

Bargain, Olivier et al.

Working Paper

Welfare, labor supply and heterogeneous preferences: Evidence for Europe and the US

EUROMOD Working Paper, No. EM5/11

Provided in Cooperation with:

Institute for Social and Economic Research (ISER), University of Essex

Suggested Citation: Bargain, Olivier et al. (2011) : Welfare, labor supply and heterogeneous preferences: Evidence for Europe and the US, EUROMOD Working Paper, No. EM5/11, University of Essex, Institute for Social and Economic Research (ISER), Colchester

This Version is available at:

<https://hdl.handle.net/10419/64904>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

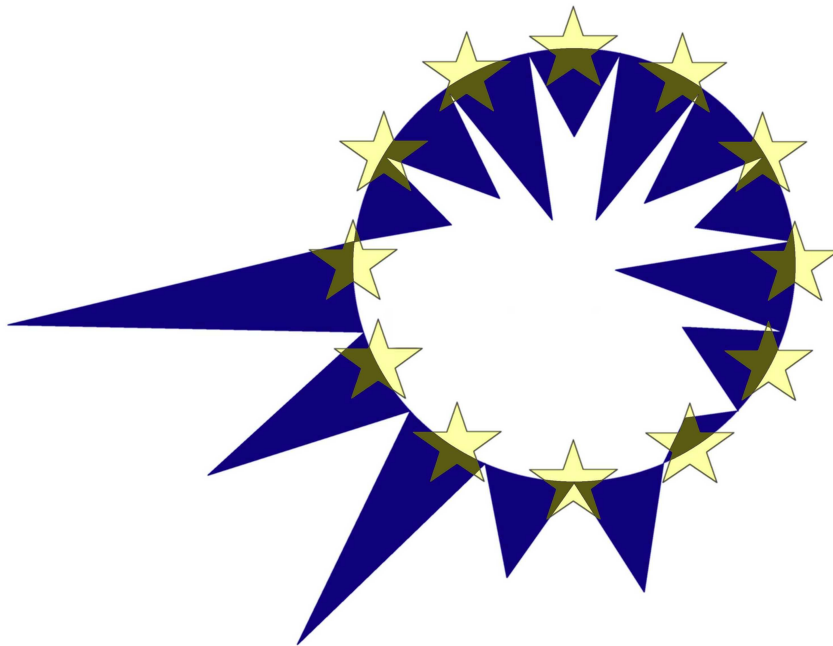
Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

EUROMOD

WORKING PAPER SERIES



EUROMOD Working Paper No. EM5/11

WELFARE, LABOR SUPPLY AND
HETEROGENEOUS PREFERENCES:
EVIDENCE FOR EUROPE AND THE US

Olivier Bargain, André Decoster,
Mathias Dolls, Dirk Neumann, Andreas Peichl
Sebastian Sieglösch

December 2011

Welfare, Labor Supply and Heterogeneous Preferences: Evidence for Europe and the US¹

Olivier Bargain, André Decoster,
Mathias Dolls, Dirk Neumann, Andreas Peichl Sebastian Siegloch²

Abstract

Following the report of the Stiglitz Commission, measuring and comparing well-being across countries has gained renewed interest. Yet, analyses that go beyond income and incorporate non-market dimensions of welfare most often rely on the assumption of identical preferences to avoid the difficulties related to interpersonal comparisons. In this paper, we suggest an international comparison based on individual welfare rankings that fully retain preference heterogeneity. Focusing on the consumption-leisure trade-off, we estimate discrete choice labor supply models using harmonized microdata for 11 European countries and the US. We retrieve preference heterogeneity within and across countries and analyze several welfare criteria which take into account that differences in income are partly due to differences in tastes. The resulting welfare rankings clearly depend on the normative treatment of preference heterogeneity with alternative metrics. We show that these differences can indeed be explained by estimated preference heterogeneity across countries – rather than demographic composition

JEL Classification: C35, D63, H24, H31, J22

Keywords: welfare measures, preference heterogeneity, labor supply, Beyond GDP

Corresponding author:

Dirk Neumann

IZA

P.O. Box 7240

53072 Bonn, Germany

E-mail: dneumann@iza.org

¹ This paper uses EUROMOD version D16. Furthermore we use TAXSIM version v9 which is based on the IPUMS CPS, an integrated set of data of the March Current Population Survey which is conducted jointly by the U.S. Census Bureau and the Bureau of Labor Statistics. EUROMOD is continually being improved and updated and the results presented here represent the best available at the time of writing. Any remaining errors, results produced, interpretations or views presented are the authors' responsibility. This paper uses data from the European Community Household Panel (ECHP) User Data Base made available by Eurostat; the Austrian version of the ECHP made available by Statistik Austria; the Panel Survey on Belgian Households (PSBH) made available by the University of Liège and the University of Antwerp; the Income Distribution Survey (IDS) made available by Statistics Finland; the Enquête sur les Budgets Familiaux (EBF) made available by INSEE; the public use version of the German Socio Economic Panel Study (GSOEP) made available by the German Institute for Economic Research (DIW), Berlin; the Living in Ireland Survey made available by the Economic and Social Research Institute (ESRI); the Socio-Economic Panel Survey (SEP) made available by Statistics Netherlands through the mediation of the Netherlands Organisation for Scientific Research - Scientific Statistical Agency; the Income Distribution Survey made available by Statistics Sweden; and the Family Expenditure Survey (FES), made available by the UK Office for National Statistics (ONS) through the Data Archive. Material from the FES is Crown Copyright and is used by permission. Neither the ONS nor the Data Archive bears any responsibility for the analysis or interpretation of the data reported here. An equivalent disclaimer applies for all other data sources and their respective providers.

² This paper was written as part of research partly conducted during Peichl's visit to ISER via the ECASS programme and supported by the Access to Research Infrastructures action (EU IHP Program) and the Deutsche Forschungsgemeinschaft (PE1675). We are indebted to all past and current members of the EUROMOD consortium for the construction and development of EUROMOD and to Daniel Feenberg and the NBER for access to TAXSIM. The authors would like to thank Bart Capéau, Koen Decancq, Marc Fleurbaey, Erwin Ooghe, Andrew Oswald as well as participants and discussants at seminars and workshops at ISER (Essex), IZA (Bonn), KU Leuven as well as the ISNE 2010 (Dublin), IT 2011 (Canazei), IMA 2011 (Stockholm), New Directions in Welfare 2011 (Paris) and IIPF 2011 (Ann Arbor) conferences for useful comments and suggestions. However, any errors and the views expressed in this paper are the authors' responsibility. In particular, the paper does not represent the views of the institutions to which the authors are affiliated.

1 Introduction

Following the report of the Stiglitz commission (Stiglitz et al., 2009), there has been a recurrent interest in measuring and comparing well-being within and especially across countries (e.g. Jones and Klenow, 2010). One main motivation of the report was to move ‘beyond GDP’ by recognizing the multi-dimensional character of welfare. In addition, recent contributions in the theory of social choice and fair allocation shed new light on how to reasonably measure and consistently compare individual well-being once certain non-market domains are considered besides income (see e.g. Fleurbaey, 2011). In the context of labor supply, the non-market domain is typically leisure time (implicitly including home production). However, while there has been substantial progress in the development of positive models in order to predict labor supply behavior, the development in normative welfare analysis lies somewhat behind. In this paper, we suggest an international comparison based on pure orderings of individual well-being and illustrate for the consumption-leisure space that the choice of how to treat heterogeneity in preferences may substantially affect the evaluation of welfare across different countries.

Precisely, when assessing multi-dimensional welfare, cross-country studies so far commonly rely on representative agent approaches to address the difficulties related to interpersonal comparisons (e.g. Fleurbaey and Gaulier, 2009). In a similar way, fixed reference preferences and prices are often used in labor supply modeling for welfare analyses in the consumption-leisure space (e.g. Aaberge et al., 2004; Aaberge and Colombino, 2008). This implies a certain ethical choice, namely that of the reference agent, while estimated preference heterogeneity is assumed away. We contrast this approach to welfare measures that fully account for different individual consumption-leisure preferences (Fleurbaey, 2006, 2008). A possible normative interpretation of these measures is in the way they differ in holding individuals responsible for their work distaste (taste for leisure).

The empirical application starts with the estimation of labor supply models for 11 European countries and the US. Focusing on married women, the group most studied in the literature, we rely on 12 representative micro-datasets and an harmonized approach for all countries in order to obtain comparable estimates of consumption-leisure preferences. We make use of a common structural discrete choice model, as used in well-known contributions for Europe (e.g. van Soest, 1995) or the US (e.g. Eissa and Hoynes, 2004). We also account for the comprehensive effect of tax-benefit systems on household budgets, which contributes to the identification of the preference parameters. Then, we obtain indifference curves for all individuals of all countries - and take only this ordinal information on well-being to derive an international ranking of individual situations for each of the alternative welfare metrics.

The main results of our analysis go as follows. First, we contrast the standard approaches of using pure income or classic money metric utilities based on a reference household to that of taking preference heterogeneity into account. Second, once heterogeneity in tastes is accounted for, our findings suggest that the resulting ranking of individuals across countries remarkably depends on the normative choice related to the metric at use. Precisely, with a metric that shows little sensitivity to work aversion (holding individuals “minimally” responsible for effort distaste), households from countries where average working hours are relatively high (as in the US and the Nordic countries) perform better on average compared to a ranking based on income only. Inversely, for countries where average working hours are relatively low (as in most Continental European countries, Ireland and the UK), the same holds true with a metric showing high sensitivity to work aversion (holding individuals “maximally” responsible for that distaste). This leads to substantial reranking across nations when moving from the former to the latter type of criteria – with remarkable changes in average individual percentile positions of at least 15 percentage points for 7 out of 12 countries. Third, we decompose marginal rates of substitution (MRS) to extract the role of different sources of heterogeneity for this result. We find that different rankings across welfare metrics are mainly due to heterogeneous work preferences across countries – rather than demographic composition. Thus, the analysis shows very clearly that respecting preference heterogeneity may have substantial influences when comparing well-being in an international context. We believe that these concerns should precede any attempts to compare countries on the basis of social welfare functions (SWF) or other forms of aggregated indices.

The rest of the paper is structured as follows. Section 2 gives an overview of the related literature. In Section 3 we review the welfare criteria and their normative interpretation. Section 4 describes the empirical implementation, including the labor supply model, the data and descriptive information. In Section 5 we present and discuss the main results together with some robustness checks. Section 6 concludes.

2 Related literature

Related to the present paper, several studies have recently attempted to provide international comparisons of welfare levels relying on an equivalent income approach when accounting for non-material aspects of well-being.¹ Becker et al. (2005) correct growth rates for life expectancy (as an indicator for quality of life). Fleurbaey and Gaulier (2009) consider leisure, risk of unemployment, health and household composition besides GDP

¹For a comprehensive overview on general attempts to construct measures of social welfare alternative to GDP, see Fleurbaey (2009). Kassenboehmer and Schmidt (2011) critically assess the additional value of taking into account alternative components to GDP.

in OECD countries. For a large set of 134 countries over time, Jones and Klenow (2010) focus on consumption rather than income when accounting for several other dimensions of well-being. Importantly, all these studies have in common that they compute equivalent incomes at the country level assuming identical preferences across individuals (i.e., relying on a representative agent approach). Aggregation and comparison across countries follows by use of a SWF. However, as already pointed out by Fleurbaey and Gaulier (2009, p. 620), for “*an accurate application of this methodology, one needs survey data on income and on the additional dimensions of consumption [...], as well as on preferences [...], at the individual level and for all the countries studied.*” This is precisely the path we take in the present paper.

As standard in the literature, we retrieve individual and cross-country specific preference heterogeneity relying on a structural discrete choice labor supply model. Naturally, such models respect individual differences in the taste for consumption versus leisure when estimating preference parameters. However, when it comes to welfare analyses, we typically observe that preference heterogeneity is neglected. The main reason is the well-known trade-off between ensuring interpersonal comparability and respecting individual preferences (see e.g. Brun and Tungodden, 2004). In empirical labor supply modeling, two main approaches emerged (besides the simple – but still prominent – use of income as a welfare index). One is to mention, but de facto neglect the comparability and aggregation problems in presence of preference heterogeneity and to report averages of individual – uncomparable – equivalent or compensating variations (e.g. Aaberge et al., 1995, 2000; Dagsvik et al., 2009) or to aggregate them using a certain SWF (e.g. Creedy and Héroult, *forthcoming*; Eissa et al., 2008; Fuest et al., 2008).² In contrast, a second approach explicitly addresses the comparability issue using a reference household for welfare analyses. Following King (1983), classic individual money-metric utilities are derived by means of a fixed preference function at fixed reference prices (e.g. Aaberge et al., 2004; Aaberge and Colombino, 2008; Ericson and Flood, 2009). However, with this approach, preferences of a certain reference household build the basis for comparing individual well-being, which are hence no longer individual specific but unified and determined by the social planner.³

In the present paper we adopt an approach that allows to fully respect individual preferences in welfare analyses (Fleurbaey, 2006, 2008, 2011; Fleurbaey and Maniquet, 2006). Here, the aforementioned dilemma is solved by limiting comparability to subsets

²Indeed, reference prices (wages) for calculations of equivalent and compensating variations are naturally individual and thus, variable. Aggregated indices based on equivalent or compensating variations are therefore inconsistent as long as they are not based on a representative agent approach.

³Then, welfare changes are usually evaluated using a certain SWF over individual money-metric utilities. This generated another stream of criticism, initiated by Blackorby and Donaldson (1988): a SWF over equivalent incomes usually fails to be quasi-concave in commodity consumptions which is incompatible with a minimal preference for equality.

of the consumption-leisure space which are tangent to the individual indifference sets. Choosing certain references in order to compare individual situations (e.g. a specific net wage rate), different metrics can be defined in terms of these subsets. In this way, they will be clearly ordered for different preferences and individual situations are unambiguously comparable. While this is similar to the derivation of classic equivalent incomes, the choice of reference values is grounded on specific fairness considerations. This makes the normative priors of the interpersonal comparison explicit – as, e.g., requested by Atkinson (2011). So far, measures of this kind have not been implemented empirically except in Decoster and Haan (2010) and the present paper. While these authors address preference heterogeneity within a country (Germany), we compute equivalent incomes for individuals of 12 countries and analyze how international rankings vary with the use of alternative welfare metrics. In particular, we focus on the extent to which welfare evaluation is affected by that part of heterogeneous work preferences which is genuinely country-specific.⁴ In addition, we assess the role of different sources of heterogeneity for the resulting differences in welfare rankings.

3 Theoretical framework

In order to respect preference heterogeneity in the consumption-leisure space, we follow Fleurbaey (2006) and look at individual welfare measures which specifically differ in the way they treat heterogeneity in tastes. In the following, we introduce these measures and their underlying normative rationales. We refer to Fleurbaey (2006) for the axiomatic derivation and to Decoster and Haan (2010) for a more detailed illustration. However, while both studies define the metrics in the usual social choice terminology (i.e. as means of budget sets), we “translate” them into the language of classical demand theory (using indirect utility and expenditure functions).

The setup. Assume that individual preferences are defined in the (c, h) -space with consumption c and labor time h . By ordering R_i , individual i weakly prefers bundle (c_i, h_i) over bundle (c'_i, h'_i) , with use of a preference representation function u_i leading to $(c_i, h_i)R_i(c'_i, h'_i) \Leftrightarrow u_i(c_i, h_i) \geq u_i(c'_i, h'_i)$. Observed preference heterogeneity is given by an individual ordering being dependent on an individual specific characteristic vector \mathbf{z}_i , $R_i = R(\mathbf{z}_i)$, and thus $u_i(c_i, h_i) = u_i(c, h, \mathbf{z}_i)$. The chosen bundle (c_i, h_i) results from a classic

⁴This can also be motivated by a prominent debate about what determines differences in labor supply behavior across countries, particularly between Europe and the US. Prescott (2004) states that different labor supply elasticities are almost only due to differences in tax-transfer systems. This view has been criticized by Blanchard (2004) who – in line with Alesina et al. (2005) – argues that different preferences for leisure indeed play a role and are maybe due to cultural differences. Our findings, which control for country-specific consumption-leisure preferences, tend to support the latter view.

individual utility maximization problem, $(c_i, h_i) = \max [u(c, h; \mathbf{z}_i) | c \leq f(I_i, w_i h), h \leq 1]$, with a function $f(\cdot)$ representing the tax-transfer system that transforms gross non-labour income I_i and gross labor income $w_i h$ (with w_i denoting individual i 's gross wage) into net income c (the total amount of time available is normalized to 1). Hence, the observed bundle of consumption and leisure results from individual choices subject to preferences and a budget constraint.

The welfare metrics. Assume the individual's utility function $u_i(c, h, \mathbf{z}_i) = u_i(c_i, h_i)$ to be well-behaved, i.e. continuous and increasing in their arguments as well as quasiconcave (quasiconvex) in $c(h)$. Furthermore, assume tax-transfer rules $f(\cdot)$ determining individual budget sets $c \leq f(I_i, w_i h)$ to be non-linear – as generally observed in reality. Then, for each chosen bundle (c_i, h_i) on a given individual indifference curve IC_i , the budget constraint can be linearized to $c \leq \tilde{w}_i h + \mu_i$ with virtual non-labor income μ_i determined by virtual net wage \tilde{w}_i - as illustrated for bundles a and b in the upper-left panel of Figure 1.⁵ The associated individual indirect utility function is defined as $v_i(\tilde{w}_i, \mu_i) = \max [u_i(c_i, h_i) | c_i - \tilde{w}_i h_i \leq \mu_i]$ and the expenditure function is given by $e_i(u, \tilde{w}_i) = \min [c_i - \tilde{w}_i h_i | u_i(c_i, h_i) \geq u]$ with a fixed level of utility u . In this setting, the different metrics can be formulated by means of *hypothetical, linear* budget constraints.

First, the **“wage”** metric is defined as the slope of the tangent through the origin at the given indifference curve, equaling the wage rate \tilde{w}_i of individual i when the value of the virtual non-labor income is set to a reference value of 0, i.e. $\mu_i = \mu^r = 0$. This is illustrated in the upper-right picture of Figure 1. By use of the metric $\nu_i^W(u, \mu^r)$, individuals can be unambiguously ordered from better to worse off even though preferences differ:

$$\nu_i^W(u, \mu^r = 0) = \min_{\tilde{w}_i} [\tilde{w}_i | v_i(\tilde{w}_i, \mu^r = 0) \geq u] \quad (1)$$

Second, the **“rent + reference wage”** criterion compares individual situations depending on a certain reference value for the virtual net wage, $\tilde{w}_i = w^r$. Then, the resulting welfare metric $\nu_i^{RW}(u, w^r)$ is the value of the corresponding virtual non-labor income (bottom-left panel of Figure 1):

$$\nu_i^{RW}(u, w^r) = e_i(u, \tilde{w}_i) = \min_{c_i, h_i \geq 0} [c_i - w^r h_i | u_i(c_i, h_i) \geq u] \quad (2)$$

Third, the **“rent”** metric directly emerges by setting $\tilde{w}_i = w^r = 0$. As far as we assume well-behaved utility functions, this is equivalent to hours worked being set to a

⁵Note, that apart from this illustration, the virtual budget constraint $c \leq \tilde{w}_i h + \mu_i$ only applies to the logic of the metrics, hence, to *hypothetical* choices of the individuals which might be only by accident consistent with observed choices. The latter are always determined by the budget constraint $c \leq f(I_i, w_i h)$ which follows from the tax-transfer rule.

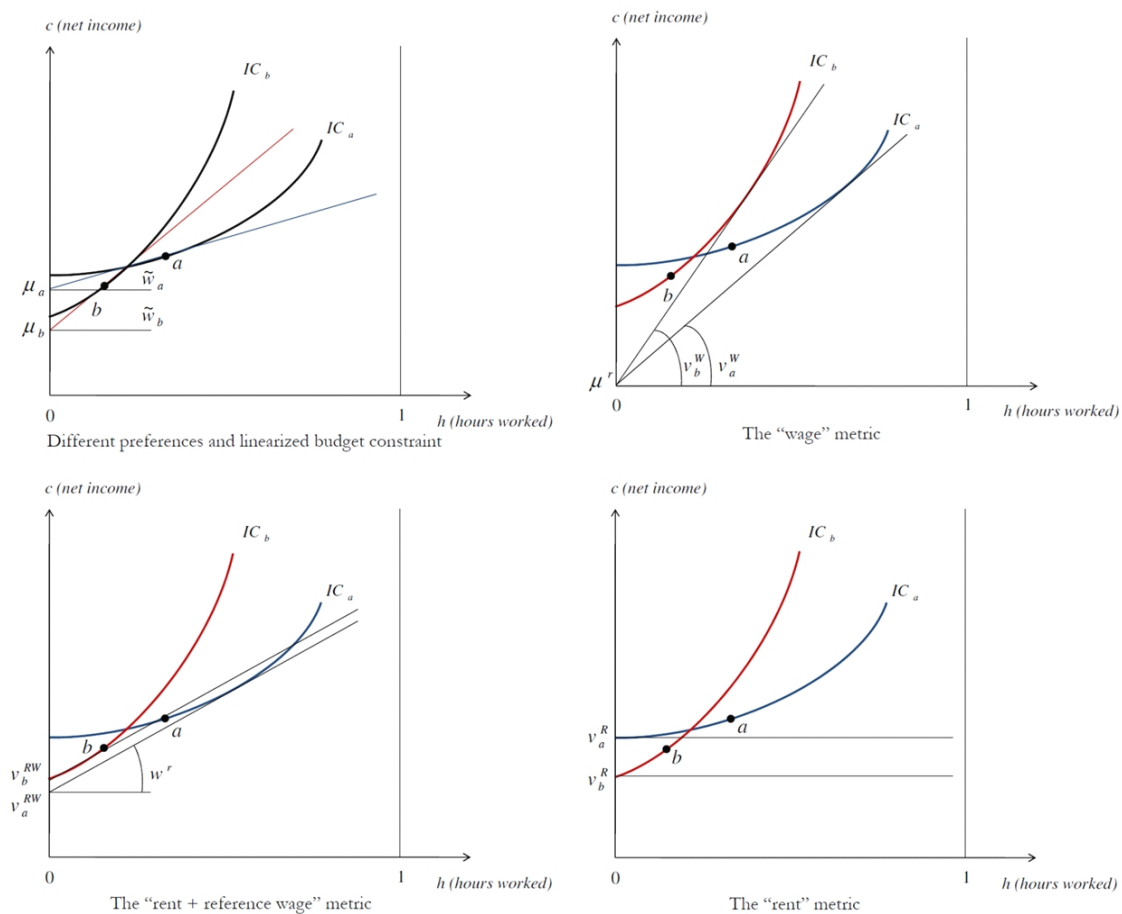


Figure 1: The different welfare metrics graphically

reference value of $h_i = h^r = 0$. The resulting metric $\nu_i^R(u, h^r)$ hence is the value of the intersection of the indifference curve with the ordinate, equaling the corresponding virtual non-labor income (bottom-right panel of Figure 1):

$$\nu_i^R(u, h^r = 0) = \min_{c_i} [c_i | u_i(c_i, h^r = 0) \geq u] \quad (3)$$

In sum, all measures guarantee that individuals with crossing preference profiles can be unambiguously compared in terms of who is better and worse off: hypothetical budget constraints are clearly nested for a given criterion. For the purpose of this paper, we keep the metrics on a purely individual basis and compare individual well-being only grounded on this ordinal information.

Normative interpretation. Generally, allowing for preference heterogeneity creates serious problems for interpersonal comparisons of well-being. It especially raises a question of fairness, i.e., who is to be considered better and worse off (and thus, who should redistribute towards whom). To deal with this, the literature on *responsibility-sensitive*

egalitarianism keeps individuals responsible for their tastes, but not for endowed circumstances (Fleurbaey, 2008). In a consumption-leisure space, tastes are reflected by preferences for work while skill levels (as reflected in gross wages) and non-labor income are assumed to be exogenous to the individuals. The welfare measures defined above evaluate individual situations according to hypothetical reference amounts of external resources (wages, non-labor income) that would allow individuals to reach their current utility level – while individual responsibility characteristics (preferences) are unchanged. When equalizing external resources in this situation, individuals will work and earn at their convenience only. The direction of redistribution, however, might depend on the choice of the reference, i.e. on the specific welfare metric at use.⁶

First, the “wage” criterion gives an answer to the hypothetical question which wage rate would leave the individual indifferent from her current utility level if she had zero (virtual) non-labor income.⁷ As can be seen from the upper-right picture of Figure 1, person *b* with a relatively lower inclination to work is evaluated to be better off compared to person *a* who is less work averse. As a consequence, redistribution (in order to equalize hypothetical wage rates) would be justified from *b* to *a*. Thus, due to the specific reference situation with zero non-labor income, the “wage” metric can be implicitly interpreted as holding individuals maximally responsible for their taste for leisure versus consumption.

Second, the “rent” metric does not ask for a certain wage rate but for the amount of (hypothetical) net income which would be enough to remain equally well off compared to the initial situation if one did no longer have to earn it. The resulting metric is simply the level of consumption when working zero hours, which hence can be seen as the maximal amount of non-labor income theoretically possible given an individual indifference curve. The bottom-right picture of Figure 1 shows, that in this case, we are more likely to judge the apparent lazy as worse off compared to hardworking individuals, and thereby, would grant redistribution towards them. Thus, we might hold people with a strong aversion to work only minimally responsible for that distaste.

Third, the “rent + reference wage” criterion asks which amount of (virtual) non-labor income would make the individual equally well off compared to her actual situation when

⁶Choosing reference values based on certain fairness considerations is the actual novelty of the fair allocation approach compared to classical demand theory when deriving equivalent incomes. See Preston and Walker (1999), for instance, who derive a similar set of metrics in line with the latter. More popular, however, has been the alternative of exploring reference price independent comparisons of individual welfare (see e.g. Roberts, 1980, Slesnick, 1991 and Blackorby et al., 1993).

⁷The underlying fairness criterion corresponds to the “laissez-faire” counterfactual defined in Fleurbaey and Maniquet (2006): in an hypothetical world with equal wage rates and zero non-labor income, differences in consumption-leisure bundles would not call for redistribution as they can only be due to differences in preferences. Hodler (2009) uses this metric to study the effect of redistribution on inequality in a highly stylized setting when a population is heterogeneous in abilities and work-leisure preferences. A variant of the metric is applied in Ooghe and Peichl (2010) to derive optimal taxes when agents only have partial control over certain effort variables. Technically, wage equivalents have been first introduced by Pencavel (1977).

receiving a certain (hypothetical) reference wage equal to w^r . The higher this reference wage is, the worse off individuals with a higher inclination to work will be evaluated. More precisely, in the bottom-left picture of Figure 1, the critical value for the reference wage (i.e., where the evaluation of individuals a and b changes), is given for $w < w^r$, assuming that w^r defines the tangent at the intersection point of IC_b with the ordinate.

4 Empirical approach

Our empirical application benefits from a unique setting whereby household preferences are estimated in an harmonized way for all countries. To do so, we use a common labor supply model, tax-benefit simulations for all countries and comparable datasets, as presented below. We focus on married women. Indeed, the labor supply of single individuals or married men is more often contaminated by demand-side restrictions (Bargain et al., 2010). Married women also show lots of variation in work duration and is the group most studied in the literature (see e.g. Blundell and MaCurdy, 1999). The empirical model is directly compatible with the theoretical framework presented in the previous section. The only difference is that we consider “unitary” households rather than individuals, i.e., couples are assumed to behave as a single decision maker regarding the trade-off between consumption and female labor supply (male labor supply is kept fixed).

Specification of preferences. In order to assess the welfare metrics described in the theoretical part, we must retrieve indifference curves for each household in our sample and, hence, estimate utility functions. To do so, we opt for a structural model of labor supply with discrete choices, which has become quite standard in the literature on tax reforms (see e.g. Aaberge et al., 1995, van Soest, 1995 or Blundell et al., 2000).⁸ This method requires the explicit parameterization of consumption-leisure preferences as it assumes that labor supply decisions can be reduced to choosing among a discrete set of possibilities (e.g., inactivity, part-time and full-time). A common specification over all countries is applied for reasons of comparability. We denote c_{ij} the net income (or consumption, in a static framework) of household i and h_{ij} the wife’s working hours at choice $j = 1, \dots, J$ where the household is assumed to obtain a utility level:

$$V_{ij} = u_i(c_{ij}, (T - h_{ij})) + \epsilon_{ij}, \quad (4)$$

with $(T - h_i)$ the wife’s “leisure time” (which may include time for domestic production), i.e., total time-endowment T minus formal hours of work. For the deterministic part of

⁸Relying on structural models is also the only way to obtain comparable preference estimates across countries. It seems indeed difficult to find natural experiments that would allow performing this task.

the utility function, we rely on a Box-Cox specification, that is:

$$u_i(c_{ij}, (T - h_{ij})) = \beta_c \frac{c_{ij}^{\alpha_c} - 1}{\alpha_c} + \beta_{li} \frac{(T - h_{ij})^{\alpha_l} - 1}{\alpha_l}. \quad (5)$$

This specification is frequently used for welfare assessments (see e.g. Aaberge et al., 1995, 2000, 2004; Aaberge and Colombino, 2008; Blundell and Shephard, *forthcoming*; Decoster and Haan, 2010). Importantly for our purpose, it is easy to check that monotonicity and concavity conditions on consumption and leisure are satisfied (respectively $\beta_c > 0$ and $\beta_{li} > 0$ for monotonicity and $\alpha_c < 1$ and $\alpha_l < 1$ for concavity). Indeed, tangency conditions are necessary for measuring and interpreting the welfare metrics in a straightforward way. We nonetheless check robustness with respect to different specifications in Section 5.4. A crucial point for our analysis is the source of heterogeneity across households. The first obvious difference is that α and β parameters are country-specific, i.e., they are estimated separately for each country. The second source is household-specific heterogeneity through the leisure term, which is specified as follows:

$$\beta_{li} = \beta_{l0} + \beta_{lz} \mathbf{z}_i, \quad (6)$$

with \mathbf{z}_i a vector of taste shifters including the age of both spouses, education of the women, presence of children younger than 3, between 3-6 or 7-12 years old and regional information. The deterministic utility is completed by i.i.d. error terms ϵ_{ij} for each choice, assumed to represent possible observational errors, optimization errors or transitory situations. Under the assumption that error terms follow an extreme value type I (EV-I) distribution, the probability for each household of choosing a given alternative has an explicit logistic form, which is a function of deterministic utilities at all choices. Then, the likelihood of a sample of observed choices can be derived from these probabilities as a function of the preference parameters whose estimates are obtained by maximum likelihood techniques (see van Soest, 1995).

Note that we keep the labor supply model as simple as possible in order to ensure a straightforward implementation and clear interpretation of the welfare metrics. This particularly implies that we do not model potential demand side restrictions on the labor market nor fixed costs of work. This is further discussed in Sections 5.4 and 6.

Data, selection and tax-benefit simulation. For our empirical application, we focus on a selection of 11 European countries and the US. For each country we use microdata based on standard household surveys which provide information on incomes and demographics. For EU countries, we rely on datasets combined with the simulation of national tax-benefit systems for years 1998 or 2001 as described in Bargain et al. (2011). For the

US, we use 2006 IPUMS-CPS (Integrated Public Use Microdata Series; Current Population Survey) data containing information for the year 2005. As mentioned above, we focus on the subpopulation of married couples and estimate the labor supply of the women. To keep the sample relatively homogeneous and avoid too much variation in household’s non-labor income (especially including husbands’ labor income), we select households where husbands at least work 30 hours/week and exclude those with extreme amounts of capital income. Furthermore, we keep households where women are aged between 18 and 59 and available for the labor market, i.e., neither disabled nor retired nor in education. In order to maintain a comparable framework while respecting possible variation in the hours distribution across countries, we adopt a discretization with $J = 7$ hours categories including non-participation, two part-time options, two full-time and two over-time categories (0 to 60 hours/week with a step of 10 hours). Net income at each discrete choice $j = 1, \dots, J$ is calculated as a function $c_{ij} = f(w_i h_{ij}, I_i, \mathbf{x}_i)$ of female earnings $w_i h_{ij}$ and household non-labor income I_i (i.e., household capital income and husbands’ earnings). Female wages w_i are predicted for all observations using calculated wage rates of the workers and the usual correction for selection bias. The function $f(\cdot)$ represents how gross income is transformed into net income, i.e., the impact of taxes and benefits which also depends on certain household demographic characteristics \mathbf{x}_i .⁹ It is calculated numerically using microsimulation models EUROMOD for EU countries and the NBER’s TAXSIM for the US.¹⁰

Empirical welfare metrics. We empirically compute welfare measures for each household in the sample. Given the nonlinear nature of the labor supply model, we must calculate *expected* values. We do so by taking 300 draws from the EV-I distribution of the error terms and searching for the individual optimal choice for each draw. Then, averaged maximum utility \bar{u} is used to derive formulas for individual indifference curves $IC_{\bar{u}}$. Equivalent incomes are computed as follows. For the “rent” metric, an analytical solution is simply obtained by setting h to zero into the formula for $IC_{\bar{u}}$ and retrieving the corresponding level of consumption (hence, the intersection level of $IC_{\bar{u}}$ with the ordi-

⁹Using predicted wages for all observations helps to reduce some of the bias due to measurement errors on wages if calculated on basis of yearly income information (division bias). Also, accounting fully for existing tax-benefit rules completes the identification. Indeed, individuals face different effective tax-benefit schedules because of their different circumstances and socio-demographic characteristics (e.g. age, family compositions, region or levels of non-labor income). This creates variation in net wages between people with the same gross wage. Using nonlinearities and discontinuities generated by the tax-benefit system in this way is a frequent identification strategy in the empirical literature based on static discrete models and cross-sectional data (e.g. van Soest, 1995, Blundell et al., 2000). See Bargain et al. (2011) for a more thorough discussion on this point.

¹⁰For an introduction to EUROMOD, descriptive information of taxes and transfers in the EU and robustness checks for tax-benefit calculation, see Sutherland (2007). An introduction to TAXSIM is provided by Feenberg and Coutts (1993). Both calculators have been already used in several empirical studies (see e.g. Immervoll et al., 2007 for EUROMOD or Eissa et al., 2008 for TAXSIM).

nate). For the “rent + reference wage” metric, we numerically search for the point (c, h) on $IC_{\bar{u}}$ for which the $MRS_{c,h}$ equals the reference wage w^r . The virtual non-labor income corresponding to this tangent is the value for the metric. Finally, the “wage” metric is numerically searched as the slope of $IC_{\bar{u}}$ for which the $MRS_{c,h}$ equals $\frac{c}{h}$. For the two last metrics, we rely on a precise iterative procedure by incrementing hours using very small steps (0.01 hours/week). Note that this is different from moving across discrete categories $j = 1, \dots, J$ as used for the labor supply estimation.

Descriptive information. In Table 1, we present summary statistics for the sample under analysis. The first two columns show the average weekly household net and non-labor income by countries (recall that household non-labor income essentially includes husband’s earnings). Next, female average wage rates, weekly working hours as well as participation rates are presented. Income and wages are de-/inflated to the reference year 2001 and transferred into comparable Purchasing Power Parities (PPP)-USD.

Table 1: Income and employment statistics

<i>Country</i>	Net income per week (1)	Non-labor income per week (2)	Female wages per hour (3)	Female hours per week (4)	Female participation rates (5)
AT	777	618	11.5	17.9	0.60
BE	823	618	13.9	25.1	0.77
DK	793	562	12.3	30.2	0.84
FI	627	427	9.6	32.3	0.85
FR	688	508	10.9	23.8	0.72
GE	696	545	13.3	19.7	0.64
IE	883	683	10.5	19.3	0.63
NL	804	635	12.4	18.2	0.71
PT	517	370	6.7	28.2	0.76
SW	708	489	11.2	31.3	0.92
UK	798	593	9.5	23.1	0.75
US	1158	857	18.4	27.2	0.71

Note: The whole sample consists of 42975 households with the husband at least working 30 hours/week. By specification, household’s non-labor income includes husband’s earnings. Income and hours are averages/week, wages are averages/hour. Income and wages in 2001 PPP-USD. *Source:* Own calculations based on EUROMOD and TAXSIM.

Women from the US show the highest net wages per hour and clearly work more than average weekly hours across countries (27.2 hours/week). Together with husbands’ earnings, this results in the highest household net income on average per week in the sample (1158 PPP-USD). However, females from the Nordic countries (Denmark, Finland, Sweden) show the highest inclination to work (all above 30 hours/week and participation

rates larger than 80%). Also, Portuguese married women, the well-known exception out of the Southern European countries, tend to work more than US females - even though their wages are by far the lowest across countries. In contrast, women from Germany, Ireland, Austria and the Netherlands show relatively low participation rates and hours.

5 Results

This section presents results of the empirical analysis in four steps. First, we outline estimated household and country specific preference heterogeneity. Then, we present information on cross-country orderings for the different individual welfare measures. Next, a decomposition of total heterogeneity into estimated preferences and demographic composition is performed. Finally, we present some robustness checks.

5.1 Estimated preference heterogeneity

We first present estimation results for the utility function, separately retrieved for each country with the same empirical specification. For lack of space and to summarize preference heterogeneity across countries, we focus on average *MRS* between consumption and leisure as reported in Table 2.¹¹ For all observations i , *MRS* are computed at a fixed consumption-labor bundle in order to exclusively capture the shape of different preference structures rather than the impact of different actual locations (c, h) along individual indifference curves.¹² In Table 2, fixed (c, h) -bundles correspond to the average and to certain percentiles of the global c and h distributions ($p10$, median and $p90$ values). *MRS* substantially differ across countries. They are particularly large in Ireland, Germany, Austria and the Netherlands, countries known for low participation levels among married women (see Table 1). Inversely, Nordic countries, Portugal, Belgium and the US show the relatively lowest *MRS* on average. Given our focus on the role of heterogeneity in welfare evaluations, we shall decompose the variations of *MRS* with respect to country demographics and country preferences in Section 5.3.

¹¹Precise estimation tables are available from the authors upon request. The impact of taste shifters (age, children etc.) can directly be seen from detailed *MRS* reported in Table A.1 in the Appendix. We find that the compensation needed in income to outweigh one additional hour of work is clearly higher for women with young children or lowly educated females compared to the average. That is, *MRS* are declining in age of children and level of education. For instance, the average *MRS* in column 1 for women with children younger than 3 years old is about 5 PPP-USD higher compared to the average *MRS* of the whole sample (14.2 versus 9.1 PPP-USD).

¹²As a preliminary check, we have verified that *MRS* are always positive and increasing as required from Section 3 – i.e., for all countries, we find that $\beta_c > 0$, $\alpha_c < 1$ and $\alpha_l < 1$; for the term β_{li} which incorporates heterogeneity, no more than 1% of the observations per country violates the monotonicity condition on leisure – these observations are excluded from the sample.

Table 2: Marginal rates of substitution (between consumption and leisure) by countries

	MRS (\bar{c}, \bar{h}) (1)	Standard error (2)	MRS (c^{p10}, h^{p10}) (3)	MRS (c^{p50}, h^{p50}) (4)	MRS (c^{p90}, h^{p90}) (5)
Full sample	9.1	(5.5)	5.3	8.3	14.8
<i>Country</i>					
AT	13.5	(5.4)	9.7	13.0	18.6
BE	7.3	(2.2)	4.5	6.8	11.4
DK	5.8	(0.6)	3.1	5.2	9.9
FI	4.0	(0.5)	2.0	3.6	7.1
FR	10.1	(3.2)	4.8	8.8	18.6
GE	13.8	(8.4)	7.6	12.4	22.7
IE	18.2	(7.7)	10.8	16.9	29.0
NL	13.9	(5.3)	7.3	12.4	24.1
PT	3.8	(1.0)	2.4	3.6	5.8
SW	5.7	(0.8)	2.1	4.7	13.0
UK	10.0	(4.7)	5.9	9.2	15.9
US	7.0	(3.3)	4.5	6.6	10.6

Note: (\bar{c}, \bar{h}) is the bundle with global means \bar{c} and \bar{h} , (c^{p10}, h^{p10}) contains $p10$ -values for c and h , (c^{p50}, h^{p50}) the median- and (c^{p90}, h^{p90}) $p90$ -values accordingly. c -values in 2001 PPP-USD. *Source:* Own calculations based on EUROMOD and TAXSIM.

5.2 Cross-country welfare rankings

We first pool households from all countries into one sample and compare individual ranks for the different metrics by use of correlation plots. Moving closer to country comparisons, we then investigate how average country positions change by choice of the metric.

Rank correlations. For the pooled country sample, Figure 2 shows empirical rank correlations between individual positions in the percentile distribution of the different metrics. For the sake of comparison, the two upper panels show correlations when identical preferences are assumed (instead of allowing for full heterogeneity). This corresponds to the prominent approach in empirical welfare analysis described above. Precisely, for all households in the pooled sample, we fix their preferences to that of the global median household (in terms of MRS) while retaining their actual (c, h) -choices and non-preference related characteristics (net wages and non-labor income). The metrics are recalculated under these conditions. As indicated in the upper-left panel of Figure 2, any metric can be used at this stage without altering the correlation (which is independent of the choice of the reference household and verified in the upper-right panel).¹³ Note that overall reranking due to the account of leisure in the money metrics is fairly modest when agents

¹³Indeed, this illustrates nothing else than what Roberts (1980) proved, namely, that individual welfare orderings are reference price independent when preferences are homothetic across individuals.

do not differ in preferences. This could of course vary with the choice of the reference household and is checked in the robustness analysis in Section 5.4.

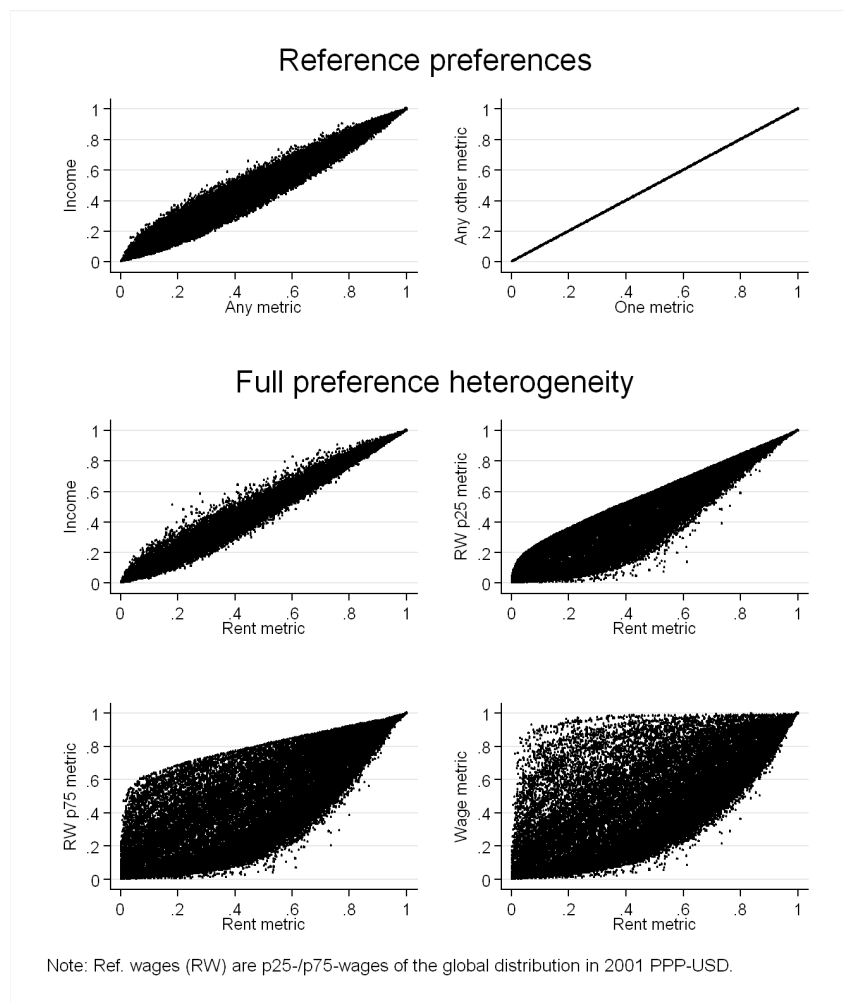


Figure 2: Rank correlations of empirical welfare metrics using reference preferences vs. full heterogeneity in preferences

The next four panels of Figure 2 compare rank distributions for two measures at a time when full heterogeneity in preferences is accounted for. We observe substantial reranking of individual positions between the metrics. While the center-left panel of Figure 2 still reveals a quite strong correlation between the individual positions under pure income and the “rent” metric (similar to the upper-left picture), the correlation between the “rent” and the further metrics in the following three panels sequentially decreases when taking distaste for leisure increasingly into account. In the bottom-right panel, only a weak correlation remains between the “rent” and the “wage” metric, showing the relatively largest reranking between individual situations. The next paragraph analyzes to which extent these rerankings affect cross-country orderings of individual welfare.

Welfare rankings. As a preliminary exercise, we compare cumulative distribution functions (CDF) of the different metrics for two illustrative countries, namely the US and Ireland. The upper-left panel of Figure A.1 in the Appendix shows that US households are relatively best off in terms of income or under the “rent” criterion. However, moving to the “rent+reference wage” metric, CDFs start to cross and households from the US become worse off. For the “wage” criterion, Irish households are now clearly best off. In the following, we analyze for the pooled country sample how these differences in (cardinal) welfare metrics translate into different cross-country welfare rankings.

In Table 3, we use the global distribution of individual ranks to compare countries on basis of the average percentile position of households for each measure. Our focus is on how the country ranking changes with the definition of the metric, i.e. with different normative rationales about how to treat heterogeneity in tastes. When using the pure income measure in column 1, the taste for leisure is simply neglected. Here, US households clearly rank first on average (63rd percentile), due to high average working hours and wage rates.¹⁴ In the second column, individual heterogeneity in consumption-leisure preferences is neglected and identical preferences are assumed (reference household). Recall that, corresponding to the previous paragraph, individual positions (and thus, also average percentile positions by countries) do not change by definition of the metric under these conditions. For instance, we see that Irish (US) households rank slightly better (worse) on average under the metrics than under pure income – simply, because a money metric accounts for leisure on top of income while Irish (US) women work relatively less (more) than the average.¹⁵

Once heterogeneous work preferences are fully respected, the rankings will change by choice of the metric. This is reflected by columns 3-7 in Table 3. For instance, US households also rank first under the “rent” metric while the average percentile is even slightly increased. That is, some US households are replaced at the bottom of the distribution by households from countries like Ireland, where a higher preference for leisure is observed (percentile 47 on average for the “rent” metric after 53 for income). The picture gradually changes when moving to the “rent + reference wage” criteria and finally, to the “wage” metric which shows the highest sensitivity to work distaste. In the latter case, US households rank at the 57th percentile on average versus Irish households at the 74th. Changes in the same direction as for the US are even more pronounced among Nordic countries while changes in the opposite direction are particularly strong for Austria, Germany and the Netherlands. The difference between average ranks under the “rent” and the “wage” metric is presented in the last column, with remarkable changes of

¹⁴Country rankings for net income are also broadly in line with respective GDP rankings.

¹⁵However, recall from the previous paragraph that this result is dependent on the specification of the reference household. “Extreme” reference preferences in terms of very large (small) *MRS* will lead to rank reversals. See Section 5.4.

Table 3: Average percentile position of households in the global welfare ranking - by country and metrics

<i>Country</i>	Income (1)	<i>Ref. preferences</i>	<i>Full heterogeneity in preferences</i>					Δpp
		Any metric (2)	Rent (3)	RW <i>p</i> 25 (4)	RW <i>p</i> 50 (5)	RW <i>p</i> 75 (6)	Wage (7)	Rent-Wage (8)
AT	43.6	47.3	41.3	49.1	54.4	58.1	61.0	19.7
BE	49.2	48.6	49.9	47.9	45.4	43.3	42.1	- 7.9
DK	47.2	42.6	48.0	39.9	35.2	32.2	31.3	- 16.7
FI	29.7	24.0	34.3	18.6	15.5	13.7	13.9	- 20.4
FR	34.4	34.5	34.1	35.5	36.1	37.1	37.3	3.2
GE	36.3	38.9	35.9	40.4	43.8	46.7	50.4	14.5
IE	53.1	56.1	46.5	53.8	60.6	66.5	73.9	27.4
NL	47.6	51.3	47.4	53.0	57.1	60.4	64.6	17.2
PT	19.1	17.9	21.8	15.4	13.9	12.8	12.3	- 9.5
SW	38.1	33.5	41.9	29.1	25.8	24.0	23.8	- 18.2
UK	45.0	45.7	44.2	46.2	47.1	47.7	48.4	4.2
US	63.3	62.3	63.4	61.7	60.1	58.5	56.7	- 6.7

Note: For each metric, we compute the percentile position of each household in the global ranking and average them across all households from the respective country. Reference wages for the “rent + reference wage” metrics (RW) are *p*25-, *p*50- and *p*75-wages of the global distribution in 2001 PPP-USD. *Source:* Own calculations based on EUROMOD and TAXSIM.

at least 15 percentage points for 7 out of 12 countries. The magnitude of rank reversals is all the more striking as our selection of countries is quite homogeneous, focusing on the relatively wealthy EU countries (Continental and Nordic Europe plus the two Anglo-Saxon countries) and the US.¹⁶ Thus, this result suggests that heterogeneous consumption-leisure preferences are the driving factor for individual rerankings across countries. In addition, note that international rankings are affected by population size, which probably limits the extent of rank reversals for large countries. The same is true for natural differences in household non-labor income (husband’s earnings) and female wages across countries (given individual choices).

Interpretation. As explained in Section 3, the society should redistribute towards the apparent work-averse (work-loving) under normative views that hold individuals only minimally (maximally) responsible for work distaste. If we attempt an analogy to our results in terms of country comparisons, we may cluster households according to certain groups of countries. For instance, households from apparently “work-loving countries” (as Den-

¹⁶The case of Portugal is an exception. It is different from other Southern countries in the sense that female participation is very high. However, wage rates are extremely low (among the lowest in Europe). This explains why ranking differences between the metrics for Portuguese households themselves exist as expected while there are simply too few households changing their relative international position to push Portuguese households on average out of the bottom of the global distribution.

mark and the US) are better off on average than households from apparently “work-averse countries” (e.g., Austria and Ireland) under the “rent” criterion. Then, under normative views that assign low responsibility for work aversion, international redistribution would be justified from the former group of countries to the latter. Inversely, under the “wage” metric, the latter are better ranked and should redistribute toward the hard-working nations. These statements simply illustrate – by means of the normative rationales underlying the different metrics – the consequences in terms of welfare comparisons across relatively similar countries when accounting for a very basic source of heterogeneity in the economic literature, i.e., the individual consumption-leisure trade-off.

5.3 Assessing the differences in welfare rankings

Finally, we check what among the direct components of the labor supply model, i.e. estimated α and β parameters or country differences in socio-demographic composition (taste shifters), can explain the differences in the welfare rankings. Recall from Section 4 that both factors determine overall heterogeneity in consumption-leisure preferences and are solely responsible for ranking differences *between* the metrics (Figure 2). We isolate all components related to the two factors, recalculate metrics and check each time how international rankings are affected. Results reported in Table 4 first show the coefficient of variation for *MRS*. Variation in *MRS* is taken as an indicator for the extent to which a certain factor contributes to overall taste differences. Columns 2 to 6 present how empirical rank correlations between income and the different metrics change for the different scenarios.

In the baseline scenario (first row), we assume reference preferences, i.e. preference parameters and characteristics are taken from the median *MRS* household as defined above. The coefficient of variation for *MRS* equals zero by construction and the correlation between income and any metric, 0.98, corresponds to the aforementioned result in the top-left panel of Figure 2. Rows 2-5 introduce heterogeneity in socio-demographic characteristics. That is, all preference parameters are held constant but some characteristics are allowed to change across countries. In row 2, age differences across countries are the only source of variation. Obviously, it cannot explain much of the variation in *MRS* and leaves the empirical correlations across metrics barely unchanged. Education levels and especially the presence of children seem to explain more of the variation in *MRS* (rows 3 and 4); as a result, rank correlations between income and the metrics become weaker when moving towards the “wage” metric. These effects cumulate when heterogeneity in all three characteristics is allowed (row 5).

In rows 6-9, country-specific differences in preferences are considered. First, all socio-demographic characteristics are kept constant and only differences in estimated preference

Table 4: Variation in MRS and correlation between metrics by different sources of preference heterogeneity

Source of preference heterogeneity:		Coeff. var. in MRS (1)	Rank correlation of income with				
<i>Pref. parameters</i>	<i>Socio-demographics</i>		Rent (2)	RW <i>p</i> 25 (3)	RW <i>p</i> 50 (4)	RW <i>p</i> 75 (5)	Wage (6)
Identical	Identical	0.00	0.98	0.98	0.98	0.98	0.98
Identical	Age only	0.04	0.98	0.98	0.98	0.98	0.98
Identical	Education only	0.21	0.97	0.96	0.95	0.93	0.91
Identical	Children only	0.31	0.98	0.96	0.93	0.91	0.89
Identical	All	0.35	0.98	0.95	0.91	0.86	0.82
Country-specific	Identical	0.39	0.96	0.95	0.91	0.86	0.82
Country-specific	Age only	0.37	0.96	0.95	0.92	0.88	0.83
Country-specific	Education only	0.39	0.96	0.94	0.90	0.84	0.78
Country-specific	Children only	0.58	0.98	0.95	0.88	0.79	0.69
Country-specific	All	0.61	0.99	0.94	0.85	0.75	0.64

Note: *MRS* are calculated for a fixed bundle (\bar{c}, \bar{h}) and averaged where \bar{c} and \bar{h} are global means and \bar{c} in 2001 PPP-USD. The median household in terms of this *MRS* serves as the reference household. *Source:* Own calculations based on EUROMOD and TAXSIM.

parameters determine heterogeneity in tastes. That is, α and β parameters are the only source of variation across countries while characteristics \mathbf{z}_i are set according to the reference household. The magnitude of the effect is very similar to that of accounting for all socio-demographic characteristics in the case before. Thus, country-specific work preferences already explain a good deal of the observed variation in *MRS* and between the metrics. Second, country differences in socio-demographics are combined with variation in different characteristics in rows 7-9. The presence of young children and the level of education (to a lesser extent) have a substantial impact on the variation across countries. When allowing for full heterogeneity in characteristics and estimated preference parameters (last row), results are close to that with heterogeneity in family size (children) only. A standard variance decomposition (ANOVA) for *MRS* and differences in individual ranks across metrics supports the above findings. That is, country-specific preferences as well as the correlation between country-specific preferences and family size are most important and significant factors of variation (detailed results are available upon request).

In addition, we reproduce welfare rankings, again in terms of average percentiles, for two counterfactual situations reflecting the different types of heterogeneity. In the Appendix, Table A.2(a) only maintains the heterogeneity in preference parameters while in Table A.2(b), only differences in socio-demographic characteristics are allowed. The differences between metrics and across countries in Table A.2(a) are by and large similar

to the orderings in Table 3. In contrast, Table A.2(b) only reveals a very small influence of demographics on average ranking positions. This confirms the intuition from the previous results that the ranking of individuals across countries in Table 3 is primarily affected by country-specific preferences (rather than by demographic composition).¹⁷

5.4 Robustness checks

We perform necessary robustness checks with respect to the labor supply specification, the calculation of the empirical welfare metrics and the decomposition analysis. We provide here the main findings (while detailed results are available from the authors upon request).

Labor supply model. For the illustrative purpose of this paper, an interpretationally simple specification for the labor supply model has been used. A Box-Cox specification for the deterministic part of the utility function – as often used in the normative literature – seemed particularly suitable since monotonicity and concavity conditions are usually fulfilled and can easily be checked ex-post. Using a more flexible functional form (e.g. quadratic) is more frequent in the empirical literature on labor supply and taxation. However, notice that the gains from flexibility are partly lost in the present context given that tangency conditions must be imposed (which can be done by adding monotonicity and concavity requirements as constraints directly into the likelihood maximization). This is checked for a subset of the countries under analysis, for which we find very similar results as with the Box-Cox specification.¹⁸

Calculation of welfare metrics. We calculate welfare metrics by using indifference curves based on estimated preference parameters and corresponding to a certain level of utility. This level of welfare is taken as the expected value over a large number of draws for the EV-I errors (while calculating the optimal choice at each draw). However, alternative ways of computation can be suggested. First, metrics for each optimal utility level at each draw can be calculated and averaged over all draws. Second, we can compute the metric for the utility level at the expected labor supply choice (according to predicted probabilities). While these alternative procedures necessarily change cardinal measures, we find that they do not affect substantially the resulting orderings compared to the results presented above.

¹⁷There are few exceptions. For France, the trend in Table 3 is more similar to Table A.2(b), suggesting that the demographic composition drives the result for this country. Also Belgium shows a reverse influence of demographics, which, however, does not outweigh the impact of estimated preferences.

¹⁸Starting with 7 hours choices for both specifications as described above, we also find that estimation results are robust to choosing an even narrower choice set with 13 categories (0 to 60 hours/week with a step of 5 hours). See also Bargain et al. (2011).

Specification of the reference household. For the decomposition analysis in Section 5.3, the reference household in the baseline scenario was specified according to the median *MRS*. However, variation in *MRS* and, hence, correlation between the metrics when partly introducing preference heterogeneity, might be sensitive to that specification. Thus, as a robustness check, further specifications for the reference household have been set with respect to p_{10} and p_{90} values in the global distribution of *MRS*. The same has been done for net income. Average *MRS* and the coefficient of variation for *MRS* of course change quantitatively with the specification. Yet, our core results do not change, i.e. relative differences in rank correlations between metrics and orderings across countries are stable.

6 Concluding discussion

The aim of this paper was to contribute to the ‘beyond GDP’-debate in terms of how to compare well-being in several dimensions across countries. We have departed from standard income rankings by the inclusion of leisure, hence, respecting one of the most primary specifications of welfare in the normative literature. Our main focus was to illustrate for the consumption-leisure space the use of welfare metrics that take preference heterogeneity into account. Our results suggest that taste differences – and their normative treatment – might matter substantially when evaluating individual welfare in an international context. Precisely, households from apparently “work-loving countries” are on average better off under criteria that advise minimal responsibility for work aversion (e.g. the US or Denmark) and, in principle, should redistribute towards the more “work-averse nations”, that are worse off (e.g. Austria or Ireland). Inversely, if the latter group is better off under ethical views that hold individuals maximally responsible for high work distaste, the redistribution scheme would be reversed. The reranking between nations when moving from the former to the latter types of welfare criteria is substantial, which is noticeable given that we consider a relatively homogenous set of countries and since the welfare measures only add one dimension to income (“leisure”). A decomposition analysis showed that cross-country differences in consumption-leisure preferences are driving this result.

For the sake of illustration and implementation of the welfare metrics, we intended to keep the empirical framework of this paper simple. However, a lot remains to be done to bring empirical estimations closer to the possibility of sound normative evaluations. In particular, the fit of labor supply models is often improved by the introduction of a term accounting for fixed costs of work. Thus it is possible to rationalize the non-participation of some people in terms of fixed costs rather than through steep indifference curves – and introducing fixed costs would certainly reduce some of the apparent differences in

MRS across household types and countries. However, fixed costs of work are usually not identified from preferences, as shown by Van Soest et al. (2002), but, if introduced in the model, they may in fact capture some elements of work disutility (or even work utility, i.e., negative fixed costs, if inactivity is a source of despair, as shown by Clark and Oswald, 1994). Importantly, the construction and especially interpretation of welfare metrics as used in the present paper is clearly more complicated under these conditions, i.e., in presence of non-regular and possibly discontinuous indifference curves. We leave these considerations for further research. Also, we have chosen to model married women's labor supply since variability in work hours of this group is more likely to reflect true choices in the consumption-leisure space (and responses to financial incentives) compared to other groups. Of course, a more complete welfare analysis across countries should first include other subgroups as well and second, consider further dimensions of individual well-being besides income and leisure.

A Appendix:

Table A.1: Marginal rates of substitution (between consumption and leisure) by subgroups

	MRS (\bar{c}, \bar{h}) (1)	Standard error (2)	MRS (c^{p10}, h^{p10}) (3)	MRS (c^{p50}, h^{p50}) (4)	MRS (c^{p90}, h^{p90}) (5)
Full sample	9.1	(5.5)	5.3	8.3	14.8
<i>Socio-demographics</i>					
Children younger 3	14.2	(6.9)	8.3	13.1	23.0
Children between 3 and 6	14.0	(7.3)	8.2	12.9	22.7
Children between 7 and 12	11.2	(6.0)	6.5	10.3	18.1
No young children	6.5	(3.1)	3.7	6.0	10.6
Low education	13.1	(6.1)	7.3	11.9	21.7
Medium education	9.4	(5.3)	5.4	8.6	15.5
High education	7.6	(4.8)	4.5	7.0	12.1
Wife younger 25	7.6	(4.9)	4.5	7.1	12.3
Wife between 25 and 55	9.2	(5.6)	5.3	8.5	15.1
Wife older than 55	8.0	(4.5)	4.6	7.3	12.9
Husband younger 25	7.0	(4.0)	4.2	6.5	11.2
Husband between 25 and 55	9.3	(5.6)	5.4	8.5	15.1
Husband older than 55	8.0	(3.9)	4.7	7.4	13.0

Notes: See Table 2. Source: Own calculations based on EUROMOD and TAXSIM.

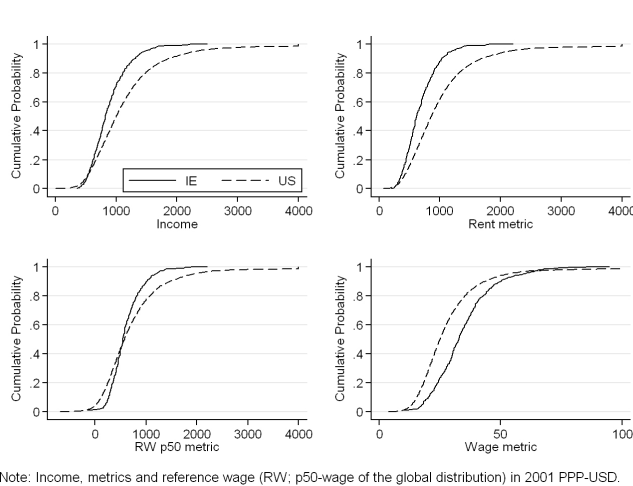


Figure A.1: Cumulative distribution functions (CDF) by metrics for 2 selected countries

Table A.2: Average percentile positions for different sources of preference heterogeneity

(a) Source of preference heterogeneity: differences in estimated preference parameters

<i>Country</i>	Income (1)	Rent (2)	RW <i>p</i> 25 (3)	RW <i>p</i> 50 (4)	RW <i>p</i> 75 (5)	Wage (6)
AT	43.6	42.2	50.9	54.5	56.8	57.8
BE	49.2	50.8	47.1	44.1	41.6	40.6
DK	47.2	50.1	38.3	33.5	30.1	29.7
FI	29.7	35.2	18.2	15.1	13.1	13.6
FR	34.4	36.5	33.8	32.9	32.5	32.5
GE	36.3	31.8	41.6	47.3	51.9	54.7
IE	53.1	44.2	54.4	62.9	69.7	77.3
NL	47.6	45.1	53.3	58.3	62.6	66.8
PT	19.1	21.2	15.9	14.5	13.5	13.3
SW	38.1	42.0	29.5	26.7	25.1	25.2
UK	45.0	43.4	47.0	47.8	48.5	48.5
US	63.3	64.3	61.5	59.6	57.9	56.5

(b) Source of preference heterogeneity: differences in socio-demographic composition

<i>Country</i>	Income (1)	Rent (2)	RW <i>p</i> 25 (3)	RW <i>p</i> 50 (4)	RW <i>p</i> 75 (5)	Wage (6)
AT	43.6	47.7	47.4	47.2	47.1	47.3
BE	49.2	45.9	48.7	49.8	50.7	51.4
DK	47.2	41.1	41.1	41.0	41.1	41.5
FI	29.7	24.1	22.6	22.5	22.6	22.6
FR	34.4	33.7	35.1	35.9	36.6	37.0
GE	36.3	39.7	39.3	39.3	39.5	40.1
IE	53.1	55.5	56.5	57.1	57.5	58.2
NL	47.6	51.0	52.5	53.1	53.6	54.4
PT	19.1	17.2	18.2	18.8	19.5	18.8
SW	38.1	33.5	32.0	31.5	31.3	31.3
UK	45.0	45.8	45.9	45.8	45.8	45.8
US	63.3	62.4	61.9	61.6	61.2	60.9

Notes: See Table 3. *Source:* Own calculations based on EUROMOD and TAXSIM.

References

- Aaberge, R. and Colombino, U. (2008). Designing Optimal Taxes with a Microeconomic Model of Household Labour Supply, ICER Working Paper No. 19/2008.
- Aaberge, R., Colombino, U. and Strøm, S. (2000). Labor Supply Responses and Welfare Effects from Replacing Current Tax Rules by a Flat Tax: Empirical Evidence from Italy, Norway and Sweden, *Journal of Population Economics* **13**(4): 595–621.
- 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**: 767–785.
- Aaberge, R., Dagsvik, J. and Strøm, S. (1995). Labor Supply Responses and Welfare Effects of Tax Reforms, *Scandinavian Journal of Economics* **97**(4): 635–659.
- Alesina, A., Glaeser, E. and Sacerdote, B. (2005). Work and Leisure in the United States and Europe: Why So Different?, *NBER Macroeconomics Annual* **20**: 1–64.
- Atkinson, A. B. (2011). The Restoration of Welfare Economics, *American Economic Review: Papers and Proceedings* **101**(3): 157–161.
- Bargain, O., Caliendo, M., Haan, P. and Orsini, K. (2010). Making Work Pay’ in a Rationed Labour Market, *Journal of Population Economics* **21**(1): 323–351.
- Bargain, O., Orsini, K. and Peichl, A. (2011). Labor Supply Elasticities in Europe and the US, IZA Discussion Paper No. 5820.
- Becker, G. S., Philipson, T. J. and Soares, R. R. (2005). The Quantity and Quality of Life and the Evolution of World Inequality, *American Economic Review* **95**: 277–291.
- Blackorby, C. and Donaldson, D. (1988). Money Metric Utility: A Harmless Normalization?, *Journal of Economic Theory* **46**: 120–129.
- Blackorby, C., Laisney, F. and Schmachtenberg, R. (1993). Reference-price-independent welfare prescriptions, *Journal of Public Economics* **50**: 63–76.
- Blanchard, O. (2004). The Economic Future of Europe, *Journal of Economic Perspectives* **18**: 3–26.
- Blundell, R., Duncan, A., McCrae, J. and Meghir, C. (2000). The Labour Market Impact of the Working Families’ Tax Credit, *Fiscal Studies* **21**(1): 75–104.

- Blundell, R. and MaCurdy, T. (1999). Labor Supply: A Review of Alternative Approaches, *in* O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Vol. 3A, Elsevier, Amsterdam, pp. 1559–1695.
- Blundell, R. and Shephard, A. (forthcoming). Employment, Hours of Work and the Optimal Taxation of Low Income Families, *Review of Economic Studies* .
- Brun, B. C. and Tungodden, B. (2004). Non-welfaristic theories of justice: Is "the intersection approach" a solution to the indexing impasse?, *Social Choice and Welfare* **22**: 49–60.
- Clark, A. E. and Oswald, A. J. (1994). Unhappiness and Unemployment, *Economic Journal* **104**(424): 648–659.
- Creedy, J. and Héroult, N. (forthcoming). Welfare-improving income tax reforms: a microsimulation analysis, *Oxford Economic Papers* .
- Dagsvik, J., Locatelli, M. and Strøm, S. (2009). Tax Reform, Sector-specific Labor Supply and Welfare Effects, *Scandinavian Journal of Economics* **111**(2): 299–321.
- Decoster, A. and Haan, P. (2010). Empirical welfare analysis in random utility models of labour supply, KU Leuven, CES Discussion Paper Series 10.30.
- Eissa, N. and Hoynes, H. W. (2004). Taxes and the labor market participation of married couples: the earned income tax credit, *Journal of Public Economics* **88**(9-10): 1931–1958.
- Eissa, N., Kleven, H. J. 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**: 795–816.
- Ericson, P. and Flood, L. (2009). A Microsimulation Approach to an Optimal Swedish Income Tax, IZA Discussion Paper No. 4379.
- Feenberg, D. R. and Coutts, E. (1993). An Introduction to the TAXSIM Model, *Journal of Policy Analysis and Management* **12**(1): 189–194.
- Fleurbaey, M. (2006). Social welfare, priority to the worst-off and the dimensions of individual well-being, *in* F. Farina and E. Savaglio (eds), *Inequality and Economic Integration*, London: Routledge.
- Fleurbaey, M. (2008). *Fairness, Responsibility and Welfare*, Oxford University Press.

- Fleurbaey, M. (2009). Beyond GDP: The Quest for a Measure of Social Welfare, *Journal of Economic Literature* **47**: 1029–1075.
- Fleurbaey, M. (2011). Willingness-to-pay and the equivalence approach, *Revue d'économie politique* **121**(1): 35–58.
- Fleurbaey, M. and Gaulier, G. (2009). International Comparisons of Living Standards by Equivalent Incomes, *Scandinavian Journal of Economics* **111**: 597–624.
- Fleurbaey, M. and Maniquet, F. (2006). Fair Income Tax, *Review of Economic Studies* **73**(1): 55–83.
- Fuest, C., Peichl, A. and Schaefer, T. (2008). Is a flat tax reform feasible in a grown-up democracy of Western Europe? A simulation study for Germany, *International Tax and Public Finance* **15**(5): 620–636.
- Hodler, R. (2009). Redistribution and Inequality in a Heterogeneous Society, *Economica* **76**: 704–718.
- Immervoll, H., Kleven, H., Kreiner, C. and Saez, E. (2007). Welfare Reform in European Countries: A Micro-Simulation Analysis, *The Economic Journal* **117**(516): 1–44.
- Jones, C. I. and Klenow, P. J. (2010). Beyond GDP? Welfare across Countries and Time, NBER Working Paper 16352.
- Kassenboehmer, S. C. and Schmidt, C. M. (2011). Beyond GDP and Back: What is the Value-Added by Additional Components of Welfare Measurement?, IZA Discussion Paper No. 5453.
- King, M. (1983). Welfare effects of tax reforms using household data, *Journal of Public Economics* **21**: 183–214.
- Ooghe, E. and Peichl, A. (2010). Fair and Efficient Taxation under Partial Control: Theory and Evidence, IZA Discussion Paper No. 5388.
- Pencavel, J. (1977). Constant-Utility Index Numbers of Real Wages, *The American Economic Review* **67**(2): 91–100.
- Prescott, E. C. (2004). Why Do Americans Work So Much More Than Europeans?, *Federal Reserve Bank of Minneapolis Quarterly Review* **28**(1): 2–13.
- Preston, I. and Walker, I. (1999). Welfare measurement in labour supply models with nonlinear budget constraints, *Journal of Population Economics* **12**(3): 343–361.

- Roberts, K. (1980). Price-Independent Welfare Prescriptions, *Journal of Public Economics* **13**(3): 277–298.
- Slesnick, D. T. (1991). Aggregate deadweight loss and money metric social welfare, *International Economic Review* **32**: 132–146.
- Stiglitz, J., Sen, A. and Fitoussi, J.-P. (2009). Report by the Commission on the Measurement of Economic Performance and Social Progress, Technical Report.
- Sutherland, H. (2007). Euromod: the tax-benefit microsimulation model for the European Union, in A. Gupta and A. Harding (eds), *Modelling Our Future: Population Ageing, Health and Aged Care*, Vol. 16 of *International Symposia in Economic Theory and Econometrics*, Elsevier, pp. 483–488.
- Van Soest, A. (1995). Structural Models of Family Labor Supply: A Discrete Choice Approach, *Journal of Human Resources* **30**(1): 63–88.
- Van Soest, A., Das, M. and Gong, X. (2002). A Structural Labour Supply Model with flexible Preferences, *Journal of Econometrics* **107**: 345 – 374.