Studies* 50, 99–110.

Varian, H.R. (1985) Non-parametric analysis of optimizing behavior with measurement error. *Journal of Econometrics* 30, 445–458.

Varian, H. R. (1990), Goodness-of-fit in optimizing models. *Journal of Econometrics* 46, 125–140.

Epstein, L.G. and A. Yatchew (1985) Nonparametric hypothesis testing procedures and applications to demand analysis. *Journal of Econometrics* 30, 149–169.

Zhou, J.L. A.L. Tits, and C.T. Lawrence (1997) User's guide for FFSQP Version 3: a Fortran code for solving optimization programs, possibly Minimax, with general inequality constraints and linear equality constraints, generating feasible iterates. Institute for Systems Research, University of Maryland, Technical Report SRC-TR-92-107r5, College Park, MD.

https://doi.org/10.1017/S1365100505040381 Published online by Cambridge University Press