Deutsches Institut für
Wirtschaftsforschung

DIW BERLIN

Discussion Papers

1074

André Decoster • Peter Haan

Empirical Welfare Analysis in Random Utility Models of Labour Supply

Berlin, November 2010

Opinions expressed in this paper are those of the author(s) and do not necessarily reflect views of the institute.

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN print edition 1433-0210
ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
http://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
http://ideas.repec.org/s/diw/diwwpp.html
http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

Empirical welfare analysis in random utility models of labour supply[*]

André Decoster[†]and Peter Haan[‡]

November 2010

Abstract

The aim of this paper is to apply recently proposed individual welfare measures in the context of random utility models of labour supply. Contrary to the standard practice of using reference preferences and wages, these measures preserve preference heterogeneity in the normative step of the analysis. They also make the ethical priors, implicit in any interpersonal comparison, more explicit.

On the basis of microdata from the Socio Economic Panel (SOEP) for married couples in Germany, we provide empirical evidence about the sensitivity of the welfare orderings to different normative principles embodied in these measures. We retrieve individual and household specific preference heterogeneity, by estimating a structural discrete choice labor supply model. We use this preference information to construct welfare orderings of households according to the different metrics, each embodying different ethical choices concerning the preference heterogeneity in the consumption-leisure space. We then discuss how sensitive the assessment of a hypothetical tax reform is to the choice of metric. The chose tax reform is similar to a subsidy of social security contributions.

Keywords: Welfare measures; labour supply; random utility; preference heterogeneity

JEL: C35 D63 D78 H24 H31 J22

[*] We are grateful to Rolf Aaberge, Koen Decancq, Marc Fleurbaey, Erwin Ooghe, Erik Schokkaert and participants of the seminars "Basic Income Policies and Tax Reforms: Alternative Concepts and Evaluations" at the University of Turin, Nov 19-20 2009, "Fairness, Equality of Opportunity and Public Economics" at CORE, Louvain-la-Neuve, April 9-10 2010 and the EUROMOD workshop at ISER (University of Essex) September 23-24 2010, for useful comments on the results presented in this paper. The usual disclaimer applies.

[†] Department of Economics KULeuven Belgium. E-mail: `andre.decoster@econ.kuleuven.be`

[‡] Research Associate at DIW-Berlin and Goethe Universität Frankfurt. E-mail: `phaan@diw.de`.

1 Introduction

The last fifteen years, substantial progress has been made in modelling individual labour supply. Random utility models based on the structural specification of preferences have become standard. Their feasibility to account for complicated real world budget constraints, and their ease of interpretation make them especially attractive for the ex ante evaluation of policy reforms in the tax benefit sphere. For an overview, see Creedy and Kalb (2005). However, it is somewhat surprising that this proliferation has not been matched by comparable progress or interest as far as the normative interpretation and/or use of the positive results of these models is concerned.

Of course, the normative analysis stricto sensu, exemplified by e.g. the optimal tax literature, has taken the progress in labour supply modelling on board (see among others Saez 2001 and 2002; Choné and Laroque 2005, 2009, Aaberge and Colombino 2008 and Blundell and Shephard 2009). In applied work however, many users of the models either completely eschew normative interpretations, or report conventional measures of 'welfare' which are not necessarily consistent with the underlying model. Indeed, many applied papers report only aggregate labour supply changes or, when unable to avoid distributional analysis, present changes in labour supply and/or changes in disposable income for deciles of the gross wage distribution. There is, of course, nothing wrong in neglecting leisure and focussing on disposable income (or consumption) alone when constructing an individual welfare measure.[1] The impression prevails however, that the predominant use of disposable income as a welfare measure in applied work, is more inspired by relative neglect than based on a conscious and deliberate conceptual and normative choice. The aim of this paper is to provide empirical evidence that the choice of the normative framework within which policy reforms, affecting the labour-leisure choice, are evaluated, strongly affects the welfare analysis of the reform.

Of course, many papers do recognise the need to account for leisure in the normative step of the analysis. In classical applied welfare analysis, individual welfare metrics such as equivalent or compensating variations, are known well enough.[2] In a context of individuals with heterogeneous preferences however, both the interpretation of these welfare metrics, and certainly their aggregation quickly faces serious difficulties. A

[1] Referring to income based poverty measures as alternative social welfare objectives to the standard utilitarian ones in optimal tax theory, Creedy and Hérault (2009, p.3) call the use of disposable income, be it as an input to a poverty measure, a non welfarist approach:

"*But 'non-welfarist' forms are sometimes used. For example, social welfare may be based solely on an income-based measure of poverty, which can give quite different results. Non-welfarist objectives may go further than simply attaching no value to leisure, in that they may prefer to encourage labour supply (whereas in a welfarist approach the existence of non-workers is acceptable in an optimal structure).*"

[2] In this paper we only deal with welfare metrics at the individual or household level. But the 'rebirth' of money metrics is also prominent in aggregate analyses such as ranking countries by means of alternatives to GDP (see Fleurbaey and Gaulier 2007 or Jones and Klenow 2010 and for an overview Fleurbaey 2009).

criterion derived from a simple aggregation of equivalent or compensating variations has been shown to be neither a sufficient, nor a necessary condition to identify potential Pareto improvements, let alone social improvements according to a well defined social welfare function (see Boadway and Bruce, 1984 Chapter 9 or Auerbach, 1985). And in any case, the use and aggregation of this kind of welfare metrics implicitly introduces comparability assumptions which would preferably have been made on an explicit basis.

To deal with these problems, one can identify two tracks in the relevant literature. The first one simply neglects the comparability and aggregation issues of the classical individual welfare metrics in a context of preference heterogeneity. This is done e.g. by simply sticking to the simple aggregation of compensating variations, and recognizing the problem in an apologizing footnote (Eissa, Kleven and Kreiner, 2008 p.804, footnote 5). This seems to us quite unsatisfactory if one wants to take preference heterogeneity seriously.

More creditworthy is the approach of Aaberge et al. (2004) and Aaberge and Colombino (2008). To simulate labour supply responses to tax reforms, they estimate preferences which are heterogeneous across households. When moving from the positive into the normative step of the analysis however, they follow King (1983) in rejecting the interpersonally incomparable equivalent or compensating variations. They impose comparability by evaluating chosen bundles by means of one fixed preference ordering (the so-called reference household) at reference prices. It is true that these preferences of the reference household are estimated on a sample of individuals or households with heterogeneous preferences, though this does not diminish the fact that in the normative part of the analysis preference heterogeneity itself is removed from the scene.[3] The normative literature on interpersonal comparisons has therefore christened this procedure as 'Perfectionism'. It escapes the clash between, on the one hand, forms of interpersonal comparability (e.g. Pigou-Dalton criteria, or bundle dominance) and, on the other hand, Paretianity, by removing preference heterogeneity and imposing preferences determined by the social planner.

Yet, precisely the research into this clash between interpersonal comparability and Paretianity (or respecting individual preferences) in a context of preference heterogeneity has proven to be fruitful to discover new and complimentary perspectives in designing individual welfare metrics in heterogeneous environments. In a series of papers, Fleurbaey and co-authors show how to construct a normative framework which maximally retains preference heterogeneity, and how individual welfare metrics follow from this analysis. In this paper we demonstrate the usefulness of these individual welfare metrics in the context of empirically estimated heterogeneous preferences. We will also illustrate one of the major advantages of these individual welfare metrics, to wit that they

[3] Also, a sensitivity analysis does not introduce genuine preference heterogeneity into the normative analysis. In each step of the sensitivity analysis, *all* individuals or households are endowed with the same preference ordering.

bring the normative choices, inevitably present in all interpersonal comparisons, clearer to the surface. In this respect, our paper can be read as a complement to Preston and Walker (1999). These authors lined up many of the measures used below, in a list of possible individual welfare metrics taking into account both consumption and leisure. The measures proposed and used in this paper are, therfore, not new at all. What is novel, is that the empirical rank correlations of welfare orderings based on these different measures, can now be interpreted as showing the sensitivity of welfare orderings to ethical choices about how to deal with preference heterogeneity. Moreover, the empirical nature of our paper complements results from similar exercises in Hodler (2009) or Luttens and Ooghe (2007), where the application of a proposed normative analysis in societies with heterogeneous preferences is confined to numerical simulations in highly stylised settings.

In order to provide this empirical evidence, we use microdata from the Socio Economic Panel (SOEP) for married couples in Germany. We retrieve individual and household specific preference heterogeneity, by estimating a structural discrete choice labour supply model as e.g. in Aaberge et al. (1995) or van Soest (1995). We use this preference information to construct welfare orderings of households according to different metrics of welfare, each embodying different ethical choices concerning the preference heterogeneity in the consumption-leisure space. We then move beyond the more descriptive analysis and discuss the different welfare implication of the welfare measures when analysing a hypothetical tax reform, similar to a subsidy of social security contributions.

The rest of the paper is structured as follows. In section 2 we briefly overview the problem of making interpersonal comparisons when preferences differ. We show how well-understood money metrics can help fix the dilemma between respecting preferences and making interpersonal comparisons. We focus on the normative interpretation of these metrics. In Section 3 we present the structural model of labour supply, calculate the welfare metrics, compare the welfare orderings, and discuss the sensitivity of welfare impact of a policy reform with respect to the choice of welfare metric. Section 4 concludes.

2 The welfare metrics and their normative interpretation

2.1 Preference heterogeneity and welfare comparisons

In the following we discuss welfare comparisons when individuals have different preferences. For the exposition of the measures, we focus on single individuals, though the same arguments apply to the household context. Observed bundles of consumption and leisure result from individual choices. Choice is explained by means of preferences and constraints. We define preferences in the (c, l)-space where c stands for consumption (or net income) and l for labour supply, and denote the fact that individual i weakly prefers bundle (c_i, l_i) over bundle (c'_i, l'_i) by the ordering R_i:

$$(c_i, l_i)\, R_i\, (c'_i, l'_i) \Leftrightarrow u_i(c_i, l_i) \geqslant u_i(c'_i, l'_i) \tag{1}$$

where the right-hand of (1) shows the notation by means of the preference representation function $u_i(c_i, l_i)$. Preference heterogeneity, revealed by the subscript i beneath R, plays a major role in this paper. We parameterise preferences as $R_i = R(\mathbf{z}_i)$, where vector $\mathbf{z}_i$ contains observable variables, partly explaining heterogeneity in preferences. The explanatory variables, therefore, appear in the preference representation function $u(c, l; \mathbf{z}_i)$. In the empirical application we will find that this deterministic part of the preferences (captured by observable vector $\mathbf{z}_i$) explains only part of the variation in choices for individuals facing the same constraints. The rest of the variation is due to 'unexplained heterogeneity'. At this stage, however, we do not elaborate the normative treatment of this unobserved heterogeneity. This means that we assume that two individuals with the same vector $\mathbf{z}$ do have the same preferences.

The chosen bundle (c_i, l_i) by individual i is rationalized as a choice of his most preferred bundle, given his choice set:

$$(c_i, l_i) = \arg\max \left[u(c, l; \mathbf{z}_i) \, | c \leq f\left(I_i, w_i l\right), \; l \leq 1 \right], \tag{2}$$

where $f(.)$ is a function representing the tax benefit system, transforming non labour income I_i and labour income $w_i l$, with w_i denoting the gross wage for individual i, into net income c.

In this framework, differences in outcomes for different individuals are explained by differences in preferences (vector $\mathbf{z}_i$), differences in gross wages (scalar w_i), and differences in non labour incomes (scalar I_i). We illustrate a typical configuration for two individuals (the, by now, more or less mythical Ann and Bob), denoted by subscripts a and b in Figure 1, where for simplicity we have assumed away the tax benefit system.

Ann has a lower preference for leisure, in that, compared to Bob, she requires less compensation to work more hours. She also has a higher non labour income than Bob, but a lower wage. The choices made by Ann and Bob are represented by bundles a and b respectively. Ann works more and has a higher net income and less leisure. Bob works less, has more leisure, but a lower net income. The question at hand is: how to compare welfare levels of Ann and Bob? Or: how to choose the (or a) metric $m(c_i, l_i; R_i, w_i, I_i)$ which takes into account both preferences and constraints of individuals, and allows to order individuals from worse to better off?

That this is not an easy task has been well known for decades. The difficulty also formally appeared in the literature in the form of an incompatibility between two sets of axioms (see e.g. Fleurbaey and Trannoy, 2004). On the one hand there is Paretianity, requiring that if all individuals weakly prefer social state x over social state y, the social ordering should also express a preference of x over y. In the following we will refer to this intuition as 'Respecting Preferences' or 'Non Paternalism'. On the other

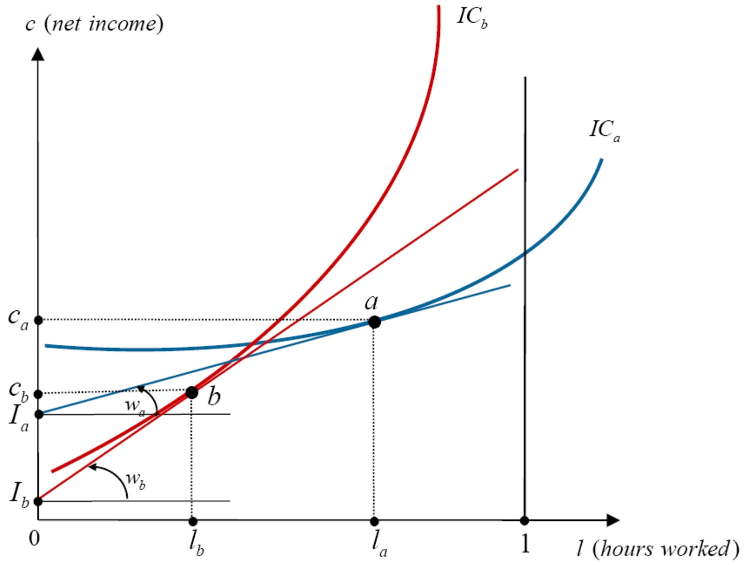

Figure 1: The choice of Ann and Bob in the $(c, l)-$space with different preferences, different unearned incomes and different wages

hand, one has the axioms which embody some form of interpersonal comparability, like Pigou-Dalton (in multidimensional settings), or dominance of bundles. We illustrate the incompatibility with the well-known figure 2, where we use bundle dominance for the sake of illustration.

On the one hand, bundle dominance, used to interpersonally compare Ann and Bob and applied to bundles a and b, allows the social planner to conclude that bundle a for Ann is to be preferred over bundle b for Bob. Respecting the preferences of Ann, who is indifferent between bundles a and a', also allows one to conclude that bundle a' for Ann is to be preferred over bundle b for Bob (Conclusion I). On the other hand, bundle dominance applied to bundles a' and b', leads to a ranking by the social planner of bundle b' for Bob being better than bundle a' for Ann. And since Bob is indifferent between b' and b, we also conclude that bundle b for Bob is to be preferred over bundle a' for Ann (Conclusion II). Obviously Conclusion I and II cannot be simultaneously true, illustrating the clash between axioms which express interpersonal comparability and those that embody respect of preference heterogeneity.

The classical way out of this problem is to put aside the requirement of respecting the heterogeneous preferences. Indeed, it has been well known for decades - although not always honoured in practice - that broadly used concepts from applied welfare economics such as equivalent or compensating variations, are only well defined for a given preference ordering, and for a given price vector. In a context of preference heterogeneity this means that the money metrics are calculated by inserting the chosen bundle

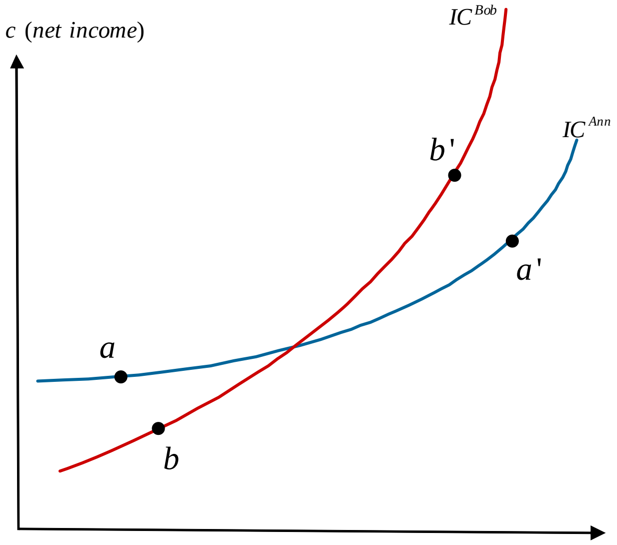

Figure 2: The incompatibility between bundle dominance and Paretianity

into a reference preference ordering, using reference prices (which are the same for all individuals). This is the approach followed by Aaberge, Colombino and Strøm (2004) and Aaberge and Colombino (2008). Normatively spoken, this boils down to impose some kind of objective criterion of welfare, which might be called a 'perfectionist' view of wellbeing. The analyst (or policy maker) introduces interpersonal comparability by fixing the welfare criterion independently from the preferences of the individuals. Otherwise stated, although preferences continue to play their full role in the determination of the chosen bundles, preference heterogeneity is de facto assumed away in the normative phase of the analysis, to wit, in the step where interpersonal comparisons are introduced.

Yet, the incompatibility described above, suggests that there is also another possibility. A recent and rapidly growing strand of the (mainly normative) literature also explores the possibility to give priority to Paretianity and to fully respect preference heterogeneity. The incompatibility result then inevitably points to the necessity of restricting the way one introduces interpersonal comparability. Recent proposals in Fleurbaey (2006, 2008) amount to restrict the interpersonal comparability by means of what the author calls *Subset Dominance.* Interpersonally comparable individual welfare levels are obtained by measuring individual welfare by means of nested sets, B_λ, where the set B_λ is implicitly defined by:

$$u(c_i, l_i; \mathbf{z}_i) = \max \left[u(c, l; \mathbf{z}_i) \,|(c, l) \in B_\lambda \right]. \tag{3}$$

The chosen bundle (c_i, l_i) on a given indifference curve is evaluated by indexing the curves by means of these equivalent sets, where $\lambda \leq \lambda'$ if and only if $B_\lambda \subseteq B_{\lambda'}$ and the situation of individual i is better the higher λ. We illustrate this way to index

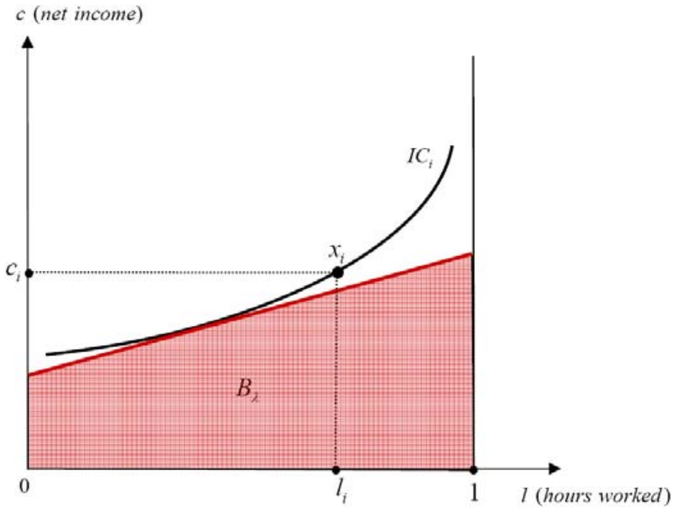

Figure 3: Indexing the indifference curve by means of the equivalent set B_λ

indifference curves in Figure 3.

This specific way to introduce interpersonal comparability not only allows to escape the incompatibility problem. It also brings the implicit normative intuitions, embedded in any interpersonal comparison, clearer to the surface. The normative principles embedded in the choice of the welfare metric and in the way interpersonal comparisons are made, are expressed by means of the choice of the equivalent set B_λ, and we will discuss them in the next section. We first introduce the three different metrics used in our empirical application as specifications of the set B_λ in (3).

The first metric is based on a specification of the equivalent set as:

$$B_{\lambda^{LF}} = \left\{ (c,l) \left| c \leq \lambda^{LF} l, \quad l \leq 1 \right. \right\}, \tag{4}$$

with a corresponding welfare metric for individual i equal to $m_i^{LF} = \lambda^{LF}(c_i, l_i)$. The superscript LF refers to the 'Laissez Faire' description of this choice in Fleurbaey and Maniquet (2006). We illustrate the metric in Figure 4.

The chosen bundles a and b lead to interpersonal comparable welfare levels m_a^{LF} and m_b^{LF} by calculating the slope of the ray through the origin which delineates the subset of the (c,l)-space to which the indifference curve through the chosen point is tangent. In fact this choice of the equivalent set amounts to the real wage criterion of Pencavel (1977), and the real wage metric W_5 in the list of Preston and Walker (1999).

The second class of examples rests on equivalent sets defined by

$$B_{\lambda^{REF}} = \left\{ (c,l) \left| c \leq \lambda^{REF} + \widetilde{w} l, \quad l \leq 1 \right. \right\}. \tag{5}$$

In this case the indifference curves are indexed by means of equivalent sets which depend on a chosen reference net wage $\widetilde{w}$ and an unearned income λ^{REF}, where the

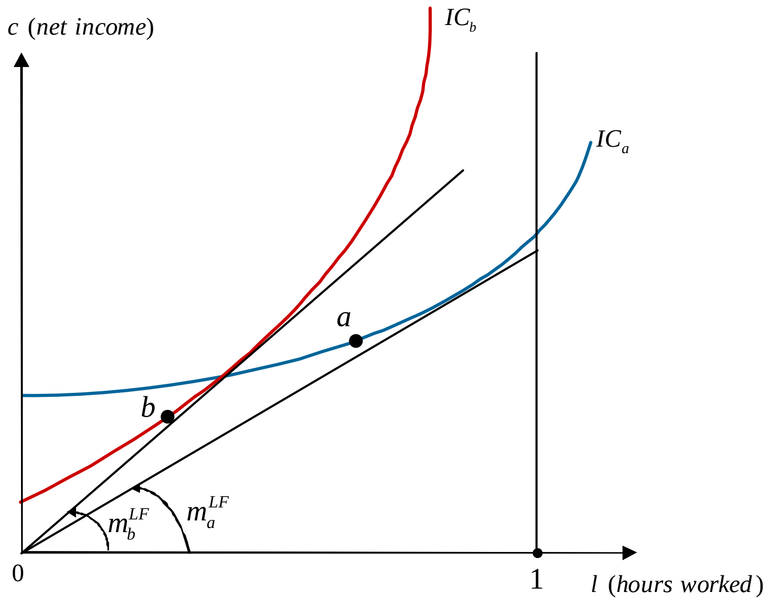

Figure 4: The "Laissez-faire" metric

corresponding individual welfare metric is then chosen to be this unearned income: $m_i^{REF} = \lambda^{REF}(c_i, l_i, \widetilde{w})$. Figures (5) and (6) illustrate the welfare metric for two choices of $\widetilde{w}$, where Figure (6) contains the special case of a reference net wage equal to zero. This specific case of $m_i^{RENT} = \lambda^{REF}(c_i, l_i, \widetilde{w} = 0)$ is called the 'Rente criterion' by Fleurbaey (2006), and coincides with the intercept income of Preston and Walker (1999).[4]

2.2 Normative interpretation of the different metrics

As such, escaping the incompatibility between Paretianity and some form of interpersonal comparison, by giving priority to respecting perferences, is of course not superior to the choice of giving up Paretianity and imposing one specific utility function. Therefore other arguments must be found to choose for the subset dominance approach. A convincing one is given by Fleurbaey (2008) when countering the objection that the choice of reference prices and characteristics in the money metric utility approach is 'arbitrary':

> "if the equivalence approach depends on reference parameters, it can avoid arbitrariness if it develops an ethical theory of the choice of the ref-

[4] In this case the equivalent set comes close to an interpretation of the imposition of interpersonal comparability in terms of reference bundles (as in Schokkaert et al. 2009). When the indifference curve is sloping upwards at $l = 0$, the tangency point of the equivalent set for a net wage equal to zero, becomes the corner solution. The Rente criterion, therefore, introduces interpersonal comparability by comparing individuals in the counterfactual situation 'as if they do not work', that is in terms of the reference bundle $(c, 0)$.

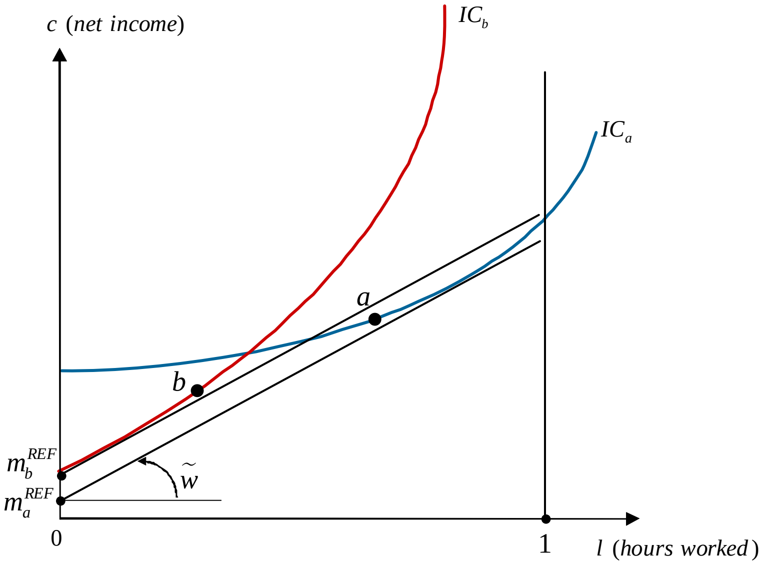

Figure 5: The reference wage metric

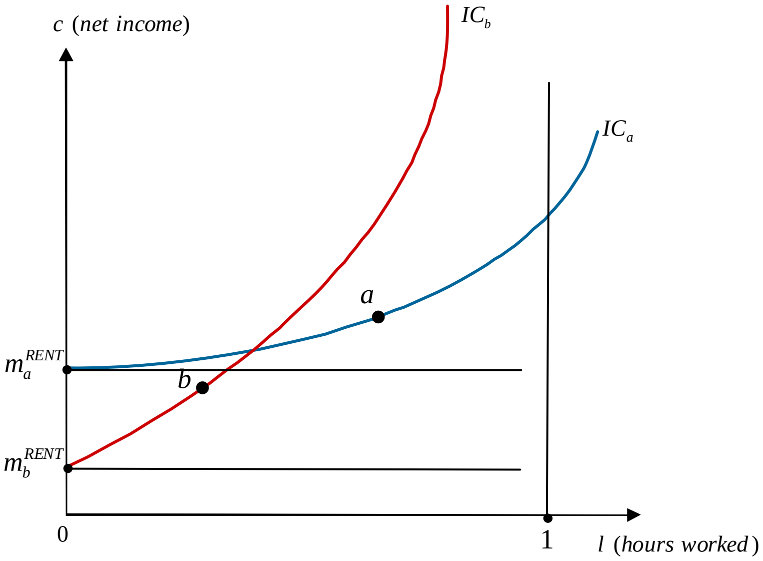

Figure 6: The Rente criterion

> erence. Some examples in the literature on fair social orderings show that rather natural axioms of fairness may force to adopt certain reference parameters". Fleurbaey (2008, p. 10).

Otherwise stated, it might be easier to think about the ethical priors in terms of choosing these equivalent sets, than in terms of a common utility function. Fleurbaey (2005) gives the example of the metric designed to measure welfare in the multidimensional space of income and health. In that case, it seems natural (though not compelling) that one restricts interpersonal comparisons to the subset of the space where all individuals are healthy (instead of in bad health). And Schokkaert et al. (2009) argue that when constructing a measure of job satisfaction along the lines of subset dominance, one can better restrict interpersonal comparability to the subset of space where all individuals have a good job instead of when they have bad jobs. Hence, also the choice of the equivalent sets described in equations (4) and (5) should ultimately be guided by a deliberate choice between different normative principles underlying the different metrics. What are the implicit normative choices in the three individual metrics m_i^{LF}, m_i^{REF}, and m_i^{RENT}?

First note that all three metrics fully respect preferences. That means that all metrics will increase when the individual moves to a bundle on a higher indifference curve of his *own preference ordering.* The difference between the metrics is to be found in the way differences in preferences play a role in the ranking of individuals. Indeed, 'Respecting preferences' does not tell us anything about how to eventually weigh people with different preferences differently. Under the Laissez Faire criterion m^{LF} of Figure 4 e.g. we judge two individuals as equally well off when they have the same hypothetical net wage rate, irrespective of the choices they make. In terms of a responsibility-compensation cut, this criterion holds people fully responsible for differences in their tastes for leisure, and only wants to compensate them for differences in their wages. In the m^{LF}-measure, differences in preferences, leading to different choices, are considered not to be a sufficient reason for redistributing, or for ranking people as worse or better off.

When choosing the Rente-criterion, on the other hand, as is shown in Fleurbaey (2006), we offer maximal protection for people who have a larger distaste for working. With Bob's indifference curve cutting Ann's one from below in Figure 6, we will always judge Bob to be worse off than Ann if they face the same constraint. From this perspective, choosing the Rente criterion as the welfare metric implements a normative choice of holding people with a strong aversion to work minimally responsible for these preferences.

By moving away from the zero reference wage in the Rente criterion to the m^{REF}-metric with a strictly positive reference wage $\widetilde{w}$, it is easy to check graphically that, for a given constellation of preferences (such as the ones of Ann and Bob in Figure 1) the reference wage $\widetilde{w}$ in fact defines the subsets of metrics in the m^{REF}-set which will judge Ann to be better off than Bob (i.e. those metrics using a reference wage below $\widetilde{w}$) and the ones which will judge Ann to be worse off than Bob (i.e. those metrics using a reference wage higher than $\widetilde{w}$). Increasing the reference wage $\widetilde{w}$, therefore, is to be interpreted as moving the redistributive concerns. If we use the reference wage metric m^{REF}, we implicitly use social preferences in which we build in a redistributive bias in favour of distaste for work for all individuals with wages exceeding $\widetilde{w}$ (by ranking them lower), and against apparent laziness for all individuals with wages below $\widetilde{w}$ (by ranking them higher).

The empirical application on which we report in the next two sections, is meant to answer the question how sensitive welfare orderings are with respect to the choice of the metric by means of which individuals are ordered, and hence to the normative choices made by the policy maker concerning preference heterogeneity. More precisely, we derive welfare orderings for the different measures derived above and show how sensitive the answer to the question "who are the poor? who are the rich?" is to the chosen metric. We also investigate the sensitivity of a ranking of gainers and losers of a stylised tax reform, similar to a subsidy of social security contributions which increases the incentives to participate in the labour market.

3 Estimated preference heterogeneity

To apply the above metrics in a real world context we use German microdata from the Socio Economic Panel (SOEP), which contains detailed information about the socio-economic situation of households. The dataset is used as the input dataset for the Microsimulation model STSM (Steiner et al. 2008) which describes in detail the German tax and transfer system. For a given gross wage, STSM allows to determine net income of the household for any chosen amount of labour supply. These detailed real world budget constraints are combined with the observed choices of the individuals in the dataset to estimate a static structural labour supply model. Since the structural character of this labour supply model consists of a specification of the functional form of the preference representation function, this technique allows us to give empirical content to the preference heterogeneity of the previous sections. We first describe the labour supply model and the functional specification chosen for the preferences, then we give some information about the underlying data.

3.1 Specification of household preferences

We estimate household preferences by means of a static structural discrete choice model of labour supply, similar to Aaberge et al. (1995) or van Soest (1995). The model is structural, because it starts from a specification of the utility function. And it is a discrete choice model because it reduces the choices of the individual (in this case the number of hours worked) to a finite number of discrete alternatives. The main advantage of this discrete specification over the continuous framework is the possibility to account for the non-linearities in the budget set and to cope with the endogeneity of net-household income in a relative straightforward way.

The discrete choice model starts from an empirical counterpart of the utility function in (2), by specifying the utility level of household i at a finite number of discrete chosen levels of labour supply. We index the discrete points by means of the subscript $j = 1, ..., J$. The state specific level of utility of household i, denoted v_{ij}, at the $j = 1, ..., J$, discrete states consists of a deterministic and a stochastic part:

$$v_{ij} = u(c_{ij}, (1 - l_{ij}); \mathbf{z}_i) + \epsilon_{ij}, \tag{6}$$

where $u(c_{ij}, (1-l_{ij}); \mathbf{z}_i)$ represents the deterministic part, and ϵ_{ij} is a stochastic random error term which varies independently between the individuals and the discrete points. Preference heterogeneity is captured by vector $\mathbf{z}_i$. Note that we will limit the analysis to observed preference heterogeneity (see below) and hence neglect household specific heterogeneity which is unobserved. We assume that all unobservable effects are captured by the stochastic term ϵ_{ij}.

In this specific empirical application, we focus on the population of married households only. Moreover, we only consider the labour supply decision of the spouse, and assume that labour supply of husbands is exogenously determined.[5] That means that l_{ij} in (6) stands for female labour supply in household i (with $L_{ij} = 1 - l_{ij}$ denoting leisure time of the wife in household i), whereas c_{ij} refers to household net income. The latter consists of labour income of the wife, and puts the exogeneously determined labour income of the husband into non labour income.

Similar to Aaberge et al. (2004) we use a Box-Cox functional form to specify the deterministic part of the utility function in (6):

$$u(c_{ij}, (1 - l_{ij}); \mathbf{z}_i) = \beta_c \frac{c_{ij}^{\alpha_c} - 1}{\alpha_c} + \beta_L(\mathbf{z}_i) \frac{(1 - l_{ij})^{\alpha_L} - 1}{\alpha_L}, \tag{7}$$

where preference heterogeneity is introduced by means of taste-shifters in the following form:

$$\beta_L(\mathbf{z}_i) = \beta_{L0} + \beta'_{L1}\mathbf{z}_i, \tag{8}$$

and vector $\mathbf{z}_i$ includes the age of both spouses, educational dummies, the number and age of children and a regional dummy. Preferences are determined by the parameters β_c, β_{L0}, β'_{L1}, α_c and α_L. The β-parameters determine the marginal utility of consumption and leisure, whereas the α-parameters determine the concavity of the utility function (see the appendix).

[5] We choose to focus on married couples since the economic literature, e.g. Blundell and MaCurdy (1999), has shown that behavioural labour supply responses of married women are particularly important.

The estimation procedure is based on the assumption that the error terms ϵ_{ij} are i.i.d. and follow an extreme value distribution. This gives an expression of the probability for each discrete working alternative, which results in the well known conditional logit framework that can be estimated by maximum likelihood. We want to focus on the calculation of the welfare metrics, and not on the most sophisticated labour supply model, as e.g. in Aaberge et al. (2004) or Blundell and Shephard (2009). Therefore, we make some simplifying assumptions in the estimation procedure. As already announced above, we do not account for unobserved heterogeneity. Haan (2006) has shown that unobserved heterogeneity does not significantly affect the labour supply elasticities when using a similar specification with cross sectional data. Nor do we model potential restrictions on the labour market as in Aaberge et al. (2004) or Bargain et al. (2010). The findings of Bargain et al. (2010) imply that demand side constraints bias elasticities in particular for men and single women, but tend to be less severe for the labour supply decision of married women.

3.2 Data and descriptive statistics

SOEP is a representative household survey for Germany with sufficient socioeconomic information to derive the budget line of a household, i.e., the net household income, and to estimate labour supply behaviour.[6] For this analysis we use the data collected in 2005, with income information about the tax year 2004. We restrict the sample to married households with a wife aged between 20 and 60 who is not self-employed, retired or in full-time education. Moreover we consider only households in which the husband is working full time, i.e., more than 30 hours per week. This gives us a sample of 2076 households. For female labour supply, we define $J = 5$ discrete working alternatives: non-participation, two part time alternatives, full-time work and over-time.[7]

To derive net household income according to the tax legislation in Germany in 2004 at each discrete alternative of working hours, we use the microsimulation model STSM (Steiner et al. 2008). More precisely, for each discrete hours point we calculate gross household earnings as the sum of observed earnings of the husband and the state specific earnings of the wife. Gross earnings of the women are simply the state specific hours multiplied by her expected market wage. For working women we take the observed wage information as their market wage, while for the non-working we impute their expected market wage using an estimated wage equation with selection correction.[8] The information on gross earnings is the key input for the microsimulation model which describes, in detail, all relevant transfer programmes, social security contributions and

[6] For a detailed description of the SOEP, see Haisken-DeNew and Frick (2005).

[7] The median of the empirical distribution in the following intervals define the discrete points: 0, [0 - 15], [16 - 34], [35 - 40], > 40. The estimation results are robust to changes in the approximation of the distribution of working hours.

[8] Estimation results for the wage equation can be obtained by the authors upon request.

income taxation and which delivers the state specific net-household income c_{ij}. Leisure time at each hours point is simply the time endowment $T = 80$ minus working time.

Table 1 shows the overall distribution of the households at the five alternatives. We also show average working hours and average monthly net household income and the shares by region, by education level and by the presence of children younger than 3 years old. The data reveal the relatively low labour market attachment of married women. About 29% of all married women are not working, another 29% works part time and less than a quarter of all married women work regular hours or more. Since in our sample, husbands work at least 30 hours, the income distribution between the 5 discrete states is not very unequal. In addition, this is partly related to the joint taxation with full splitting which leads to high marginal tax rates for the secondary earner.

Table 1: Discrete employment states

	Employment	Share in %	Working Hours per week	Net Income per months	East Germans in %	Education in years	Child younger than 3 years
1	not working	29.06	0	2744	13.07	11.68	27.28
2	0 - 15 hrs	18.00	10	3107	6.33	11.49	10.29
3	16 - 34 hrs	29.01	23	3398	18.99	11.91	4.09
4	35 - 40 hrs	18.33	38	3805	38.60	12.34	2.59
5	>40 hrs	5.60	42	3943	48.31	13.35	3.38

Notes: The sample consists of 2076 married households where the husband is working at least 30 hours. The second column gives median working hours for the intervals 0, [0 - 15], [16 - 34], [35 - 40], > 40, and this median is used to define the discrete employment states.
The share of East German households in the population is 20%, 11% of all women are low educated, i.e. 9 years of school or less, and 11.5% of all households have a child younger 3 years.

Source: SOEP, wave 2005 and STSM

Table 1 shows interesting differences in the distribution across the employment states by region, education, and family composition. In our sample roughly 20 % of all households live in East Germany, but we only find 13% East Germans amongst the non-working women, and even less among part time work. On the other hand the share of East Germans in the subset of households where the wife is working fulltime is close to 40%. For over time work the overrepresentation of East-Germans is even larger. By education we find that women who work more hours tend to have more years of education. The opposite holds for the family composition. Close to 30% of non-working women have a child younger than three years, as apposed to only 3% of those working full time or more hours.

3.3 Estimation Results

Table 2 presents the estimated parameters of the Box-Cox utility function in (7).

Table 2: Estimated parameters of Box-Cox utility function

	Coefficient	Standard Error
Preferences for Consumption		
β_c	3.47	0.59
α_c	0.20	0.14
Preferences for Leisure		
β_{L0}	0.64	0.27
β'_{L1} (taste shifter dummies)		
Age of wife	1.79	0.95
Age of husband	-1.02	0.86
Child younger 3	1.75	0.41
Child between 4 and 6	0.95	0.23
East Germany	-0.64	0.15
Low Education	0.40	0.15
Medium Education	0.28	0.10
α_L	-1.82	0.33

Notes: α_c and α_L determine the concavity of the utility function with respect to consumption and leisure. β_c and β_L determine the marginal utility of consumption and leisure.

Source: SOEP; Number of observations: 2076

Parameters α_c and α_L, both smaller than 1, indicate that the utility function is concave with respect to consumption and leisure time. For consumption, the curvature comes close to a logarithmic functional form (which would be the case if $\alpha_c = 0$) and the concavity is more pronounced for leisure. As expected, households value consumption positively ($\beta_c = 3.47$ being positive) and - on average - women also value leisure time positively ($\beta_{L0} = 0.64$). However, we find significant preference heterogeneity by observable characteristics. In line with previous studies we find that the taste for leisure increases with the presence of children, in particular for children younger than 3 years. We find positive effects of the educational dummies, where the reference category is high education. This implies that *ceteris paribus* women with low and medium education have a higher preference for leisure than women with the highest educational degree. Finally, we find important differences between women in East and West Germany. In line with the descriptive statistics of table 1, women in West Germany have a significantly lower inclination to work. This different pattern in female employment behaviour has often been analysed and is mainly explained by the different history and socialisation of the two parts of Germany before the reunification.

In table 3 we present the preference heterogeneity by means of the variation in the marginal rates of substitution for different subgroups. For all households in the sample, we calculated the slope of the indifference curve at the same bundle of 40 hours of weekly labour supply, and a net monthly income of 2000 euros. The results are striking. On average the MRS in this bundle is 8.5 euros, though there is large variation. According to the estimated preferences, East German women are willing to work an additional hour for less than half the compensation asked by West German women (3.9 compared to 9.6). The presence of young children increases the distaste for work dramatically. The slope of the indifference curves for lower educated people is steeper than for higher educated ones, and contrary to what one would expect, the preference for work is not lower, but higher for females above 55.

Table 3: Marginal rates of substitution for different groups

	Marginal Rate of Substitution	Standard error
Whole Sample	8.5	5.1
West German household	9.6	4.8
East German household	3.9	3.9
children younger than 3	19.8	3.7
children younger than 6	15.7	5.6
low education	11.0	4.3
medium education	9.8	4.4
high education	7.5	5.3
female younger than 25	12.2	7.8
female between 25 and 55	13.2	6.8
female older than 55	8.4	5.1
Labor Supply Elasticities of 1% increase in gross wages		
Change in Participation Rate (in %)	0.16	
Change in Working Hours (in %)	0.34	

Notes: Marginal rates of substitution were calculated in the bundle $(c, l) = (2000, 40)$.
Labour supply elasticities were obtained by increasing female gross wages by 1%

Source: SOEP; Number of observations: 2076

At the bottom of table 3 we also provide information about the size of the behavioural responses with respect to changes in financial incentives by simulating labour supply elasticities. In particular, we increase female gross wages by 1% and given the estimated parameters, we simulate relative changes in expected average participation rates and the relative change in weekly working hours. The magnitude of the elasticities is very much in line with previous studies and suggests that women only modestly respond to changes in their budget line.

3.4 Empirical welfare metrics

To calculate the welfare metrics defined in section 2 for the preferences estimated in this section, we took 100 random draws from the extreme value distribution of the stochastic component. For each of the draws we determine labour supply behaviour of the female in the household by selecting the discrete choice which gives the highest utility. For each of the draws, we also calculate the corresponding net income and the welfare metric by means of the analytical or numerical procedure described in the appendix. Finally, we calculate expected labour supply, expected net income, and the expected value of the welfare metric by averaging over the 100 draws.

4 Who are the poor? Who are the rich? Who are the gainers? Who are the losers?

We present the sensitivity of the welfare ordering to the chosen normative framework for individual welfare measurement in three stages. First, we compare the ordering of households from worst to best off for each welfare metric in a stylised setting of households who only differ in their preferences. Next, we produce an analogous picture for our real world sample of households, where differences in preferences interact with differences in gross wage rates and non-labour income. Finally, we also investigate the sensitivity of a distribution of gainers and losers of a stylised tax reform for the chosen welfare metrics.

4.1 Results for 24 stylised households

We defined a set of stylised households by fixing the female gross wage at €10 in a household where the husband is working full time (38 hours a week) at a gross wage of €15 per hour. With given gross female wage, and a given non labour income these stylized households only differ in their preferences. The combination of two regional values (E for East and W for West German), the possibility that children younger than 3 are present (K if present, N if not), three levels of education (L for low, M for medium and H for high), and two selected ages (25 and 45) produces 24 typical households.

Figure 7 shows the results of simulating labour supply for the females in these households, and the corresponding monthly net income. All results are in expected values. The preference heterogeneity induces large variations in labour supply behaviour, ranging from about 6 hours a week, to nearly 30 hours a week. All households choose a bundle on the budget constraint, and figure 7 clearly reveals the upward shift of the

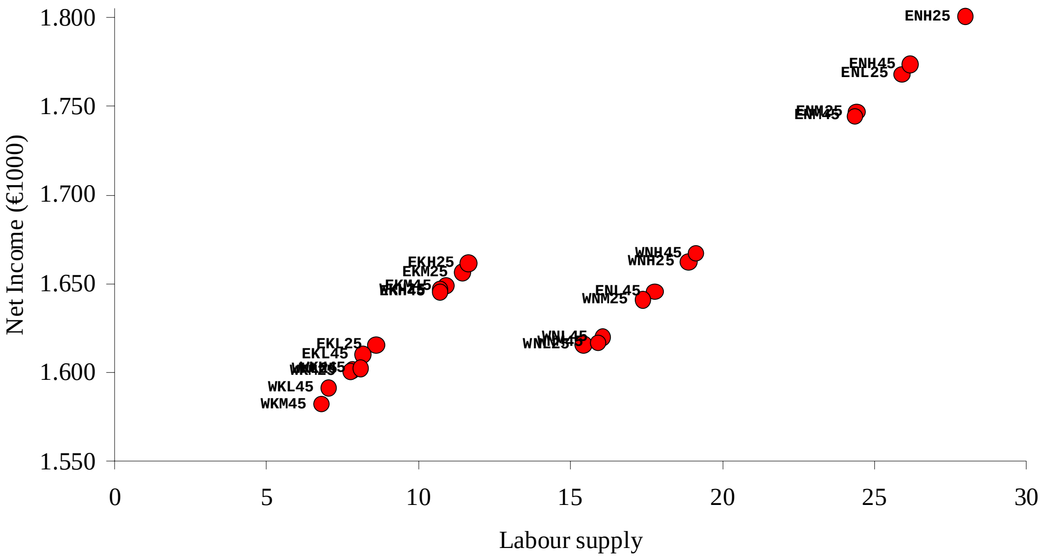

Figure 7: Expected labour supply and net income for 24 stylised households

budget constraint due to the presence of child allowances in the tax benefit system. Besides the effect of young children, the figure mainly illustrates that females in Eastern German households, in general, work more than Western German ones. The whole North-East part of Figure 7 is made up of East German households. Only if they received less education and are older (in this case 45 years old, see label E-N-L-45), they reduce their labour supply.

The different choices in Figure 7 obviously lead to different net incomes for the households. Apart from child allowances, working more also leads to a higher net income of the household, since all households have the same gross wage and the same non labour income. Therefore, the young East German household with no kids and high education who works most (label E-N-H-25) is considered to be the best-off in terms of net income, whereas the older West German household with kids and a middle education level (label W-K-M-45) who supplies the lowest amount of labour is considered to be the worst off in income terms. This is presented in Table 4. For different individual welfare metrics we give the position in the welfare ordering, with "1" indicating the poorest household, and "24" the richest one.

The sensitivity of the answer to the popular and obviously relevant policy question "who are the poor? who are the rich?" to the normative choices underlying the different welfare metrics, is tremendous. Household W-K-M-45 is the poorest in terms of income, but quickly moves up the ladder of the welfare distribution when leisure is taken into account. Moreover, its position heavily depends on how the policy maker or social analysts weighs its preference characteristics relative to households who have preferences

Table 4: Position in the welfare ordering of 24 stylized households

	labour	net		Position in welfare ordering based on				
Household	supply	income	net	Rente	m^{REF} with Reference			Wage
Type	hours/	€/	income	criterion	wage $\widetilde{w}$ =			criterion
	week	month		m^{RENT}	€7	€12	€20	m^{LF}
E-N-L-25	25.9	1768	22	20	24	13	8	7
W-N-L-25	15.4	1616	8	10	1	3	9	10
E-K-L-25	8.6	1616	7	4	2	10	20	16
W-K-L-25	7.8	1602	4	1	3	18	11	21
E-N-L-45	17.8	1646	13	19	20	22	2	8
W-N-L-45	16.0	1620	10	12	6	8	6	14
E-K-L-45	8.2	1610	6	3	5	16	21	17
W-K-L-45	7.0	1592	2	2	9	19	16	24
E-N-M-25	24.4	1747	21	23	21	2	12	4
W-N-M-25	17.4	1641	11	13	13	4	10	6
E-K-M-25	11.4	1657	16	9	10	11	22	13
W-K-M-25	7.8	1601	3	6	4	17	14	20
E-N-M-45	24.3	1745	20	18	17	7	3	11
W-N-M-45	15.9	1617	9	15	8	5	13	12
E-K-M-45	10.9	1649	15	11	7	14	23	15
W-K-M-45	6.8	1582	1	7	15	23	17	22
E-N-H-25	28.0	1801	24	24	22	1	4	3
W-N-H-25	18.9	1662	18	21	23	24	5	2
E-K-H-25	11.6	1662	17	8	16	9	15	23
W-K-H-25	10.7	1647	14	14	12	21	18	18
E-N-H-45	26.2	1774	23	22	14	12	1	1
W-N-H-45	19.1	1667	19	17	19	6	7	5
E-K-H-45	10.7	1645	12	16	18	15	24	9
W-K-H-45	8.1	1602	5	5	11	20	19	19

Notes: the labels of the households consists of four characteristics, West/East, Kids/No kids, Low, Medium or High education, and age of the female in the household.

that are more favourable to supply labour. With the wage criterion e.g., which explicitly ignores differences in net incomes resulting from differences in preferences if gross wages are equal, the same household W-K-M-45 ends up in the third position of the welfare distribution. The reverse holds for the household which is classified as best-off in net income terms (E-N-H-25). With the wage criterion this richest household is considered to be one of the worst-off (with the criterion m^{REF} and a reference wage of €12 it is even the absolutely poorest household). These rerankings in the welfare ordering, based on clearly specified individual welfare metrics for this subset of households who only differ in their preferences, are striking. Preference heterogeneity not only matters in the positive analysis (to predict behaviour as precise as possible), it also matters in the normative phase of the analysis. Once the policy maker has chosen to respect preferences, he also has to make his weighing of differences in preferences explicit. Not unexpectedly, the degree to which he holds people responsible for their distaste for work dramatically determines the welfare ordering.

4.2 Welfare metrics for the population

The results of the previous subsection are exacerbated if, besides preference heterogeneity, we also introduce differences in gross wages and non labour incomes. This is illustrated in Figure 8 which compares the welfare orderings for the different welfare metrics. More precisely, for each metric we calculate the relative position of each household in the welfare ordering and compare the different rankings by means of a scatter plot. If all individuals are ranked in the same position for two metrics, the scatter is displayed as a diagonal one. We compare all measures with the net income criterion.

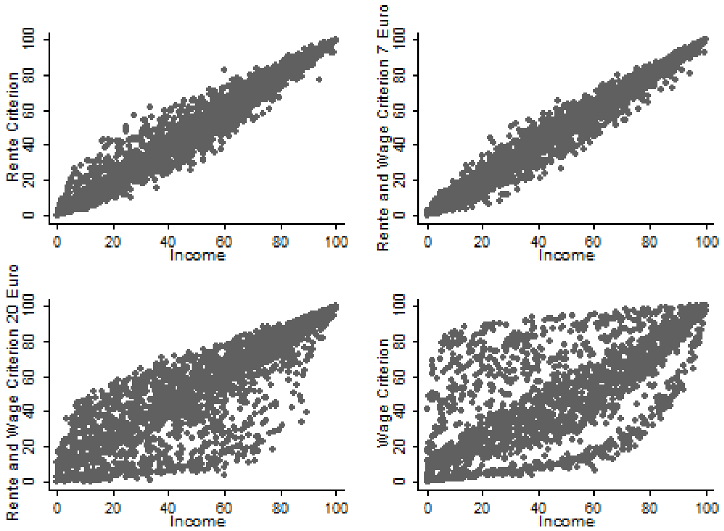

Figure 8: Rank correlation of individual welfare measures

The upper left panel, with the comparison between the Rente Criterion and the pure net income measure, shows that, not surprisingly, taking leisure into account clearly matters. Although there is some concentration on the diagonal, the orderings of the two measures clearly differ, but the introduction of variation in ethical priors about how to weigh differences in preferences is obviously even more important. The m^{REF}-criterion with a reference wage of €7, still correlates quite well with the Rente criterion itself.

Once we move to a wage of €20 however, and certainly to the Wage criterion m^{LF}, the correlation is weak, or even non existent.

The normative significance of this finding is further illustrated in table 5. There we answer the same question "who are the poor?" and "who are the better-off?" by describing the presence of households with certain characteristics in the different quintiles of the welfare distribution based on a given metric. We consider three characteristics which are closely related to preference heterogeneity: living in East Germany, having young children, and being lowly educated.

Table 5: Composition of quintiles of the welfare ordering for the different welfare metrics

Quintiles	Welfare ordering based on					
	net income	Rente criterion m^{RENT}	m^{REF} with Reference wage $\widetilde{w}$ = €7	€12	€20	Wage criterion m^{LF}
	Share of East German households (20%)					
1	0.31	0.22	0.33	0.47	0.61	0.62
2	0.21	0.18	0.20	0.17	0.16	0.18
3	0.17	0.20	0.17	0.15	0.07	0.14
4	0.17	0.24	0.15	0.12	0.10	0.05
5	0.17	0.19	0.16	0.11	0.08	0.04
	Share of households with low education (11%)					
1	0.23	0.24	0.21	0.17	0.11	0.09
2	0.14	0.14	0.14	0.18	0.20	0.14
3	0.12	0.09	0.12	0.11	0.13	0.19
4	0.05	0.05	0.05	0.07	0.07	0.08
5	0.03	0.02	0.03	0.03	0.03	0.06
	Share of hh's with children younger than 3 (11.5%)					
1	0.22	0.29	0.23	0.12	0.03	0.00
2	0.16	0.12	0.15	0.17	0.11	0.02
3	0.07	0.08	0.09	0.12	0.19	0.04
4	0.09	0.06	0.07	0.10	0.15	0.18
5	0.05	0.03	0.03	0.06	0.10	0.33

The results are striking when reading the table across the different columns. Take the first row, which shows the presence of East Germans in the bottom quintile of the welfare distribution, and remember that about 20% of the sample is living in East-Germany. When the welfare ordering is based on disposable income alone, East Germans are clearly overrepresented in the poorest quintile. They do work more, but seemingly, their gross wages and their non-labour incomes are lower. Moving to the second column (the Rente criterion) is a move toward a criterion which also takes into account leisure. And yet, the harder working East-Germans do not move down the welfare ranking because they work more. The reason is that, under the Rente criterion, they are pushed out of the bottom of the welfare distribution by those individuals who have a more pronounced distaste for working. The Rente criterion offers maximal protection with respect to this preference characteristic, by ordering individuals with a distaste for work, ceteris paribus, lower. Moving further to the right in the first row, across the columns of the table, shows how sharply the share of East Germans increases in the bottom quintile, when changing the ethical priors. When we hold individuals more responsible for their preferences w.r.t. the labour leisure choice, and only consider differences in wage rates a legitimate reason for redistribution, the policy analyst will find that the bottom quintile of the welfare ordering is filled with 62% East Germans, which is three times as large as in the Rente criterion.

The same story holds for the other characteristics. The share of households with a lowly educated female in the bottom quintile, drops from 24% under the Rente criterion, to 9% under the Laissez Faire criterion. And the 29% of the bottom quintile which consists of households with children younger than three disappears completely from the bottom of the distribution. They appear to be predominantly well off (33% of the top quintile) when the policy analyst considers their lower preference for work not as a legitimate reason for redistribution.

The interpretation of these striking changes in the composition of the quintiles of the distribution in table 5 can, of course, be contaminated by correlation between the different characteristics. In table 6 we, therefore, investigate whether the above findings are robust when we control for this correlation. We present results from multivariate regressions of the different welfare metrics on observed characteristics, viz. by region, education, presence of young children and non-labour income.[9]

The Rente Criterion and the different metrics of the m^{REF}-criterion are defined in terms of monthly non-labour income. The Wage criterion m^{LF} is expressed in its monthly full-time equivalent. The coefficients can therefore be interpreted in monetary terms, although a direct comparison of the wage criterion with the other ones requires caution. Overall, the findings of table 5 seem to be robust even after controlling for correlation between the characteristics. We find strong and significant differences in the welfare metrics by observed demographics which can be related to preference heterogeneity. *Ceteris paribus* net income is higher for women in East German households, lower for lowly educated females, and lower for females with young children.[10] When the policy analyst moves to the Rente criterion, these effects are strongly amplified. East German women are judged to be even more better off than when using net income, and lowly educated females and females with young children are considered more worse

[9] Note that for comparability we always use expected rather than observed household income. Expected net income is calculated as net income in the optimal working alternative of the wife averaged over the 100 draws from the extreme value distribution.

[10] The positive effect of the East German dummy on net income follows from the fact that we control for non-labor income (i.e. mainly the income of the husband), which is higher for West German households. A regression without this non-labor income as explanatory variable gives the expected negative sign for the East German dummy on net income.

Table 6: Regression of the different welfare metrics on demographic characteristics

		Welfare ordering based on				
	net	Rente	m^{REF} with Reference			Wage
	income	criterion	wage $\widetilde{w}$ =			criterion
		m^{RENT}	€7	€12	€20	m^{LF}
East Germany	109	339	73	-135	-421	-203
	(22)	(32)	(28)	(28)	(27)	(11)
Low Education	-173	-366	-224	-182	-108	1.5
	(27)	(39)	(35)	(35)	(33)	(13.9)
Child younger 3	-199	-650	-372	-172	142	452
	(29)	(42)	(37)	(37)	(35)	(15)
Child between 3 and 6	-244	-594	-387	-251	-31	182
	(26)	(37)	(33)	(33)	(32)	(13)
Age wife	3.3	-2.6	2.6	6.4	11.5	8.9
	(2.3)	(3.3)	(3.0)	(3.0)	(2.8)	(1.2)
Age husband	5.7	13.0	8.2	5.7	1.9	-0.4
	(2.2)	(3.2)	(2.9)	(2.8)	(2.7)	(1.1)
Non labour income in (1000)	451	508	614	657	722	255
	(8)	(11)	(10)	(10)	(10)	(4)
Constant	194	1200	463	185	-396	-57
	(60)	(87)	(78)	(77)	(74)	(31)

Note: Coefficients are obtained by multivariate regressions of the welfare metric in monetary terms on demographic characteristics. All welfare measures are expressed in Euros/1000 per months. Welfare effects are derived based on the estimated coefficients and draws from the extreme value distributed error terms.

off. The amplification of the welfare differences is erased again when switching to the reference wage criterion with a wage of €7. But, and this is even more striking, even when we control for other observable characteristics, we do find rank reversal. East Germans e.g. are, ceteris paribus, considered worse off when using reference wages of €12 or €20, and also when using the wage criterion. This rightmost column of table 6 suggests that, when measured by the wage criterion welfare is about 200 Euros lower for East Germans, ceteris paribus, whereas they were considered to be 339 euros better off by means of the Rente criterion. The opposite holds for females in households with young children, and the welfare difference between the different measures is even larger. Ceteris paribus a household with young children is considered to be 650 euros worse off with the Rente criterion, but are 450 euros better off with the wage criterion. We find these rank reversals for all characteristics. They are outspoken for the presence of children, but individuals with less education are no longer considered worse off neither, once the policy maker does no longer accept that preference characteristics, leading to a lower willingness to work, are a legitimate reason for redistribution.

Tables 5 and 6 not only illustrate the importance of taking leisure into account in the individual welfare measure. They also point to the importance of clearly specifying and founding the normative choices underlying redistributive activities in a setting where one respects preference heterogeneity.

4.3 Gainers and losers of a stylised reform in work incentives

The previous section demonstrated how sensitive the welfare distribution is to normative principles in a setting which respects preference heterogeneity. However, in practice, policy makers might be more interested in identifying gainers and losers of policy *reforms*, instead of knowing who are the poor and the rich in levels. It is possible that the change in welfare level is less sensitive to the underlying normative choices. To investigate this, we simulated a stylised policy reform similar to a subsidy of social security contributions. In particular we increased gross female wages by 1%. We used the labour supply model to determine the behavioural reaction and calculate the welfare metrics before and after the reform. The relative change in the individual welfare metric was used to rank the population in increasing order of welfare gain. This gain distribution was partitioned into quintiles and table 7 describes the composition of these quintiles in terms of characteristics that were relevant for the preference heterogeneity.

The bottom quintile in table 7 contains the households who have the smallest gain. The top quintile is populated by the households with the largest gains. According to the pure income measure which neglects leisure, East Germans are overrepresented in the highest quintile of gainers (33% of this quintile consists of East Germans). The quintile of (relative) losers of the reform are dominated by lowly educated people, and even more outspoken, by households with young children. These standard results are of course directly related to the labour market participation of these respective groups. The question is whether the identification of gainers and losers is robust with respect to choice of the individual welfare metric.

We therefore move to the right in table 7 to use metrics which take up leisure (and the change therein) in the welfare metric, and fully account for preference heterogeneity between the individuals. The overrepresentation of East Germans among the gainers of the reform further increases to 50% when using the Rente criterion to assess the impact of a gross wage increase, but it drops back to 35% when using the wage criterion m^{LF}. This illustrates the crucial role of the slope of the indifference curves (and hence the preference heterogeneity), not only in the calculation of the welfare level, but also for the welfare *difference*. A given net income change translates in a larger welfare gain (e.g. measured on the vertical axis at $l = 0$), the flatter the indifference curve is. With the Rente criterion e.g. one not only considers people with distaste for work as worse off in levels, one also considers that an increase in labour income is valued less by them. However, when the policy maker discards the low preference for work as a legitimate reason for favourable treatment, the share of East Germans in the top quintile falls back to a much lover percentage (35%). They still gain considerably because they

Table 7: Composition of quintiles of gainers and losers of a change in work incentives

Quintiles	Welfare ordering based on					
	net	Rente	m^{REF} with Reference			Wage
	income	criterion	wage $\widetilde{w}$ =			criterion
		m^{RENT}	€7	€12	€20	m^{LF}
		Share of East German households (20%)				
1	0.07	0.07	0.06	0.07	0.07	0.06
2	0.10	0.07	0.07	0.08	0.08	0.13
3	0.21	0.14	0.13	0.13	0.15	0.20
4	0.31	0.25	0.25	0.25	0.24	0.28
5	0.33	0.49	0.51	0.50	0.48	0.35
		Share of households with low education (11%)				
1	0.19	0.20	0.19	0.19	0.18	0.17
2	0.15	0.16	0.20	0.17	0.16	0.15
3	0.10	0.15	0.13	0.14	0.14	0.13
4	0.07	0.04	0.03	0.06	0.07	0.08
5	0.04	0.00	0.01	0.01	0.01	0.03
		Share of hh's with children younger than 3 (11.5%)				
1	0.40	0.45	0.44	0.44	0.37	0.26
2	0.07	0.08	0.10	0.10	0.14	0.19
3	0.06	0.03	0.02	0.03	0.04	0.07
4	0.04	0.02	0.02	0.01	0.02	0.03
5	0.02	0.00	0.00	0.00	0.00	0.03

Note: We consider a tax reform consisting of 1% increase in gross wages. Expected Welfare effects are derived from simulated labour supply behaviour under 100 draws from the extreme value distributed error terms.

Source: SOEP, wave 2005.

work a lot and hence capture the wage increase, but compared to the Rente criterion, a hard working person is no longer treated unfavourably, ceteris paribus, as compared to someone who works less.

The choice of metric also has an outspoken effect on where we classify the families with young children: the share in the lowest quintile varies between 45% and 26% when switching from the Rente to the Wage criterion. For education the effect is especially striking in the top quintile of gainers. Lowly educated households form 4% of the top quintile of the gainers distribution when using the income criterion, but they are all re-allocated to a relatively more losing position when taking leisure into account, and not holding them responsible for their preference characteristics (the Rente criterion).

5 Conclusion

Besides differences in budget sets, heterogeneity in preferences plays a crucial part in explanatory models of labour supply. But the incompatibility between the respect for heterogeneous preferences (as e.g. expressed in Paretianity of the social ordering) and interpersonal comparability, has confined applied welfare analysis to the case of comparability by means of a reference household or individual. Sensitivity analysis of the robustness of empirical results with respect to the choice of the reference household suggests that the choice of this reference preference is not very important (Aaberge et al. 2004).

Introducing a reference preference ordering is however, only one way to escape the impossibility result. In this paper we have followed a different route in the normative part of the analysis by calculating welfare metrics which fully respect preference heterogeneity but restrict the scope of interpersonal comparisons. We applied some of the measures developed in Fleurbaey (2006) and highlighted their different underlying normative priors in the empirical context of an estimated labour supply model. These by now standard discrete choice models of labour supply reveal considerable preference heterogeneity and hence are excellent candidates to illustrate the normative issues at hand. In this paper we explored how this positive information could be fed into the newly proposed metrics, and shed light on the empirical relevance of the choice to respect preference heterogeneity.

The results of the comparison of welfare orderings based on different metrics are striking. Not the inclusion of leisure into the welfare metric plays the decisive role, but the different normative treatment of the preference heterogeneity with respect to the labour-leisure choice. This indicates that the above mentioned robustness of results with respect to the choice of the reference household might have to do more with the removal of preference heterogeneity than with a robustness as such. The illustrative results have severe consequences for any policy advice which wants to incorporate distributional analyses against the background of preference heterogeneity (and respecting it). The answer to the question "who is worst off" and "who is best off" inevitably has to face the question whether one treats people with different preferences differently. Does one consider preference characteristics as legitimate sources for compensation or not? If the answer is affirmative, one might go for a normative analysis based on, what is called in this paper, the Rente criterion. In that case, the difference between welfare ordering based on disposable income and a metric which includes leisure is not very important. If, however, one only considers differences in the budget constraints, as legitimate reasons for redistribution, one has to choose for the wage criterion. The correlation between the ordering based on disposable income and this wage criterion is very weak.

6 Appendix: Recipes used to calculate the money metrics

6.1 The Box-Cox utility function and the budget constraint

The deterministic part of the utility function, with net income c and labour l as endogeneous variables reads as (see (7)):

$$u(c,l) = \beta_c \left[\frac{c^{\alpha_c}-1}{\alpha_c}\right] + \beta_L \left[\frac{(1-l)^{\alpha_L}-1}{\alpha_L}\right], \tag{9}$$

where we have omitted the subscripts i and j used in the text to refer to the household and the chosen discrete point. The available time endowment is normalised at 1 and leisure equals $L = 1 - l$.

To graph the indifference curves, we solve c for a given $\overline{u}$ and varying labour supply l in (9) :

$$c = f(\overline{u}, l) = \left[\frac{\alpha_c}{\beta_c}\left[\overline{u} - \beta_L \frac{(1-l)^{\alpha_L}-1}{\alpha_L}\right] + 1\right]^{\frac{1}{\alpha_c}} \tag{10}$$

The budget constraint follows from the tax benefit system, determining net income c from gross income wl, non labour income I and other characteristics:

$$c = n(I, wl; \mathbf{z}_i). \tag{11}$$

This non linear budget constraint (11) can be linearised by determining virtual non labour income μ for a virtual net wage ω (e.g. corresponding to the $MRS_{c,l}$ in the observed choice (c,l)-see below):

$$\mu = c - \omega.l. \tag{12}$$

6.2 First Partial derivatives

$$\frac{\partial u}{\partial c} = u_c = \beta_c.c^{\alpha_c - 1} \tag{13}$$

$$\frac{\partial u}{\partial L} = u_L = \beta_L.L^{\alpha_L - 1} \tag{14}$$

$$\frac{\partial u}{\partial l} = u_l = -u_L = -\beta_L(1-l)^{\alpha_L - 1} \tag{15}$$

Hence marginal utility is positive for resp. consumption and leisure if $\beta_c > 0$ and $\beta_L > 0$. The latter guarantees that labour has disutility.

6.3 Second derivatives

$$\frac{\partial}{\partial c}u_c = u_{cc} = \beta_c(\alpha_c - 1).c^{\alpha_c - 2} \tag{16}$$

$$\frac{\partial}{\partial l}u_l = u_{ll} = -\frac{\partial}{\partial L}u_l = -\frac{\partial}{\partial L}(-u_L) = u_{LL} = \beta_L(\alpha_L - 1).c^{\alpha_L - 2} \tag{17}$$

We have decreasing marginal utilities ($u_{cc} < 0$ and $u_{LL} < 0$) for both consumption and leisure if resp. $\alpha_c < 1$ and $\alpha_L < 1$. Cross-effects are zero. Note that the change in the marginal disutility of labour is also negative. Hence, the marginal utility of labour is negative, and becomes more negative the more we work.

6.4 The Marginal Rate of Substitution

6.4.1 Between c and leisure L

$$\begin{aligned} 0 = du &= u_c \; dc + u_L \; dL \\ \Leftrightarrow \frac{dc}{dL} &= MRS_{c,L} = -\frac{u_L}{u_c} = -\frac{\beta_L . L^{\alpha_L - 1}}{\beta_c . c^{\alpha_c - 1}} \end{aligned} \tag{18}$$

Since we have $\beta_c > 0$ and $\beta_L > 0$ if the marginal utilities are positive, the slope of the $c - L-$indifference curves will be negative.

To make the slope become less negative as L increases, the absolute value of the $MRS_{c,L}$ should decrease:

$$\frac{\partial}{\partial L} \left| MRS_{c,L} \right| = \frac{\partial}{\partial L} \left[\beta_L L^{\alpha_L - 1} \beta_c^{-1} c^{-(\alpha_c - 1)} \right],$$

where c is itself a function of L to stay on the indifference curve. Hence:

$$\begin{aligned} \frac{\partial}{\partial L} \left| MRS_{c,L} \right| &= \frac{\beta_L}{\beta_c} \left[\frac{(\alpha_L - 1) L^{\alpha_L - 2}}{c^{\alpha_c - 1}} - L^{\alpha_L - 1} (\alpha_c - 1) c^{-\alpha_c} \frac{dc}{dL} \right] \\ &= \frac{\beta_L}{\beta_c} \left[\frac{(\alpha_L - 1) L^{\alpha_L - 2}}{c^{\alpha_c - 1}} - \frac{(\alpha_c - 1) L^{\alpha_L - 1}}{c^{\alpha_c}} \left| MRS_{c,L} \right| \right], \end{aligned}$$

which can be signed as negative when $\alpha_c < 1$ and $\alpha_L < 1$.

6.4.2 Between c and labour supply l

$$\frac{dc}{dl} = MRS_{c,l} = -\frac{u_l}{u_c} = \frac{u_L}{u_c} = \frac{\beta_L}{\beta_c} \frac{(1 - l)^{\alpha_L - 1}}{c^{\alpha_c - 1}} \tag{19}$$

which is positive and increasing (the compensation needed to work more and more is increasing).

6.4.3 Virtual non labour income

Using the $MRS_{c,l}$ in the observed choice (c^0, l^0) we can determine virtual non labour income for a linearised budget constraint with net wage equal to the $MRS_{c,l}$ in the observed point. Substitute $MRS_{c,l}$ for w in (12) and solve for virtual non labour income μ:

$$\mu^0 = c^0 - MRS_{c^0, l^0} . l^0. \tag{20}$$

6.5 Calculation of the welfare measures

6.5.1 The Rente-criterion m^{RENT}

For reference wage $\widetilde{w} = 0$, labour income $= 0$. We calculate virtual non labour income such that $u^0 = u(c^0, l^0)$, given by (9), is reached in a c, l-combination where $MRS_{c,l} = 0$. We assume that the indifference curve has a positive slope at $l = 0$. Excluding negative labour supply, we end up with a corner solution at the intersection of the IC with the vertical net income axis. Hence, we calculate from (10):

$$\begin{aligned} m^{RENT}(c^0, l^0) &= f(u(c^0, l^0), 0) \\ &= \left[\frac{\alpha_c}{\beta_c}\left[u^0 - \beta_L \frac{1^{\alpha_L} - 1}{\alpha_L}\right] + 1\right]^{\frac{1}{\alpha_c}} \\ &= \left[\frac{\alpha_c}{\beta_c} u^0 + 1\right]^{\frac{1}{\alpha_c}} . \end{aligned}$$

6.5.2 The Rente+minimum wage criterion m^{REF}

For this measure we choose a reference wage $\widetilde{w}$, and look for the c, l-combination where the $MRS_{c,l}$ equals this reference wage. Hence from (19) we have:

$$\widetilde{w} = \frac{\beta_L}{\beta_c} \frac{(1-l)^{\alpha_L - 1}}{c^{\alpha_c - 1}},$$

which can be solved for c as:

$$c(l; \widetilde{w}) = \left[\frac{1}{\widetilde{w}} \frac{\beta_L}{\beta_c} (1-l)^{\alpha_L - 1}\right]^{\frac{1}{\alpha_c - 1}}, \tag{21}$$

giving all c, l-combinations satisfying $MRS_{c,l} = \widetilde{w}$. The one combination on the initial indifference curve $u^0 = u(c^0, l^0)$ is found by substituting (21) in the utility function (9):

$$u^0 = u(c^0, l^0) = \left(\frac{\beta_c}{\alpha_c}\right) \left\{ \left[\frac{1}{\widetilde{w}} \frac{\beta_L}{\beta_c} (1-l)^{\alpha_L - 1}\right]^{\frac{\alpha_c}{\alpha_c - 1}} - 1 \right\} + \beta_L \left[\frac{(1-l)^{\alpha_L} - 1}{\alpha_L}\right]. \tag{22}$$

For a given value of u^0, we solve (22) for l numerically by starting at $l = 0$ and assuming that in this point the $MRS_{c,l}$ will be lower than the required reference wage $\widetilde{w}$. We then gradually increase l, calculate the corresponding c, the $MRS_{c,l}$, and compare with $\widetilde{w}$. To sum up:

1. choose $l^{(0)}$ where the superscript between bracket denotes the iteration;

2. determine $c^{(0)}$ to be on the IC with level u^0 with this $l^{(0)}$ by using (10);

3. calculate the $MRS_{c^{(0)}, l^{(0)}}$ in this point $(c^{(0)}, l^{(0)})$ by using (19);

4. compare with $MRS_{c^{(0)}, l^{(0)}}$ with $\widetilde{w}$

- if $MRS_{c^{(0)},l^{(0)}} < \widetilde{w}$, increase labour supply with a small step and go back to step 1;
- if $MRS_{c^{(0)},l^{(0)}} \geq \widetilde{w}$, leave the iterative loop;

Denote the values of net income and labour supply when the loop is left as $(c^{(r)}, l^{(r)})$. Measure m^{REF} is determined by calculating the virtual non labour income from (20)

$$m^{REF}(c^0, l^0) = c^{(r)} - \widetilde{w}.l^{(r)}.$$

We also calculate the utility level $u(c^{(r)}, l^{(r)})$ to check its equality to u^0.

6.5.3 The wage criterion m^{LF}

For the measure m^{LF} we search for the c, l-combination on the indifference curve $u^0 = u(c^0, l^0)$, where the $MRS_{c,l}$ equals the ratio $\frac{c}{l}$ (denoted by c^{LF}, l^{LF}). From (19) we have:

$$\frac{c}{l} = \frac{\beta_L}{\beta_c} \frac{(1-l)^{\alpha_L - 1}}{c^{\alpha_c - 1}} \tag{23}$$

Following the same sequence as for m^{REF}, we first solve for c as a function of l and then substitute this c into the utility function:

$$c.c^{\alpha_c - 1} = \frac{\beta_L}{\beta_c}(1-l).l^{\alpha_L - 1}, \tag{24}$$

but again this is not analytically solvable. As for m^{REF}, we therefore start form a $(c^{(0)}, l^{(0)})$-guess, calculate the $MRS_{c^{(0)},l^{(0)}}$, compare it with the ratio $\frac{c^{(0)}}{l^{(0)}}$ and then adjust the guess. We infer where to move from the intial choice based on the sign of the virtual non labour income μ^0 calculated in (20). If this virtual non labour income is positive we know that

$$MRS_{c^0,l^0} < \frac{c^0}{l^0},$$

and hence the (c, l)-combination where both are equal must be at the right of the chosen point $l^{LF} > l^0$. If the virtual non labour income is negative we have the reverse situation:

$$MRS_{c^0,l^0} > \frac{c^0}{l^0} \Rightarrow l^{LF} < l^0.$$

We therefore start the search iteration from the initial point (c^0, l^0) (all the more because starting at $l = 0$ is numerically infeasible since the ratio c/l is undefined, and starting at $l = 1$ might also lead to an overflow for $MRS_{c,l}$). The iterations then run as follows:

1. Start with $l^{(0)} = l^0$ and $c^{(0)} = c^0$;
2. Calculate the ratio $r^{(0)} = \frac{c^{(0)}}{l^{(0)}}$ and calculate $MRS_{c^{(0)},l^{(0)}}$ from (19);
3. Compare $r^{(0)}$ with $MRS_{c^{(0)},l^{(0)}}$, with $d = MRS_{c^{(0)},l^{(0)}} - r^{(0)}$. This difference is negative for $\mu_0 > 0$ and positive for $\mu_0 < 0$.

4. Fix a variable $l_step = sign(\mu^0) * $[small increment in labour supply].

5. Check the condition $sign(\mu_0) * d < 0$

    - if true, go to step 6
    - if false, leave the iteration.

6. change $l^{(0)}$ with l_step;

7. change $c^{(0)}$ accordingly to stay on the same IC as u_0 using (10);

8. go back to step 2

When the iteration is quit, we have measure $m^{LF} = \frac{c^{LF}}{l^{LF}}$.

References

[1] Aaberge, R. and Colombino, U. (2008), Designing optimal taxes with a microeconometric model of household labour supply, ChilD WP. 06/2008.

[2] Aaberge, R., Colombino, U., and Strøm, S. (2004), Do More Equal Slices Shrink the Cake? An Empirical Investigation of Tax-Transfer Reform Proposals in Italy, *Journal of Population Economics,* 17(4), 767-785.

[3] Aaberge, R., Dagsvik, J., and Strøm, S. (1995), Labour supply responses and welfare effects of tax reforms, *Scandinavian Journal of Economics,* 97, 635-659.

[4] Auerbach, A. (1985) *The Theory of Excess Burden and Optimal Taxation*, in Auerbach, A. and Feldstein, M. (eds.) Handbook of Public Economics, Vol. 1, Elsevier Science Publishers, 61-127.

[5] Boadway R. and and Bruce N. (1984), *Welfare Economics*, Oxford, Basil Blackwell.

[6] Bargain, O., Caliendo, M., Haan, P. and Orsini, K. (2010), 'Making Work Pay' in a rationed labour market, *Journal of Population Economics* 23(1), 323-351.

[7] Blundell, R. and McCurdy T. (1999), *Labor Supply: a Review of Alternative Approaches*, in: Ashenfelter O. and Card D. (eds.), Handbook of Labour Economics, Vol. 3A., Elsevier Science Publishers.

[8] Blundell, R. and Shephard, A. (2009), *Employment, Hours of Work and the Optimal Taxation of Low Income Families*, IFS Working Papers, W08/01.

[9] Choné, P. and Laroque, G. (2005), Optimal incentives for labour force participation, *Journal of Public Economics,* 89, 395-425.

[10] Choné, P. and Laroque, G. (2009), *Optimal taxation in the extensive model*, INSEE-CREST.

[11] Creedy, J. and Kalb, G. (2005), Discrete Hours Labour Supply Modelling: Specification, Estimation and Simulation, *Journal of Economic Surveys* 19(5), 697-734.

[12] Creedy, J. and Hérault, N. (2009), *Optimal Marginal Income Tax Reforms: A Microsimulation Analysis*, Melbourne Institute Working Paper Series No. 23/09.

[13] Eissa N., Kleven H. and Kreiner C. (2008), Evaluation of four tax reforms in the United States: Labor supply and welfare effects for single mothers, *Journal of Public Economics,* 92 (3-4), 795-816.

[14] Fleurbaey, M. (2008), *Fairness, Responsibility, and Welfare*, Oxford University Press.

[15] Fleurbaey, M. (2005), Health, Wealth, and Fairness, *Journal of Public Economic Theory* 7(2), 253-284.

[16] Fleurbaey, M. (2006) *Social welfare, priority to the worst-off and the dimensions of individual well-being*, in: Farina F. and Savaglio E. (eds.) Inequality and economic integration, London, Routledge.

[17] Fleurbaey, M. (2008) Willingness-to-pay and the equivalence approach, Oxford Poverty & Human Development Initiative, OPHI Working Paper No. 25.

[18] Fleurbaey, M. (2009), Beyond GDP: The Quest for a Measure of Social Welfare, *Journal of Economic Literature*, 47(4), 1029-75.

[19] Fleurbaey, M. and Gaulier, G. (2009), International Comparisons of Living Standards by Equivalent Incomes, *Scandinavian Journal of Economics*, 111(3), 597-624.

[20] Fleurbaey, M. and Maniquet, F. (2006), Fair Income Tax, *Review of Economic Studies,* 73(1), 55-83.

[21] Fleurbaey, M. and Trannoy, A. (2003), The Impossibility of a Paretian Egalitarian, Social Choice and Welfare, 21, 243-263.

[22] Haan, P. (2006), Much ado about nothing: conditional logit vs. random coefficient models for estimating labour supply elasticities, *Applied Economics Letters*, 2006, 13, 251–256.

[23] Haisken-DeNew J. and Frick, J. (2005), *Desktop Compendium to The German Socio-Economic Panel Study (SOEP)*, DIW Berlin.

[24] Hodler R. (2009), Redistribution and Inequality in a Heterogeneous Society, *Economica*, 76(304), 704-718.

[25] Jones, C. and Klenow P. (2010), *Beyond GDP? Welfare across Countries and Time*, NBER Working Paper No. 16352.

[26] King, M. (1983), Welfare analysis of tax reforms using household data, *Journal of Public Economics,* 21, 183-214.

[27] Luttens R. and Ooghe E. (2007), Is it Fair to Make Work Pay? *Economica*, 74(296), 599-626.

[28] Preston, I. and Walker, I. (1999), Welfare Measurement in Labour Supply Models with Nonlinear Budget Constraints, *Journal of Population Economics,* 12, 343-361.

[29] Saez, E. (2002), Optimal Income Transfer Programs: Intensive Versus Extensive Labor Supply Responses, *Quarterly Journal of Economics,* 117, 1039-1073.

[30] Saez, E. (2001), Using Elasticities to Derive Optimal Income Tax Rates, *Review of Economic Studies,* 68, 205-229.

[31] Schokkaert, E., Van Ootegem, L and Verhofstad, E. (2009), *Measuring job quality and job satisfaction*, FEB working paper 2009/620

[32] Steiner, V., Wrohlich, K., Haan, P., and Geyer J. (2008), Documentation of the Tax-Benefit Microsimulation Model STSM: Version 2008, *Data Documentation 31*, DIW Berlin.

[33] Van Soest, A. (1995), Structural Models of Family Labor Supply: A Discrete Choice Approach, *Journal of Human Resources*, 30, 63-88.