DISCUSSION PAPER SERIES

IZA DP No. 2004

Nonparametric Analysis of Household Labor Supply: Goodness-of-Fit and Power of the Unitary and the Collective Model

Laurens Cherchye
Frederic Vermeulen

March 2006

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Nonparametric Analysis of Household Labor Supply: Goodness-of-Fit and Power of the Unitary and the Collective Model

Laurens Cherchye
Catholic University of Leuven
and Fund for Scientific Research - Flanders (FWO)

Frederic Vermeulen
CentER, Tilburg University
and IZA Bonn

Discussion Paper No. 2004
March 2006

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 2004
March 2006

ABSTRACT

Nonparametric Analysis of Household Labor Supply: Goodness-of-Fit and Power of the Unitary and the Collective Model*

We compare the empirical performance of unitary and collective labor supply models, using representative data from the Dutch DNB Household Survey. We conduct a nonparametric analysis that avoids the distortive impact of an erroneously specified functional form for the preferences and/or the intrahousehold bargaining process. Our analysis focuses on the goodness-of-fit of the two behavioral models. To guarantee a fair comparison, we complement this goodness-of-fit analysis with a power analysis. Our results strongly favor the collective approach to modeling the behavior of multi-person households.

JEL Classification: C14, D12, J22

Keywords: labor supply, collective model, unitary model, nonparametric analysis, revealed preferences

Corresponding author:

Frederic Vermeulen
Department of Econometrics and OR
Tilburg University
P.O. Box 90153
NL-5000 LE Tilburg
The Netherlands
Email: frederic.vermeulen@uvt.nl

* We thank the Co-Editor and two anonymous referees for numerous insightful comments and suggestions. We are also grateful to Jan Annaert, Denis Beninger, Martin Browning, Pierre-André Chiappori, André Decoster, David Margolis, Erik Schokkaert and Dirk Van de gaer, as well as seminar participants in Copenhagen, Ghent, Leuven, Tilburg, ESEM 2003 Stockholm and JMA 2004 Lille for useful remarks. Frederic Vermeulen acknowledges the financial support provided through the European Community's Human Potential Programme under contract HPRN-CT-2002-00235, (AGE).

1 Introduction

Standard microeconomic theory assumes that a household acts as if it were a single decision maker. Within this tradition, household demand is assumed to result from maximizing a unique utility function subject to a household budget constraint. However, a growing body of evidence suggests that this *unitary* model is at odds with observed household behavior; the associated restrictions of homogeneity, symmetry and negativity have been rejected at numerous occasions (e.g., Fortin and Lacroix, 1997, and Browning and Chiappori, 1998).

A more recent alternative, the so-called *collective* approach to household behavior (Chiappori, 1988, 1992), explicitly takes account of the fact that multiperson households consist of several individuals with their own rational preferences; household decisions are then the Pareto efficient outcomes of a bargaining process. This collective approach entails other behavioral restrictions than the unitary model. Interestingly enough, these restrictions seem to better fit the data than the unitary restrictions; e.g., Browning et al. (1994), Fortin and Lacroix (1997), Browning and Chiappori (1998), Chiappori et al. (2002) and Vermeulen (2005).

Still, the hitherto employed tests of the unitary and collective models are *parametric* in nature. Hence, they crucially depend on the functional form that is used for representing the preferences and/or the intrahousehold bargaining process. They do not only test the unitary or collective approach as such, but also an *ad hoc* functional specification; a rejection of the unitary model may well be due to ill-specification.

Nonparametric tests for consistency of observed behavior with utility maximization or Pareto efficiency do not require any assumptions regarding the parametric form of utility functions or the intrahousehold bargaining process; see, e.g., Afriat (1967), Varian (1982), Chiappori (1988) and Snyder (2000). These tests are solely based on revealed preference theory, which makes them particularly attractive for testing consistency of the data with theoretical behavioral models.

This directly suggests using nonparametric testing tools for comparing the empirical performance of the unitary and collective models. However, to the best of our knowledge, an in-depth nonparametric comparison has not yet been carried out. This paper wants to fill that gap, by studying the specific case of household labor supply. Conveniently, our focus on labor supply also guarantees substantial price/wage variation across individuals, which can only benefit the empirical comparison.

Our following assessment specifically concentrates on two types of (nonparametric) empirical performance measures: *goodness-of-fit* measures and *power* measures. We indeed believe that a fair comparison of the two behavioral models under study should complement a goodness-of-fit analysis with a power analysis: favorable goodness-of-fit results, indicating few violations of the behavioral restrictions, have little meaning if the behavioral implications have low power, i.e., optimizing behavior can hardly be rejected.

Our empirical evaluation uses representative Dutch data taken from the DNB

Household Survey. The data set that we focus on is divided in three subsamples: working female singles, working male singles and couples where both spouses are working. We essentially discuss two types of comparisons:
- First, we compare the empirical performance of the unitary model for singles with that for couples. The rationale of this comparison is that the standard unitary approach should always be fully applicable to singles, even if it does not well fit the observed behavior of couples. This first comparison should give us a deeper understanding of the harmless/harmful nature of the aggregation assumptions that underlie the unitary modeling of couples' behavior.
- Second, we compare the empirical results of the collective model with those of the unitary model, both applied to the data of couples. Because the collective and unitary models evidently have different implications for couples' behavior, these results should give us a better insight into which of the models does the better job in describing multi-person household consumption behavior.

Section 2 briefly reviews the nonparametric methodology for testing the unitary and the collective labor supply models. In addition, we introduce the nonparametric goodness-of-fit and power measures. Section 3 presents the results of our application to Dutch household data. Section 4 concludes.

2 Methodology

2.1 Testing the unitary model

For the sake of compactness, we only discuss unitary consistency tests for couples with two working individuals (M and F). Our discussion is directly translated to the singles' case.

The nonparametric approach starts from n observations for household consumption and the household members' labor supply. For each household i ($i = 1, ..., n$) we denote the net wage rate and leisure amount of individual I ($I = M, F$) by w_i^I and l_i^I, respectively. (The leisure amount is computed from observed labor supply $\ell_i^I = T - l_i^I$, with T the individuals' time endowment.) Next, we use y_i and c_i to respectively denote household i's nonlabor income and consumption. The household's nonlabor income aggregates the spouses' assignable unearned incomes and, possibly, unearned income that cannot be assigned to one of the spouses. It should be stressed that we focus on a standard static labor supply model; i.e., households are assumed not to save nor to draw down their assets (see, e.g., Blundell and MaCurdy, 1999). Finally, we represent the set of all observations by $S = \left\{\left(c_i, l_i^M, l_i^F, w_i^M, w_i^F, y_i\right), i = 1, ..., n\right\}$.

Within the unitary model, the decision problem of each household i boils down to maximizing a nonsatiated utility function $v\left(c_i, l_i^M, l_i^F\right)$ subject to the budget constraint $c_i + w_i^M l_i^M + w_i^F l_i^F \leq y_i + w_i^M T + w_i^F T$; without losing generality, we set the price of consumption to 1. A necessary and sufficient condition for the data to be consistent with this utility maximization problem is that there exists a function v that *rationalizes* the household data, i.e., for all

$i \in \{1, ..., n\}$ the value $v\left(c_i, l_i^M, l_i^F\right)$ equals:

$$\max_{j \in \{1,...,n\}} v\left(c_j, l_j^M, l_j^F\right) \text{ s.t. } c_j + w_j^M l_j^M + w_j^F l_j^F \leq y_i + w_i^M T + w_i^F T. \qquad (1)$$

Varian (1982) has demonstrated that such a data rationalizing utility function exists if and only if the observed set S is consistent with the *Generalized Axiom of Revealed Preference* (*GARP*). To formally state this last consistency condition, we first need the following revealed preference definition (using $\left(1, w^M, w^F\right)' = \mathbf{w}$ and $\left(c, l^M, l^F\right)' = \mathbf{l}$):

Definition 1 *A bundle* $\mathbf{l}_i$ *is* revealed preferred *to a bundle* $\mathbf{l}$*, denoted by* $\mathbf{l}_i R \mathbf{l}$*, if* $\mathbf{w}_i'\mathbf{l}_i \geq \mathbf{w}_i'\mathbf{l}_j$*,* $\mathbf{w}_j'\mathbf{l}_j \geq \mathbf{w}_j'\mathbf{l}_k$*,...,* $\mathbf{w}_m'\mathbf{l}_m \geq \mathbf{w}_m'\mathbf{l}$ *for some sequence of bundles* $(\mathbf{l}_i, \mathbf{l}_j, ..., \mathbf{l}_m)$*.*

We can now define GARP as:

Definition 2 *The observed set* S *satisfies GARP if for all* $j \in \{1, ..., n\} : \mathbf{w}_j'\mathbf{l}_j = \min_{\mathbf{l} \in \mathbf{RP}_j} \mathbf{w}_j'\mathbf{l}$ *for* $RP_j = \{\mathbf{l}_i : \mathbf{l}_i R \mathbf{l}_j; i \in \{1, ..., n\}\}$*.*

This definition expresses the idea that observation j is utility maximizing subject to its budget constraint if and only if it is expenditure minimizing over its 'better than' set; in the (empirical) GARP this last set is approximated by the 'revealed preferred' set RP_j.

Consistency of S with GARP is easily tested: we first identify the sets RP_j and subsequently check the expenditure minimization condition for each observation. See Varian (1982; p. 949) for an efficient algorithm.

2.2 Testing the collective model

We focus on a collective model with *egoistic* preferences; preferences only depend on own consumption and leisure (Chiappori, 1988). Moreover, we assume that there is no public consumption in the household.[1] Empirically, the modeling of this collective approach is somewhat more involved as the private consumption of each household member is usually not observed; labor supply data sets only reveal information on *total* household consumption (as the sum of earned and unearned incomes).

In the following, we denote individual I's private consumption by c_i^I, and the vectors $\left(1, w_i^I\right)'$ and $\left(c_i^I, l_i^I\right)'$ by respectively $\mathbf{w}_i^I$ and $\mathbf{l}_i^I$ $(I = M, F)$. Using this, we consider the case where each couple i is characterized by a pair of

[1] The analysis is in fact also applicable to individual *caring preferences*, which can be represented by a utility function of the form $f^I\left(v^M\left(c^M, l^F\right), v^F\left(c^F, l^F\right)\right)$ $(I = M, F)$; see Chiappori (1992) for a detailed discussion. Cherchye, De Rock and Vermeulen (2004) provide a nonparametric characterization of a general collective model with public goods and externalities. Given the current paper's objective, we focus on a rather simple collective model, which can be considered as a direct generalization of the unitary model. Of course, if this rudimentary collective model outperforms the unitary model for describing couples' behavior, then this will certainly be the case for any more refined collective model.

(nonsatiated) utility functions, $v^M\left(c_i^M, l_i^M\right)$ and $v^F\left(c_i^F, l_i^F\right)$, *and* a sharing rule $\phi\left(w_i^M, w_i^F, y_i\right)$ that determines the distribution of the household's nonlabor income y_i over the household members (see Chiappori, 1988):

Definition 3 *A* sharing rule ϕ *is a function which maps the vector* $\left(w_i^M, w_i^F, y_i\right)'$ *to* $\phi\left(w_i^M, w_i^F, y_i\right) = \left(y_i^M, y_i^F\right)'$ *such that* $y_i^M + y_i^F = y_i$.

The sharing rule concept allows us to model household behavior as a two-stage budgeting process. After dividing total nonlabor income in the first stage, each individual I $(I = M, F)$ of the couple i faces the maximization problem:

$$\max_{c_i^I, l_i^I} v^I\left(c_i^I, l_i^I\right) s.t.\ c_i^I + w_i^I l_i^I \leq y_i^I + w_i^I T,$$

which is formally similar to the unitary household decision problem; see (1). Chiappori (1992) demonstrated that the resulting household allocation is always Pareto efficient.

This alternative interpretation of Pareto efficient household behavior is particularly convenient within the nonparametric context, as it entails the same kind of GARP tests as for the unitary model: if we knew private consumption for each observation (c_i^M and c_i^F), then we could immediately check consistency of the observed set S by using the standard GARP tests at the level of the *household members*. In practice, however, we do *not* observe the intrahousehold allocation of total consumption. This entails the following empirical condition for the collective model (see also Chiappori, 1988):

Definition 4 *The observed set* S *is consistent with a* collective rationalization with egoistic agents *if there exist* n *pairs of real numbers* $\left(c_i^M, c_i^F\right)'$ *such that for all* $i = 1, ..., n$*:*
$c_i^M + c_i^F = c_i$,
$c_i^M, c_i^F \geq 0$,
$c_i^M + c_i^F + w_i^M l_i^M + w_i^F l_i^F \leq y_i + w_i^M T + w_i^F T$
and
GARP is satisfied at the individual level $(I = M, F)$*:*
$\forall i, j \in \{1, ..., n\}$, *if* $\mathbf{l}_i^I R \mathbf{l}_j^I$ *then* $\mathbf{w}_j^{I\prime}\mathbf{l}_j^I \leq \mathbf{w}_j^{I\prime}\mathbf{l}_i^I$.

Thus, given that the intrahousehold consumption allocation is not observed, we only need that there exists *at least one feasible* allocation entailing *individual* data $\{(c_i^I, l_i^I, w_i^I, y_i^I = c_i^I - w_i^I \ell_i^I);\ i = 1, ..., n;\ I = M,\ F\}$ that are consistent with GARP for *both* individuals.

Snyder (2000) introduced an 'all-or-nothing' nonparametric test for the collective model.[2] In that test, either data satisfy collective rationality or they do not. We follow a different approach, induced by our specific focus on the goodness-of-fit of the alternative behavioral models. Our starting point is that the collective rationalization test boils down to standard GARP tests conditional

[2] In her analysis, Snyder restricts attention to the case n=2, while we consider the more general case; e.g., in our application n=586 (see Section 3.1).

upon an intrahousehold consumption allocation (c_i^M and c_i^F). Specifically, we impute (unobserved) member-specific private consumption amounts by exploiting a systematic finding in parametric studies of collective labor supply, namely the positive correlation between the male/female member's share of total nonlabor income and the corresponding individual wage (e.g., Chiappori et al., 2002, and Vermeulen, 2005).

Using this, our nonparametric testing exercise considers the following pair of distributions for the female consumption share s_i^F ($= c_i^F/c_i$; the corresponding male share equals $1-s_i^F$): the first distribution has mean 0.40 and a cumulative probability of 95% for the values between 0.35 and 0.45; the second distribution has mean 0.60 and a cumulative probability of 95% for the values between 0.55 and 0.65. From these distributions, we draw 1000 combinations of s_i^F values ($i = 1, ..., n$): if $w_i^M \geq w_i^F$ ($w_i^F > w_i^M$) then s_i^F is drawn from the first (second) distribution. We subsequently select the combination of shares with the highest number of individual (male and female) household members passing GARP. This combination is used for comparing the empirical performance of the collective model with that of the unitary model.[3]

As a final note, we must emphasize that this approach does not guarantee the most favorable treatment of the collective model: to ensure computational tractability, our procedure restricts attention to a limited number of possible combinations of intrahousehold allocations; there may well exist other, non-investigated, combinations that are associated with an even higher number of individuals consistent with GARP. We can therefore state that our empirical analysis implicitly gives the 'benefit of the doubt' to the unitary model.

2.3 Empirical performance: goodness-of-fit

The consistency tests reviewed above are 'sharp' tests; they only tell us whether observations are *exactly* optimizing in terms of the behavioral model that is under evaluation. However, as argued by Varian (1990), *exact* optimization is not a very interesting hypothesis. Rather, we want to know whether the behavioral model under study provides a *reasonable* way to describe observed behavior; for most purposes, 'nearly optimizing behavior' is just as good as 'optimizing' behavior. Varian's argument is all the more valid in the context of comparing theoretical behavioral models: we are primarily interested in the extent to which one model 'fits' the observed data better than the other model. Therefore, our following assessment will be based on measures of *goodness-of-fit*.

Our goodness-of-fit measure is the 'improved violation index' (or 'efficiency index') proposed by Varian (1993; based on Afriat, 1973; see also Cox, 1997), which indicates the *degree* to which the data are 'optimizing' (or 'efficient') in the sense of the evaluated behavioral model. More specifically, this index gives for each observation the minimal perturbation of the expenditure level that

[3] We also experimented with alternative means for the above normal distributions (including a rule where both distributions have mean 0.50). But this did not yield a higher number of (male and female) household members passing GARP.

guarantees consistency of the observed set S with GARP. See Varian (1993) and Cox (1997) for in-depth formal discussions of this goodness-of-fit measure.

2.4 Empirical performance: power

We compute six different power measures. A first distinction relates to the consumption data that is used. The first data set (used for the measures *Power1a*, *Power1b* and *Power1c* that will be introduced below) consists of the original consumption and leisure data for the unitary model and the combination of observed and partly imputed data for the collective model. The second data set (used for the measures *Power2a*, *Power2b* and *Power2c*) multiplies the original expenditure level for each observation with the corresponding improved violation index value (at the household level for the unitary model and at the individual level for the collective model); this anticipates the question to what extent the necessary data perturbation for obtaining GARP consistency (captured by the improved violation index) would effectively impact on the power estimates.

For each data set we compute two types of power measures proposed by Bronars (1987) and one additional measure. Bronars' two measures essentially pertain to the 'mimicking' of irrational behavior *à la* Becker (1962), by means of a specific randomization procedure for constructing irrational consumption bundles. Each power measure then captures the probability of detecting that irrational behavior, which acts as the alternative hypothesis to the null hypothesis that is tested. Bronars' first measure, labeled *Power1a* (*Power2a*) for the first (second) data set, is based on the alternative hypothesis that consumers choose bundles randomly from a uniform distribution across all bundles in their budget hyperplanes. Since this first power measure may entail quite extreme behavior (e.g., consumers jumping from one 'corner' of the budget line to another 'corner'), we also applied Bronars' second measure, labeled *Power1b* (*Power2b*) for the first (second) data set.This alternative measure makes extreme (irrational) behavior less likely than the first one. (See Bronars, 1987, for formal definitions of both measures).

Our additional third power measure (labeled *Power1c* and *Power2c* when applied to the first and second data set respectively) assumes that consumers randomly draw consumption and leisure bundles from the *empirical distribution* as observed in the data. The rationale for this additional measure pertains to the observation that, e.g., singles who work the same number of hours can never be involved in a GARP violation *vis-à-vis* each other. Since observed working hours are discretely distributed with an important mass point at a weekly labor supply of 40 hours for males, and with mass points at 40 and 32 hours for females, such a situation may apply to a nonnegligible number of observations. Our third power measure accounts for this potential problem and provides a measure for its importance. The measure basically implies that we remain ignorant about the alternative hypothesis to the model under study. Still, we recognize that this power measure may be subject to some criticism. Most importantly, and contrary to Beckerian irrational behavior, random behavior based on the empirical distribution may seem logically inconsistent since it potentially mixes

both rational and irrational behavioral aspects. One should take into account that both the null hypothesis (the model under study) and the (unspecified) alternative hypothesis give rise to exactly the same distributions.

For a given data set and randomization procedure, the specific construction of the power measures first simulates irrational/random behavior for each observation, and subsequently checks consistency with GARP for each observation. In our empirical application, we repeat this procedure 200 times. The proportion of rejections of GARP (over these 200 replications) then gives the probability of detecting irrational/random behavior of each observation, given random behavior of the other observations.

Hence, for each behavioral model that we evaluate we measure power in *each* element of the observed set S. This practice contrasts with e.g. Bronars (1987) and Cox (1997), who provide overall power measures that are based on the *entire* sample. Their measures reveal the probability that random behavior of at least one observation in the sample is detected. In our opinion, evaluating power at the level of individual observations is more informative. For example, it provides a more detailed insight into the extent to which the different observations *can* cause rejection of the model under study; we believe that there is a stronger case for a model that has high power in many observations than for a model with high power in only a few observations. Also, an observation-specific power measure naturally links up with our observation-specific goodness-of-fit measure; persistently high goodness-of-fit values for a given sample of observations are all the more convincing evidence in favor of a particular behavioral model if they are complemented with high power values for the same sample.

3 Application

3.1 Data and methodological issues

We use 11 waves of the DNB Household Survey (formerly known as the CentER Savings Survey), drawn from 1995 until 2005. The data are representative for the Dutch population and are collected every year for a panel of more than 2000 households. The survey contains a rich amount of economic, socio-demographic and psychological variables. We focus on three subsamples: female singles, male singles and couples. The first two subsamples consist of female and male singles that meet the following criteria: no children, aged between 25 and 55 and employed. The third subsample consists of (de-facto) couples, where the household members meet the same criteria as the selected singles. To minimize the impact of measurement error, we trimmed out from each subsample those households that include a (female/male) member with a wage that lies above the 97.5 percentile or below the 2.5 percentile of the empirical (female/male) wage distribution. This yields samples of 522 single females, 888 single males and 586 couples. Table 3 in the Appendix reports descriptive statistics for each subsample.

Cox (1997) and Snyder (2000) conduct nonparametric tests of labor sup-

ply behavior on time-series micro-data and, hence, exclude preference variation over time. Our analysis deviates in that we assume constant preferences in each subsample (female singles, male singles and couples); in each subsample, all observations correspond to the same preferences but to different price regimes. Our motivation for this particular preference homogeneity assumption is threefold. Firstly, the DNB Household Survey was subject to substantial attrition between 1995 and 2005. Only a relatively small number of households were observed in all the waves, which implies too few households with 11 consecutive observations for robust nonparametric testing based on time-series data. Secondly, our selection criteria ensure relatively homogeneous subsamples, which makes that our equal preference assumption does not seem overly strong.[4] Finally, and importantly, recall that we focus on goodness-of-fit measures in our following analysis. Obviously, this practice anticipates some preference variation over households.

3.2 Singles versus couples

Figure 1 presents the cumulative distribution functions (c.d.f.'s) of the goodness-of-fit measures (i.e., the improved violation indexes, in ascending order) associated with the unitary model for female singles, male singles and couples.[5] When restricting to the 'sharp' GARP test, we would conclude rejection for all three subsamples; most observations have an index value that is less than 100%. We note that this result should not be very surprising in view of our preference homogeneity assumption. It seems more meaningful to look at the *entire* distribution of the goodness-of-fit measure.

[Figure 1 about here]

When considering the c.d.f.'s more closely, we observe important differences between couples and singles. Firstly, we find that 39% of the female singles and 28% of the male singles are fully efficient, as opposed to only 12% of the couples. Secondly, and more importantly, the index values of couples are generally below those of singles; the couples' distribution is stochastically dominated by the two singles' distributions. One-tailed Kolmogorov-Smirnov tests confirm this overall picture: the null hypothesis of equal distributions of couples on the one hand and male and female singles on the other hand is strongly rejected in favor of the alternative hypothesis that the couples' index systematically lies below the respective singles' indexes; see Table 1.

As discussed above, it is recommendable to complement this goodness-of-fit analysis with a power analysis. Figure 2 presents the c.d.f.'s of the individually

[4] Compare, e.g., with Famulari (1995) who analyzes consistency of observed behaviour with GARP (in the unitary framework) for homogeneous subgroups of households that are identified on the basis of similar selection criteria.

[5] For expositional convenience, the c.d.f.'s have been cut off at the 91% efficiency level since no observation has a violation index below that figure. We also explicitly distinguish between indexes that are equal to 1 and those that are less than 1.

calculated *Power1a* indexes for single females, single males and couples.[6] This figure reveals high power for most observations: 98% of the couples, 96% of the male singles and 95% of the female singles have a power index value that exceeds 95%; for these observations, irrational random behavior will be detected with a probability of at least 95%. More generally, while the overall power for couples appears to be slightly higher than for female and male singles, Figure 2 suggests that the differences remain marginal. This impression is confirmed by one-tailed Kolmogorov-Smirnov tests: we cannot reject (at any reasonable significance level) equality of the c.d.f.'s in favor of the alternative hypothesis that the power index values for female and male singles are lower than those for couples; see Table 1.

We obtain exactly the same qualitative conclusions for the power measures *Power2a*, *Power1b* and *Power2b* (see also Table 1): power index values are generally very high, while equality of the c.d.f.'s for the three subsamples cannot be rejected. Although power index values are also relatively high for the power measures *Power1c* and *Power2c*, equality of the c.d.f.'s for the three subsamples is rejected. At this point, it is worth recalling the potential criticism on the latter power measures in Section 2.4. From that perspective, these power measures (especially for singles) only give some indication about the importance of the mass points in the empirical distribution (at 40 hours for example) and the consequent failure to come to a GARP rejection for observations at these mass points.[7]

We conclude that the relatively poor performance of the unitary model for describing observed couples' behavior (when compared to singles' behavior) can hardly be attributed solely to higher power of the model for the associated couples' consistency tests. In our opinion, these findings strongly question the harmless nature of the aggregation assumptions in the unitary approach to modeling couples' behavior.

[Figure 2 about here]

[Table 1 about here]

3.3 Unitary versus collective model

Our previous findings cast doubts on the usefulness of the unitary model for analyzing couples' behavior. As a natural next step, we investigate whether the collective approach provides a better alternative for modeling couples' behavior, by comparing its empirical performance with that of the unitary model. Like before, our unitary results refer to GARP tests at the *aggregate household* level. By contrast, our collective results are obtained from applying GARP tests to the *individual members* of each couple, hereby using the intrahousehold allocations obtained by the procedure described in Section 2.2.

[6] In contrast to Figure 1, Figure 2 presents the whole c.d.f. The reason is that a few observations have low power indexes.

[7] Table 4 in the Appendix contains other descriptive statistics on the different power distributions. For the sake of brevity, these statistics are not further discussed here.

Figure 3 presents the c.d.f.'s of the goodness-of-fit measure for couples (in the unitary model) and female and male household members (in the collective model). In line with our earlier results, more individuals than aggregate households behave consistently with the utility maximization hypothesis: 26% of the men and 14% of the women are 100% efficient, while only 12% of the couples attain an improved violation index value of 100%. In fact, Figure 3 reveals a picture that is roughly similar to that in Figure 1: the (unitary) couples' distribution is stochastically dominated by the (collective) distributions of the male and female household members. The Kolmogorov-Smirnov test results in Table 2 provide further evidence in support of the collective model: the null hypothesis of equal c.d.f.'s is strongly rejected in favor of the alternative hypothesis that the couples' improved violation index systematically lies below that for women and men in the collective model.

[Figure 3 about here]

Again, we complement this goodness-of-fit analysis with a power analysis. Our power results persistently indicate that the better fit of the collective model is not due to lower power. For example, Figure 4 clearly shows that the distribution of the *Power1a* values is practically the same for couples (in the unitary model) and individuals (in the collective model). This observation is formalized in Table 2: one-tailed Kolmogorov-Smirnov tests reveal that equality of the c.d.f.'s of the power indexes cannot be rejected at any reasonable significance level. Moreover, the power indexes are generally high: 98% of the couples (in the unitary model), 98% of the females and 98% of the males (in the collective model) have a power index that amounts to at least 95%. Just like in Section 3.2, we checked the sensitivity of these power results. Interestingly, the power measures *Power2a*, *Power1b* and *Power2b* entail the same qualitative conclusions as *Power1a* (see also Table 2). As for the measures *Power1c* and *Power2c*, we only reject equality of the c.d.f.'s when comparing male individuals (in the collective model) with couples (in the unitary model); equality of the c.d.f.'s cannot be rejected when comparing female individuals with couples. Moreover, also these measures imply relatively high power indexes.

In our opinion, these results provide strong enough evidence to argue that the collective approach performs significantly better than the unitary approach for modeling couples' labor supply behavior. In fact, this argument becomes all the more convincing when taking into account our rather rudimentary procedure to model the distribution of household consumption over the different household members; more refined allocation rules can only benefit the relative performance of the collective model.[8]

[8] This conclusion is in line with that obtained by Snyder (2000), who tested unitary and collective labor supply models on couples drawn from the 1969 and 1971 National Longitudinal Survey of Men. More specifically, she focused attention on a sample of 243 couples (with a similar sample selection as ours), where each household is only two times observed (thus assuming preference homogeneity over time, but not at the cross-sectional level as we do). It turns out that collective rationality is not rejected for any couple in her sample, while the unitary model is rejected for 6 couples. Snyder (2000) did not test the unitary model on singles, nor did she conduct a power analysis.

[Figure 4 about here]

[Table 2 about here]

4 Conclusion

We compared the empirical performance of the unitary model to describe household labor supply behavior with that of the more recently developed collective model. Our findings strongly suggest using the collective model for analyzing the behavior of households consisting of multiple individuals:
- First, we found that the unitary model performs significantly worse when applied to couples than when applied to singles. As these results cannot be attributed to power differences, we conclude that they signal violations of the preference aggregation assumptions that underlie the unitary approach, i.e., that multi-person households behave as single decision makers.
- Second, and probably more importantly, a direct comparison of the collective model with the unitary model provided additional evidence to support the use of the collective model: it fits observed couples' behavior much better than the unitary model. Again, this significant difference cannot be explained by power differences. Hence, our findings do not only indicate that the unitary approach is too restrictive for modeling the behavior of multi-person households, but also that the collective model constitutes a more promising alternative.

References

[1] Afriat, Sydney (1967), "The Construction of Utility Functions from Expenditure Data", *International Economic Review*, 8, 67-77.

[2] Afriat, Sydney (1973), "On a System of Inequalities in Demand Analysis: An Extension of the Classical Method", *International Economic Review* 14, 460-472.

[3] Becker, Gary (1962), "Irrational Behavior and Economic Theory", *Journal of Political Economy*, 70, 1-13.

[4] Blundell, Richard and Thomas MaCurdy (1999), "Labor Supply: A Review of Alternative Approaches", in Orley Ashenfelter and David Card (Eds.), *Handbook of Labor Economics. Vol. 3*, Amsterdam, Elsevier Science, pp. 1559-1695.

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

[6] Browning, Martin, François Bourguignon, Pierre-André Chiappori and Valérie Lechene (1994), "Income and Outcomes: A Structural Model of Intrahousehold Allocation", *Journal of Political Economy*, 102, 1067-1096.

[7] Browning, Martin and Pierre-André Chiappori (1998), "Efficient Intra-household Allocations: A General Characterization and Empirical Tests", *Econometrica*, 66, 1241-1278.

[8] Cherchye, Laurens, Bram De Rock and Frederic Vermeulen (2004), "The Collective Model of Household Consumption: A Nonparametric Characterization", *CentER Discussion Paper*, 2004-76, Tilburg, CentER, Tilburg University.

[9] Chiappori, Pierre-André (1988), "Rational Household Labor Supply", *Econometrica*, 56, 63-89.

[10] Chiappori, Pierre-André (1992), "Collective Labor Supply and Welfare", *Journal of Political Economy*, 100, 437-467.

[11] Chiappori, Pierre-André, Bernard Fortin and Guy Lacroix (2002), "Marriage Market, Divorce Legislation and Household Labor Supply", *Journal of Political Economy*, 110, 37-72.

[12] Cox, James (1997), "On Testing the Utility Hypothesis", *Economic Journal*, 107, 1054-1078.

[13] Famulari, Melissa (1995), "A Household-based, Nonparametric Test of Demand Theory", *Review of Economics and Statistics*, 77, 372-382.

[14] Fortin, Bernard and Guy Lacroix (1997), "A Test of the Unitary and Collective Models of Household Labour Supply", *Economic Journal*, 107, 933-955.

[15] Snyder, Susan (2000), "Nonparametric Testable Restrictions of Household Behavior", *Southern Economic Journal*, 67, 171-185.

[16] Varian, Hal (1982), "The Nonparametric Approach to Demand Analysis", *Econometrica*, 50, 945-973.

[17] Varian, Hal (1990), "Goodness-of-fit in Optimizing Models", *Journal of Econometrics*, 46, 125-140.

[18] Varian, Hal (1993), "Goodness-of-fit for Revealed Preference Tests", Mimeo, Ann Arbor, University of Michigan.

[19] Vermeulen, Frederic (2005), "And the Winner is... An Empirical Evaluation of Unitary and Collective Labour Supply Models", *Empirical Economics*, 30, 711-734.

Appendix

[Table 3 about here]
[Table 4 about here]

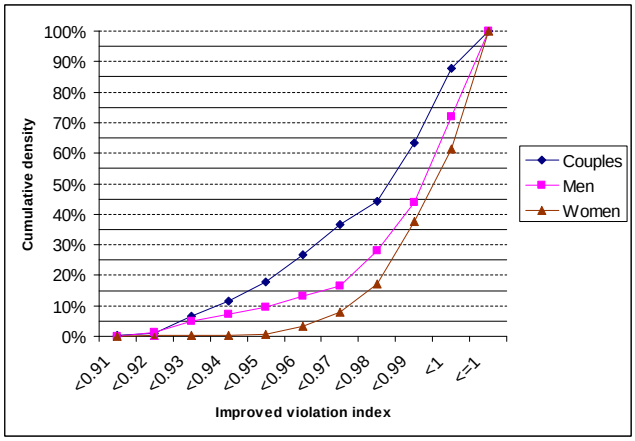

Figure 1: Unitary model singles and couples: cumulative distribution function of improved violation index

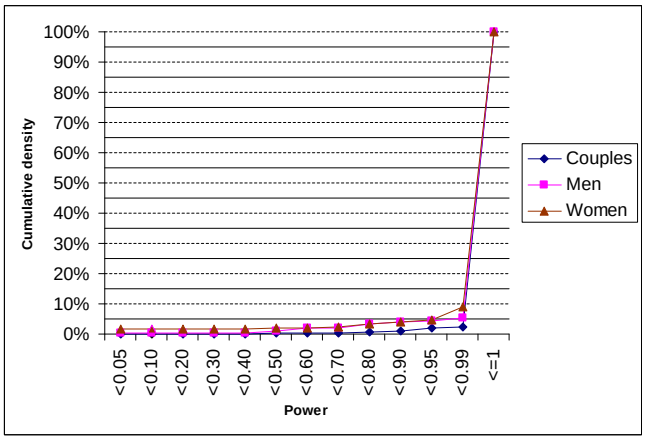

Figure 2: Unitary model singles and couples: cumulative distribution function of power

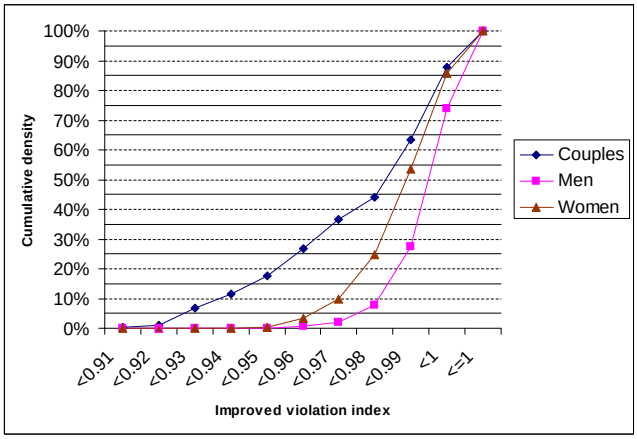

Figure 3: Unitary versus collective model couples: cumulative distribution function of improved violation index

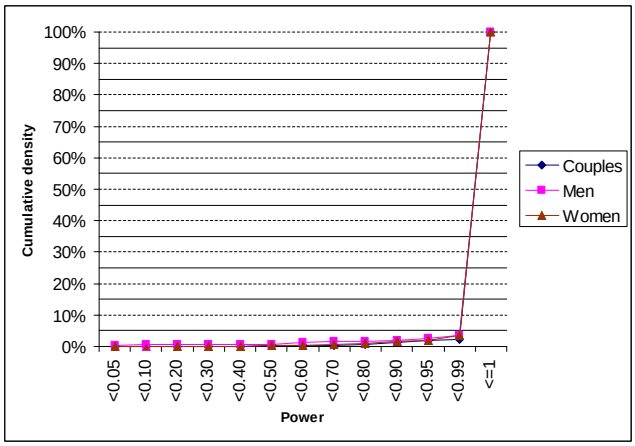

Figure 4: Unitary versus collective model couples: cumulative distribution function of power

Table 1: Nonparametric test results: Singles versus couples

	Single males versus couples	Single females versus couples
Improved violation index	0.000	0.000
Power1a	0.473	0.820
Power1b	0.826	0.278
Power1c	0.000	0.000
Power2a	0.977	0.813
Power2b	0.886	0.620
Power2c	0.000	0.000

Power1a (*Power2a*) is Bronars' (1987) first power measure applied to observed (efficiency corrected) data; *Power1b* (*Power2b*) is Bronars' (1987) second power measure applied to observed (efficiency corrected) data; while *Power1c* (*Power2c*) is based on random drawings from the empirical hours distribution and applied to observed (efficiency corrected) data. Entries show the probability that the null hypothesis of an equal distribution is not rejected, as computed on the basis of a one-tailed Kolmogorov-Smirnov test; we compare the distributions of the improved violation index and the power indexes for couples with the respective distributions for single women and single men.

Table 2: Nonparametric test results: Individuals in couples versus couples

	Men in couples versus couples	Women in couples versus couples
Improved violation index	0.000	0.000
Power1a	0.978	0.962
Power1b	0.972	0.994
Power1c	0.000	0.961
Power2a	0.962	0.994
Power2b	0.962	0.778
Power2c	0.002	0.990

Power1a (*Power2a*) is Bronars' (1987) first power measure applied to observed (efficiency corrected) data; *Power1b* (*Power2b*) is Bronars' (1987) second power measure applied to observed (efficiency corrected) data; while *Power1c* (*Power2c*) is based on random drawings from the empirical hours distribution and applied to observed (efficiency corrected) data. Entries show the probability that the null hypothesis of an equal distribution is not rejected, as computed on the basis of a one-tailed Kolmogorov-Smirnov test; we compare the distributions of the improved violation index and the power index for couples in the unitary model with the respective distributions for women and men in the collective model.

Table 3: Sample descriptive statistics

	Couples		Single males		Single females	
Variable	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.
Weekly working hours male	37.80	3.99	37.29	4.15		
Weekly working hours female	31.55	7.90			35.27	5.14
Net hourly wage rate male	12.07	4.09	12.18	4.24		
Net hourly wage rate female	9.96	3.36			11.08	3.59
Age male	40.77	9.23	40.11	7.52		
Age female	38.52	9.10			38.88	8.55
Dummy primary education male	0.13		0.09			
Dummy secondary education male	0.35		0.36			
Dummy non-academic higher education male	0.33		0.36			
Dummy academic higher education male	0.19		0.19			
Dummy primary education female	0.11				0.07	
Dummy secondary education female	0.41				0.27	
Dummy non-academic higher education female	0.35				0.47	
Dummy academic higher education female	0.13				0.19	
Weekly nonlabor income	17.49	42.53	13.39	32.13	9.81	28.88
Weekly consumption	782.46	223.17	464.35	163.74	399.38	140.28

Source: DNB Household Survey 1995-2005. Weekly consumption equals the sum of earned income and nonlabor income. Monetary variables are in 2005 euro.

Table 4: Descriptive statistics power indexes

Variable	Single males	Single females	Couples	Men in couples	Women in couples
Power1a					
Mean	98.24	97.43	99.63	98.88	99.41
Standard dev.	0.32	0.59	0.13	0.34	0.19
Median	100	100	100	100	100
Power1b					
Mean	98.62	97.17	99.39	98.67	99.20
Standard dev.	0.28	0.61	0.19	0.38	0.24
Median	100	100	100	100	100
Power1c					
Mean	85.47	78.58	97.48	93.98	98.26
Standard dev.	0.88	1.52	0.54	0.70	0.44
Median	99.00	98.50	100	100	100
Power2a					
Mean	99.61	98.94	99.58	98.87	99.41
Standard dev.	0.14	0.28	0.14	0.34	0.20
Median	100	100	100	100	100
Power2b					
Mean	99.22	97.93	99.42	98.75	99.19
Standard dev.	0.17	0.42	0.19	0.37	0.25
Median	100	100	100	100	100
Power2c					
Mean	87.90	81.16	97.80	94.74	98.12
Standard dev.	0.85	1.39	0.48	0.67	0.45
Median	100	99.5	100	100	100

Power1a (*Power2a*) is Bronars' (1987) first power measure applied to observed (efficiency corrected) data; *Power1b* (*Power2b*) is Bronars' (1987) second power measure applied to observed (efficiency corrected) data; while *Power1c* (*Power2c*) is based on random drawings from the empirical hours distribution and applied to observed (efficiency corrected) data. Entries are in percent.