

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

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

Received: 3 November 2011 / Accepted: 8 October 2012 / Published online: 19 October 2012
© Springer-Verlag Berlin Heidelberg 2012

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

O. Bargain
Aix-Marseille School of Economics, Château Lafarge, 13290 Aix-en-Provence, Les Milles, France
e-mail: olivier.bargain@univ-amu.fr

O. Bargain · M. Dolls · D. Neumann (✉) · A. Peichl · S. Sieglösch
IZA, P.O. Box 7240, 53072 Bonn, Germany
e-mail: dneumann@iza.org

M. Dolls
e-mail: dolls@iza.org

A. Peichl
e-mail: peichl@iza.org

S. Sieglösch
e-mail: sieglösch@iza.org

A. Decoster
CES, Katholieke Universiteit Leuven, Naamsestraat 69, 3000 Leuven, Belgium
e-mail: andre.decoster@econ.kuleuven.be

M. Dolls · D. Neumann · A. Peichl · S. Sieglösch
University of Cologne, Cologne, Germany

A. Peichl
ISER, University of Essex, Colchester, UK

A. Peichl
CESifo, Munich, Germany

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.

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 when individuals have different preferences over the various dimensions of life (see e.g. Fleurbaey 2011). In the economic literature, individual welfare at least depends on consumption and leisure, resulting in the consumption-leisure trade-off in, e.g., labor supply modeling. However, while there has been substantial progress in the development of positive labor supply models in terms of (structurally) estimating individual consumption-leisure preferences, the heterogeneity in preferences is usually neglected in the normative part of the analysis concerned with welfare evaluation. This is due to the difficulties related to interpersonal welfare comparisons. One way to solve this issue is to use preferences of a reference household (King 1983, and, more recently, Aaberge et al. 2004; Aaberge and Colombino 2011). Clearly, this makes individual well-being comparable, but the heterogeneity in preferences is assumed away. In this paper, we contrast this approach to welfare measures that fully account for different individual consumption-leisure preferences (Fleurbaey 2006, 2008) and suggest an international comparison based on pure orderings of individual well-being. Then, we illustrate that the choice of how to treat heterogeneity in preferences may substantially affect the evaluation of welfare across different countries.

The empirical application starts with the estimation of labor supply models, separately for 11 European countries¹ and the US. Focusing on married women, the group most studied in the literature, we rely on 12 representative micro-datasets (on household net income, hours worked and various socio-demographics) and a harmonized econometric approach for all countries in order to obtain comparable estimates of consumption-leisure preferences. We make use of a common structural discrete choice model for labor supply, as used in well-known contributions for Europe (e.g. Van Soest 1995) or the US (e.g. Eissa and Hoynes 2004). This allows us to account for the comprehensive and usually non-linear effect of tax-benefit systems on household

¹ These are Austria (AT), Belgium (BE), Denmark (DK), Finland (FI), France (FR), Germany (GE), Ireland (IE), the Netherlands (NL), Portugal (PT), Sweden (SW) and the United Kingdom (UK).

budgets, which contributes to the identification of the preference parameters. As the labor supply model is identified via a direct parametrization of the utility function, we are then able to 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. These rankings are simple index orderings reflecting interpersonal comparisons of individual utilities and are not based on any kind of a social aggregator function.

The main results 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, take the case, where—when two individuals have different preferences and get the same consumption-leisure bundle—by use of the metric we make the normative choice to consider the agent with the relatively higher willingness-to-work to be better off compared to the agent with the relatively lower willingness-to-work. In that case, we find that households from countries where average female working hours are rather high (as in the US and the Nordic countries) are ranked relatively higher on average compared to a ranking based on income only. Inversely, with a metric that considers agents with a relatively lower willingness-to-work as better off, we attribute a higher welfare level to households from countries where average working hours are rather low (as in most Continental European countries, Ireland and the UK). 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 clearly shows 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 Sect. 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 Sect. 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.² Becker et al. (2005) correct growth

² For 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.

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 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 labor supply literature, we retrieve individual and cross-country specific preference heterogeneity relying on a structural discrete choice 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. [Fleurbaey and Trannoy 2003](#); [Brun and Tungodden 2004](#)).³ 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 equivalent or compensating variations (see e.g. [Aaberge et al. 1995, 2000](#); [Dagsvik et al. 2009](#)) or to aggregate them using a certain SWF (e.g. [Eissa et al. 2008](#); [Fuest et al. 2008](#); [Creedy and Héroult 2012](#)). 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](#); [Ericson and Flood 2009](#); [Aaberge and Colombino 2011](#)). 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.⁴

In the present paper we adopt an approach from the recent social choice literature that allows to fully respect individual preferences in welfare analyses ([Fleurbaey 2006, 2008, 2011](#); [Fleurbaey and Maniquet 2006](#)). In this approach, interpersonal comparisons are conceived directly in terms of subsets of the consumption-leisure space which are nested into each other. This allows deriving a welfare metric which will be clearly ordered for different preferences, making individual situations unambiguously comparable. In the consumption-leisure context, the derivation of comparable, nested

³ A related, more practical reason might be that even if differences in individual preferences were accounted for, it could become a very complicated normative exercise to determine the weights assigned to individual utilities in order to aggregate them.

⁴ Then, 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.

subsets could e.g. be implemented by fixing a specific net wage rate or a certain amount of non-labor income. In that case, the chosen bundle on a given indifference curve is evaluated based on a subset that is tangent to the individual indifference surface. While this procedure is thus similar to the derivation of classic equivalent incomes, the choice of the reference values might be grounded on certain fairness considerations. This makes the normative priors of the interpersonal comparison more explicit—as, e.g., requested by [Atkinson \(2011\)](#).⁵ So far, measures of this kind have not been implemented empirically except in [Decoster and Haan \(2010\)](#) and the present paper. While those 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.⁶ 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 \(2006, 2008\)](#) 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 \(2006, 2008\)](#) for the axiomatic derivation and to [Decoster and Haan \(2010\)](#) for a more detailed illustration.

The setup. Assume that agent i has individual preferences over consumption c_i and labor time h_i , denoted R_i , with $c_i \in \mathbb{R}_+$, $h_i \in [0, 1]$. By R_i , agent i weakly prefers bundle (c_i, h_i) over bundle (c'_i, h'_i) . R_i is represented by a preference representation function u_i defined by the equivalence $(c_i, h_i) R_i (c'_i, h'_i) \Leftrightarrow u_i(c_i, h_i) \geq u_i(c'_i, h'_i)$. Observed preference heterogeneity is introduced via an individual specific vector \mathbf{z}_i (containing all characteristics determining individual preferences), $R_i = R(\mathbf{z}_i)$, and thus $u_i(c_i, h_i) = u(c_i, h_i; \mathbf{z}_i)$. The chosen bundle (c_i, h_i) results from a classic individual utility maximization problem. Let $f(\cdot)$ represent the tax-transfer function that transforms gross non-labor income I_i and gross labor income $w_i h$ (with w_i denoting individual i 's gross wage) into net income c , i.e. $(c_i, h_i) = \max [u(c, h; \mathbf{z}_i) | c \leq f(I_i, w_i h), h \leq 1]$. Hence, the observed bundle of consumption

⁵ Choosing 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](#); [Blackorby et al. 1993](#)).

⁶ This 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.

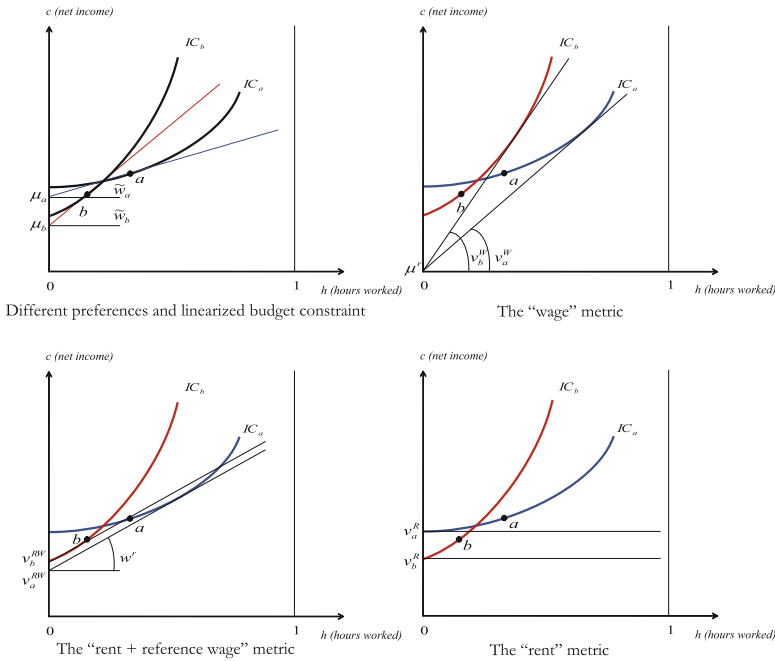


Fig. 1 The different welfare metrics graphically

and leisure results from individual choices subject to preferences and a budget constraint.

The welfare metrics. Assume the agent’s utility function $u(c_i, h_i; \mathbf{z}_i)$ to be well-behaved, i.e. continuous and increasing in its arguments as well as quasiconcave 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 $IC_i = c_i(u, h_i) = \min[c_i | u_i(c_i, h_i) \geq u]$, an associated hypothetical, linear budget constraint that would leave the agent indifferent between choosing from this or her actual budget, can be derived as $c \leq \tilde{w}_i h + \mu_i$ 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 Fig. 1. For the definition of different metrics below we further define the indirect utility function $v_i(\tilde{w}_i, \mu_i) = \max[u_i(c_i, h_i) | c_i - \tilde{w}_i h_i \leq \mu_i]$ and the expenditure function $e_i(u, \tilde{w}_i) = \min[c_i - \tilde{w}_i h_i | u_i(c_i, h_i) \geq u]$, with a fixed level of utility u . The slope of the indifference curve in a given bundle (c, h) is defined as the MRS between consumption and hours worked, $MRS_{c,h} = -\frac{\partial u / \partial h}{\partial u / \partial c}$. Preference heterogeneity shows up by the fact that, in a given bundle (c, h) , different agents are characterized by different MRS. More precisely, an agent i is characterized to be relatively more (less) work averse than an agent i' , if, in a given bundle (c, h) , her indifference curve IC_i is steeper (flatter) than the indifference curve of agent i' , $IC_{i'}$, and thus, $MRS_{c,h;i} > (<) MRS_{c,h;i'}$ (where we also assume that the indifference curves cross at most once). In this setting,

different metrics can be formulated by means of *hypothetical* budget constraints with specific choices of the virtual net wage rate or virtual non-labor income.

First, the “**wage**” metric is defined as the slope of the tangent through the origin at a given indifference curve IC_i , equaling the wage rate \tilde{w}_i of agent i when the value of the virtual non-labor income is set to a reference value of 0, i.e. $\mu_i = \mu^r = 0$. The corresponding function might be called a wage equivalent as firstly introduced by [Pencavel \(1977\)](#). The upper-right picture of [Fig. 1](#) shows that, by use of the metric $v_i^W(u, \mu^r)$, the two agents can be unambiguously ordered from better to worse off even though preferences differ:

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

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

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

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 reference value of $h_i = h^r = 0$. The resulting metric $v_i^R(u, h^r)$ 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 [Fig. 1](#)):

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

Normative interpretation. The key feature of the metrics introduced above is that they fully respect preferences: all metrics increase when the individual moves to a bundle on a higher indifference curve of her *own* preference ordering. However, once we allow for heterogeneous preferences across individuals, the non-trivial issue of making well-founded *interpersonal* comparisons of well-being re-enters the scene. More specifically, one will want to use a specific fairness concept, i.e. a device to separate who is better, and who is worse off, and thus, to determine who should eventually redistribute towards whom—when taking into account that individual outcomes not only result from endowed circumstances, but also from individual preferences. The literature on *responsibility-sensitive egalitarianism* addresses this problem by keeping individuals responsible for the latter, but not for the former ([Fleurbaey 2008](#)).

In order to operationalize this principle, two competing interpretations evolved in the economic literature, namely the *compensation* and the (*liberal*) *reward* principle. The compensation principle states that inequalities due to endowed circumstances (i.e. not due to responsibility factors) should be removed. In contrast, the principle of liberal reward states that inequalities due to individual preferences are legitimate. Although similar at a first glance, both principles are logically independent and to some extent even in conflict with each other.

The three measures defined above give priority to the compensation principle: individuals with poorer hypothetical circumstances are always ranked worse off (as can be seen from Fig. 1) and—depending on social preferences to be specified when moving from the two-agent case to real populations—might deserve compensation. When individuals have identical preferences, they are ranked in exactly the same way by all three measures. However, when preferences differ (as shown in Fig. 1 with crossing indifference curves), it becomes obvious that the metrics also introduce ethical priors about how to treat this preference heterogeneity. Thus, eventual compensation will also depend on the willingness-to-work of the individuals. This might lead to favoring either the industrious or the work averse—which obviously is in possible conflict with the reward principle.

In this paper, we are especially interested in the differences between individual welfare metrics that result from giving priority to the compensation and compromising on the reward principle. That is, we study how the underlying ethical choices in a framework of respecting preference heterogeneity systematically alter *interpersonal comparisons* of well-being—in an international empirical context. Thereby, we focus on simple index orderings of individual well-being levels and do not consider any specific *social ordering function*. The latter would go beyond the question of “who is better and worse off” and additionally requires weighting individual utilities (for this distinction see also Fleurbaey 2007).

In sum, what is relevant for interpersonal welfare comparisons is the fact that the three measures defined rank individuals with identical preferences in the same way (i.e. in accordance with how these individuals would themselves rank their bundles)—while their sole difference is in the way they treat heterogeneity in tastes. Then, once we accept the differential treatment of individual preferences, the ethical choice at hand is related to what generates this difference between the metrics, namely the choice of the references for interpersonal comparisons. In the following we explain how the metrics embody this ethical choice.

In a consumption-leisure space, individuals have different preferences for work (resulting in different levels of exhibited effort) while skill levels (as reflected in gross wages) and non-labor income are assumed to be exogenous endowments to the individuals. The welfare measures defined evaluate individual situations according to hypothetical reference amounts of those endowments such that they would allow individuals to reach their current utility level.

First, the “rent” metric asks 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 corresponds to the level of virtual non-labor income at a reference wage of zero. The bottom-right picture of Fig. 1 illustrates, that in this case, we judge the agent who gets bundle *b*, say Bob, with a relatively lower willingness-to-work to be worse off compared to the agent who has bundle *a*, and a higher willingness-to-work, say Ann. Thereby, redistribution would be granted from Ann to Bob and we might hold Bob only minimally responsible for his preferences.

Second, the “rent + reference wage” metric asks which amount of (virtual) non-labor income would make the individual equally well off compared to her actual situation when receiving a positive (hypothetical) reference wage equal to w^r . Clearly, the

higher this reference wage is, the worse off relatively industrious individuals will be evaluated, i.e. the more they will be favored for receiving protection. This is illustrated in the bottom-left picture of Fig. 1 where w^r is constructed such that the associated linear budget curve is tangent to IC_b in the intersection point with the ordinate. In this case, we evaluate Bob to be better off than Ann for any $w \geq w^r$.

Third, the “wage” metric asks which wage rate would leave the individual indifferent from her current utility level if she had zero (virtual) non-labor income.⁷ Note, that this metric differs from the previous ones in the sense that its properties are less clear in terms of favoring the industrious or work averse. In the upper-right picture of Fig. 1, work averse Bob is considered to be better off compared to industrious Ann and redistribution (in order to equalize hypothetical wage rates) would be justified from Bob to Ann. However, one might easily construct a situation where two agents with crossing preferences are evaluated in the opposite direction by means of the “wage” metric.⁸

4 Empirical approach

The theoretical section presented three specific individual welfare measures that take into account that individuals might have different preferences for consumption versus leisure. In this section, we illustrate the empirical application of the metrics. We proceed as follows. First, we collect information about individuals’ consumption and leisure in 12 countries. Second, we estimate individual preferences based on those revealed choices and various socio-demographic characteristics. Third, we calculate the welfare metrics based on individual consumption-leisure bundles and the estimated preferences (individual indifference curves).

Thereby, using data from different countries enables us to take into account cross-country differences in consumption-leisure preferences, i.e. preference profiles of different populations besides the heterogeneity in individual tastes within a country. Addressing these potential differences requires to keep other factors of the analysis (socio-demographic variation, differences in the tax-benefit systems etc.) as comparable as possible. We therefore make use of a unique setting and estimate household preferences in a harmonized way for all countries under analysis by using (a) comparable datasets with common variable definitions, (b) a common econometric approach to estimate labor supply models for each country and (c) a harmonized tax-benefit calculator to compute net incomes at different points of the household budget curves

⁷ The underlying fairness criterion is developed 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 and *laissez-faire* would be the best policy. The “wage” metric might thus be interpreted as holding individuals maximally responsible for their willingness-to-work. [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 \(2011\)](#) to derive optimal taxes when agents only have partial control over certain effort variables.

⁸ Thus, further research will be necessary to systematically determine how especially the wage metric treats agents with different preferences. In fact, in the empirical section of this paper it turns out that (on average) we are making more comparisons of the kind illustrated in Fig. 1.

as required by the nature of the labor supply model and explained below. We also focus on a specific subgroup of the population, namely married women. First, married women is the group most studied in the labor supply literature as they show lots of variation in work duration and thus also relatively considerable differences in labor supply elasticities (see e.g. [Blundell and MaCurdy 1999](#)). Since this variation partly is affected by differences in consumption-leisure preferences, it might also help to identify differences in the empirical welfare measures. Second, married women's labor supply is less likely to be contaminated by demand-side restrictions compared to single individuals or married men ([Bargain et al. 2010](#)), a factor not explicitly considered with our approach (see below).

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 empirically derive the welfare metrics, we must retrieve indifference curves for each household in our sample and, hence, estimate utility functions. To do so, we specify a structural model of labor supply with discrete choices, which is standard in the literature on tax reforms (see e.g. [Aaberge et al. 1995](#); [Van Soest 1995](#); [Blundell et al. 2000](#)).⁹ Agents are assumed to choose among a set of discrete hours alternatives rather than continuously distributed options which better corresponds to the observed distribution of available hours (non-participation and several part-time, full-time and over-time categories). Also, a discrete choice model better allows to account for the non-linear effect of tax-benefit systems on household budgets as net income needs to be determined at each discrete point. Consumption-leisure preferences are explicitly parameterized as follows while 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 the utility function, we rely on a Box-Cox specification, that is:

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

This specification is frequently used for welfare assessments (see e.g. [Aaberge et al. 1995, 2000, 2004](#); [Decoster and Haan 2010](#); [Aaberge and Colombino 2011](#); [Blundell](#)

⁹ Relying 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.

and Shephard 2012). 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. The deterministic utility is completed by i.i.d. random terms ϵ_{ij} for each choice, leading to the individual random utility function $V_{ij}(u_i, \epsilon_{ij})$. By using a random utility concept, we especially account for the fact that there will always be characteristics of the household (influencing the hours choice) that are known by the household itself while being unobserved by the econometrician. This specifically includes that for a given household, tastes may vary across opportunities which will not be captured by estimating the deterministic part of the utility function (McFadden 1974).¹⁰ As a consequence, non-concavity of u_i would not be inconsistent with random utility theory (as long as V_i is quasi-concave). However, assuming quasi-concavity of u_i is necessary in order to empirically derive (well-behaved) welfare metrics in line with the theory laid out in Sect. 3. This is explained below where we suggest a way to empirically deal with this issue. In addition, in Sect. 5.4, we check robustness with respect to a different, more flexible specification of the utility function and to alternative ways to empirically compute the welfare metrics.

Under the (standard) assumption that random 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 McFadden 1974).

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 and 6 or 7 and 12 years old and regional information.

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 Sect. 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 micro-data based on standard household surveys which provide information on incomes and demographics. For EU countries, we rely on datasets combined with the simulation

¹⁰ Besides, the random term might also capture possible observational errors, optimization errors or transitory situations.

of national tax-benefit systems for years 1998 or 2001 as described in [Bargain et al. \(2012\)](#). 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. Clearly, this assumes away potential cross effects between labor supply decisions of the spouses. However, given the illustrative purpose of the paper, this assumption seems acceptable. To keep the sample relatively homogeneous and avoid too much variation in household's non-labor income (in this context especially including husbands' labor income), we select households where husbands at least work 30 h/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 $J = 7$ h categories including non-participation, two part-time options, two full-time and two over-time categories (0 to 60 h/week with a step of 10 h). 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 estimated with 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 .¹¹ It is calculated numerically using microsimulation models EUROMOD for EU countries and the NBER's TAXSIM for the US.¹²

Empirical welfare metrics. We empirically compute welfare measures based on individual preferences for each household in the sample. Importantly, the random utility framework leads to a frequency distribution of hours choices across the discrete alternatives rather than a perfect prediction of the observed choice. Therefore, we have to compute expected values for the metrics. Yet, one might argue that using a concept of expected measures contradicts the normative background of the individual welfare measures, which essentially relies on observed preferences (derived from observed choices). In order to bring the probabilistic nature of the empirical labor supply model and individual choices together, several approaches are possible. In the baseline, we

¹¹ Using 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. \(2012\)](#) for a more thorough discussion on this point.

¹² For 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).

compute expected metrics as described below and provide robustness checks on different methods in Sect. 5.4.

First, we generate a set of $r = 1, \dots, R$ draws from the EV-I distributed random variable ϵ_j for the given fixed set of hours alternatives (including non-participation). For each draw r , we then compute each individual's utilities V for each alternative j (suppressing index i in the following). As explained above, the welfare metrics can only be empirically derived in a consistent way for well-behaved indifference curves, i.e. based on the deterministic utility. Thus, the deterministic part of the utility of the chosen alternative (the one with highest V), u_r^{max} , will form the basis of the welfare metric for each draw. Subsequently, we average over the number of draws, i.e. $\bar{u} = \frac{1}{R} \sum_{r=1}^R u_r^{max}$. This "expected optimal utility" \bar{u} is used to empirically derive individual indifference curves $IC_{\bar{u}}$, using the general function as introduced in Sect. 3 applied to the Box-Cox specification given in equation (5). Finally, equivalent incomes are computed as follows.¹³ For the "rent" metric, an analytical solution is 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 ordinate), see bottom-right panel of Fig. 1. Due to the Box-Cox specification of the deterministic utility we are not able to derive analytical expressions for the other two metrics. Hence, we must apply numerical procedures. This basically requires searching for the relevant tangency point (c, h) of $IC_{\bar{u}}$ with the hypothetical budget line corresponding to the metric of interest—along the full shape of each individual indifference curve on the hours interval $[0, T]$ (while this point, again, usually will be different from the observed bundle). Once the tangency point (c, h) is found, the value for the metric is determined as well. More precisely, for the "rent + reference wage" metric, the tangency point is 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 μ corresponding to this tangency point is the value for the metric (see bottom-left panel of Fig. 1). Finally, the "wage" metric is derived as the slope of $IC_{\bar{u}}$ for which the $MRS_{c,h}$, because of the zero virtual non-labor income, equals $\frac{c}{h}$ (see top-right panel of Fig. 1). For the numerical derivation of the two last metrics, we rely on a precise iterative procedure by incrementing hours from 0 to T for each household in the sample using very small steps (0.01 h/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. Depending on the year of the data, incomes and

¹³ We abstain from providing the relevant formulas for the concrete Box-Cox specification in order not to exacerbate the understanding of the main procedures with unnecessary technical issues (see Decoster and Haan 2010 for details). The reader may verify the proceeding directly via Fig. 1 and the formulas introduced in Sect. 3.

Table 1 Income and employment statistics

	Data year	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	1998	777	618	11.5	17.9	0.60
BE	2001	823	618	13.9	25.1	0.77
DK	1998	793	562	12.3	30.2	0.84
FI	1998	627	427	9.6	32.3	0.85
FR	2001	688	508	10.9	23.8	0.72
GE	1998	696	545	13.3	19.7	0.64
IE	2001	883	683	10.5	19.3	0.63
NL	2001	804	635	12.4	18.2	0.71
PT	2001	517	370	6.7	28.2	0.76
SW	2001	708	489	11.2	31.3	0.92
UK	1998	798	593	9.5	23.1	0.75
US	2005	1158	857	18.4	27.2	0.71

Note The whole sample consists of 42,975 households with the husband at least working 30 h/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

wages are up- or downrated to the reference year 2001 and transferred into comparable Purchasing Power Parities (PPP)-USD.

Women from the US show the highest net wages per hour and clearly work more (27.2) than average weekly hours across countries (24.7). 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 h/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 hours worked defined as the amount of net income in PPP-USD that is needed by an household to be compensated for an 1 h increase in weekly labor time. Note that MRS are of key relevance in our analysis, rather than labor supply elasticities; while the latter are determined by individual preferences and budget constraints, the former solely represent consumption-leisure tastes in the given framework.¹⁴ For all observations i , MRS_i are computed as the slope of individual indifference curves at a fixed consumption-labor bundle. By doing so, we exclusively capture the shape of different preferences rather than the impact of different actual locations (c, h) along individual indifference curves for a set of given estimates.¹⁵ In Table 2, fixed (c, h) -bundles correspond to the average and to certain percentiles of the global hours distributions ($p10$ -, $p50$ - and $p90$ -values) with accordant net incomes. 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 Sect. 5.3.¹⁶

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 positions of households by countries change by choice of the metric.

¹⁴ This should be distinguished from the fact that individual preferences and thus, MRS, might be also (indirectly) formed by the (country-specific) design of tax-benefit systems in the long run. However, this interesting topic can not be considered in the given static framework where preferences and constraints are clearly separated by construction of the labor supply model. What remains, of course, is the direct influence of tax-benefit systems in the estimation procedure which, however, is genuine as preferences are defined over leisure and net (rather than gross) income.

¹⁵ As a preliminary check, we have verified that MRS are always positive and increasing as required from Sect. 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.

¹⁶ Precise estimation tables are available from the authors upon request. The impact of taste shifters (age, children etc.) is reported in detail in Bargain et al. (2011). 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 for women with children younger than 3 years old is about 5 PPP-USD higher compared to the average MRS of the whole sample (13.7 vs. 8.7 PPP-USD).

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

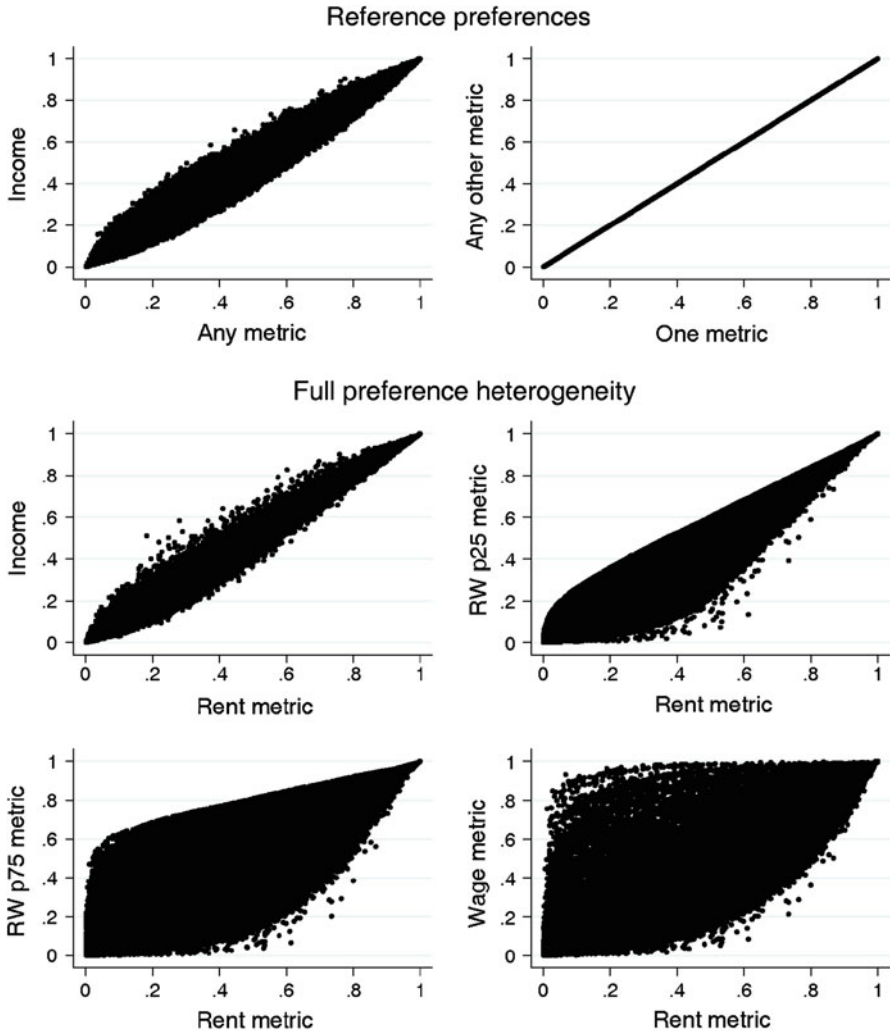
	MRS $(c(\bar{h}), \bar{h})$ (1)	Standard error (2)	MRS $(c(h^{p10}), h^{p10})$ (3)	MRS $(c(h^{p50}), h^{p50})$ (4)	MRS $(c(h^{p90}), h^{p90})$ (5)
Fullsample	8.7	(5.3)	7.0	9.6	12.0
AT	13.2	(5.3)	10.9	13.8	17.1
BE	7.1	(2.1)	5.8	7.7	9.5
DK	5.5	(0.6)	4.4	6.2	7.7
FI	3.8	(0.5)	2.9	4.3	5.5
FR	9.5	(3.1)	7.3	10.9	13.9
DE	13.2	(8.1)	10.7	14.7	17.9
IE	17.6	(7.4)	13.9	19.1	24.2
NL	13.2	(5.1)	10.3	14.8	18.8
PT	3.7	(1.0)	3.0	4.0	5.0
SW	5.3	(0.7)	3.9	6.4	8.4
UK	9.6	(4.5)	7.7	10.5	13.1
US	6.8	(3.2)	5.5	7.3	9.2

Note $(c(\bar{h}), \bar{h})$ is the bundle with global mean hours \bar{h} and corresponding net income $c(\bar{h})$. $(c(h^{p10}), h^{p10})$ contains the mean hours of the 10th percentile in the global hours distribution and the corresponding mean net income $c(h^{p10})$. For $p50$ - and $p90$ -values accordingly. c -values in 2001 PPP-USD. *Source* Own calculations based on EUROMOD and TAXSIM

Rank correlations. For the pooled country sample, Fig. 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 heterogeneity in preferences). 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 $MRS_{c(\bar{h}), \bar{h}}$) 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 Fig. 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).¹⁷ Note that, compared to the pure income measure, overall reranking due to the account of leisure in the money metrics is fairly modest when agents 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 Sect. 5.4.

The next four panels of Fig. 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 Fig. 2 still reveals a quite strong correlation between the individual positions under

¹⁷ Indeed, this illustrates nothing else than what Roberts (1980) proved, namely, that individual welfare orderings are reference price independent when preferences are homogeneous across individuals.

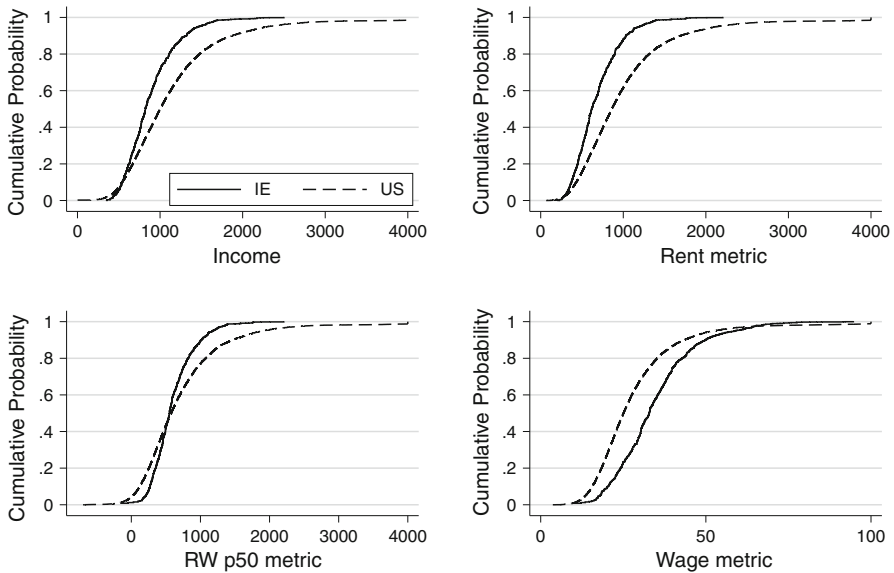


Note: Ref. wages (RW) are p25–p75–wages of the global distribution in 2001 PPP–USD.

Source: Own calculations based on EUROMOD and TAXSIM.

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

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 preferences 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.



Note: Income, metrics and reference wage RW (p50–wage of the global distribution) in 2001 PPP–USD.
Source: Own calculations based on EUROMOD and TAXSIM.

Fig. 3 Cumulative distribution functions (CDF) by metrics for 2 selected countries

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 Fig. 3 shows that US households are relatively better 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 better off. In the following, we analyze for the pooled country sample how these differences in CDFs 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 preferences. When using the pure income measure in column 1, consumption-leisure preferences are simply neglected. Here, US households clearly rank first on average (63rd percentile), due to high average working hours and wage rates.¹⁸ In the second column, individual heterogeneity in consumption-leisure preferences is neglected and identical preferences are assumed (according to the reference household specified above). 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

¹⁸ Country rankings for net income are also broadly in line with respective GDP rankings. Accordant figures are available upon request.

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

	Income (1)	<i>Ref. preferences</i>		<i>Full heterogeneity in preferences</i>			Wage (7)	Δpp Rent-Wage (8)
		Any metric (2)	Rent (3)	RW p_{25} (4)	RW p_{50} (5)	RW p_{75} (6)		
AT	43.6	47.4	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	42.5	48.0	39.9	35.2	32.2	31.3	-16.7
FI	29.7	23.9	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	56.2	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	17.8	21.8	15.4	13.9	12.8	12.3	-9.5
SW	38.1	33.4	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	62.2	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

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.¹⁹

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 successively changes when moving to the “rent + reference wage” criteria and finally, to the “wage” metric. 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 at least 15 percentage points for 7 out of 12 countries. The extent 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

¹⁹ However, 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 affect absolute percentile values. See Sect. 5.4.

Table 4 Average percentile position of the income poor (lowest quintile) in the global welfare ranking—by country and metrics

	Income (1)	Rent (2)	<i>Full heterogeneity in preferences</i>			Wage (6)	Δpp Rent-Wage (7)	ΔMRS ($c(\bar{h}), \bar{h}$) (8)
			RW p_{25} (3)	RW p_{50} (4)	RW p_{75} (5)			
AT	11.8	9.2	20.9	29.6	36.8	41.3	32.1	1.3
BE	20.5	22.2	21.2	20.9	21.1	21.5	-0.7	1.0
DK	20.4	23.9	15.1	13.2	12.3	12.6	-11.3	0.6
FI	6.2	9.6	2.7	2.5	2.4	2.6	-7.0	0.3
FR	5.5	5.7	8.0	9.9	11.9	11.0	5.3	1.7
DE	9.0	10.1	15.2	20.4	25.1	29.4	19.3	0.5
IE	18.2	11.3	22.5	33.8	44.5	55.3	43.9	3.0
NL	15.8	17.4	25.0	31.0	36.5	41.4	24.0	1.5
PT	1.0	1.1	1.0	1.0	1.0	1.0	-0.1	-0.2
SW	13.4	18.0	5.9	5.3	5.1	5.4	-12.6	0.8
UK	10.7	10.4	14.1	17.6	20.8	22.3	11.9	1.3
US	18.6	18.3	18.0	19.1	20.6	21.2	2.9	1.6

Note See Table 3. For each metric, we take the percentile position of each household of the lowest income quintile in the respective country and average. ΔMRS is the difference of $MRS_{c(\bar{h}), \bar{h}}$ for the income poor to the country average as reported in Table 2. *Source* Own calculations based on EUROMOD and TAXSIM

the US.²⁰ 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 may even limit 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).²¹

While Table 3 illustrates potential rerankings of all households in a country on average across the different metrics, a similar analysis for the (income) poor might lead to different conclusions—especially, when preferences of a country's worst-off sufficiently differ from the average preferences in that country. Table 4 therefore shows how the poorest quintile of a country's households in terms of income on

²⁰ The 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.

²¹ Note that, given the ordinal framework, we do not need any information about the differences in the levels of the metrics to answer the question of who will be considered better or worse off. However, we also checked if the differences in the average ranks correspond to (economically) significant differences in the levels of the metrics. With view to column (8) in Table 3, displaying average changes in ranks, we also find considerable average differences in levels between the "rent" and the "wage" metric (defined as full-time equivalent for better comparison), ranging from 18 PPP-USD per week for Finland to 695 PPP-USD for Ireland. Also across countries—exemplarily comparing the US and Ireland (as in Fig. 3)—we find that the absolute average differences across the metrics are substantial, ranging from 331 PPP-USD for the "rent" metric to -245 PPP-USD for the full-time "wage" equivalent.

average is reranked across the different metrics in the global distribution.²² For most countries, results show by and large the same direction as in Table 3, again summarized by the difference between average ranks under the “rent” and the “wage” metric in the last column. The extent of rerankings, however, differs. For instance, the income poor in Portugal find themselves in the lowest percentile of the global distribution and thus, unsurprisingly fare also worst under the remaining metrics, with a marginal improvement for the “rent” metric only. Contrary, rerankings are even more significant for households from countries with a relatively lower preference for work, as e.g. Ireland. For Belgium, there is barely an effect and most interestingly, the ranking of the poor in the US changes in the opposite direction compared to Table 3. These effects might be somewhat explained with view to clearly higher MRS for this group in both countries compared to the average (column (8) in Table 4).

Interpretation. As explained in Sect. 3, the metrics applied differ only in the way they treat heterogeneity in consumption-leisure preferences. As a result, agents with different willingness-to-work might be evaluated very differently depending on the metric. Then, the first and most important question is, who will be considered better and worse off under the various criteria. Therefore, we focused on a pure index ordering for each metric based on individual percentile positions in the accordant global distribution. In terms of country comparisons, we may cluster households according to certain groups of countries. For instance, households from apparently “work-loving countries” (as Denmark and the US) are better off on average than households from apparently “work-averse nations” (e.g., Austria and Ireland) under the “rent” criterion. The reason is that with the “rent” metric, the policy maker tends to evaluate an agent with a higher willingness-to-work to be better off compared to another agent with a lower willingness-to-work (assigning low responsibility for work aversion). Thus, the latter would eventually be favored to receive redistribution from the former and on average, we more often conclude from our interpersonal comparisons that households from Ireland should be “favored” over those from the US. Contrary, under the “wage” metric, we obviously more often favor households from the US over those from Ireland (due to higher responsibility assigned to work aversion). However, these considerations are based on the average percentiles for all households while we might conclude differently when looking at subgroups (income quintiles) of a countries population, as additionally considered in the previous paragraph.

²² This should be distinguished from a feature of the metrics applied which has not been mentioned so far. In fact, it has been shown that the normative principles underlying the different metrics also single out a specific way of how to aggregate them, namely using a maximin (leximin) SWF with infinite aversion to inequality and thus, focusing on the worst-off (Fleurbaey 2008). Again, as this paper is about *interpersonal comparisons* and not about *social evaluation*, we do not consider any type of an aggregator function for our analysis. However, looking at how the poor of each country fare in the world distribution might be worth for answering the question of who will be better or worse off under which metric.

5.3 Assessing the differences in welfare rankings

Finally, we check what among the direct components of the labor supply model, i.e. estimated country-specific α and β parameters or country differences in socio-demographic household composition (taste shifters), can explain the differences in the welfare rankings. Recall from Sect. 4 that both factors determine overall heterogeneity in consumption-leisure preferences and are solely responsible for ranking differences between the metrics (Fig. 2). We start with identical preferences imposed for each household in the sample, isolate all components related to the two factors and separately introduce heterogeneity based on the set of estimates as derived in Sect. 5.1—while keeping individual budgets and observed choices (c , h) fixed. We thus do not re-estimate the models but perform a pure decomposition analysis with respect to observed heterogeneity. Under these conditions, we recalculate the metrics and check each time how international percentile distributions are affected. Results reported in Table 5 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 the “rent” and the further metrics change for the different scenarios (equivalent to the correlation plots presented in Fig. 2).

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

Table 5 Variation in MRS and correlation between metrics by different sources of preference heterogeneity

Source of preference heterogeneity:		Coeff. var. in <i>MRS</i> (1)	Rank correlation of Rent metric with			
<i>Pref. parameters</i>	<i>Socio-demographics</i>		Income (2)	RW <i>p</i> 25 (3)	RW <i>p</i> 75 (4)	Wage (5)
Identical	Identical	0.00	0.98	1.00	1.00	1.00
Identical	Age only	0.04	0.98	0.99	0.99	0.99
Identical	Education only	0.20	0.97	0.99	0.95	0.94
Identical	Children only	0.31	0.98	0.97	0.90	0.88
Identical	All	0.35	0.98	0.96	0.86	0.82
Country-specific	Identical	0.40	0.96	0.95	0.82	0.77
Country-specific	Age only	0.40	0.97	0.95	0.84	0.79
Country-specific	Education only	0.40	0.96	0.95	0.81	0.75
Country-specific	Children only	0.65	0.98	0.92	0.72	0.60
Country-specific	All	0.60	0.99	0.91	0.70	0.59

Note *MRS* are calculated for a fixed bundle $(c(\bar{h}), \bar{h})$ with global mean hours \bar{h} and corresponding net income $c(\bar{h})$ and averaged. *c*-values 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

above.²³ The coefficient of variation for MRS equals zero by construction and the correlation between the “rent” metric and income equals 0.98, while being perfect for the other metrics (which corresponds to the aforementioned results in the top panels of Fig. 2). Rows 2–5 introduce heterogeneity in socio-demographic characteristics. That is, all preference parameters are held constant according to the reference household but some characteristics are allowed to change across countries and households. In row 2, age differences are the only source of variation. Obviously, this cannot explain much of the variation in MRS and barely changes the empirical correlations across metrics. 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 introduced (row 5).

In rows 6–9, country-specific differences in preferences are considered. First, all socio-demographic characteristics are kept constant and solely differences in estimated preference 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 consumption-leisure 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. Here, especially the presence of young children has a substantial impact on the variation across countries, which seems to account for most of the variation when allowing for full heterogeneity in characteristics and estimated preference parameters (last row). A standard variance decomposition (ANOVA) for MRS and differences in individual ranks across metrics supports these findings. That is, country-specific preferences as well as the correlation between country-specific preferences and family size (children) are most important and significant factors of variation (detailed results are available upon request).

While the results presented so far only give an indication about what affects overall correlation between the ordinal metrics, nothing is said yet about which factors actually drive the observed differences in individual cross-country rankings. Therefore, we additionally reproduce welfare rankings, again in terms of average percentiles, for the two main counterfactual scenarios reflecting the different sources of heterogeneity. Table 6(a) only maintains differences in socio-demographic characteristics while in Table 6(b), only the heterogeneity in preference parameters is accounted for. As can be seen, the differences between metrics and across countries in Table 6(b) are by and large similar to the orderings in Table 3. In contrast, Table 6(a) only reveals a very small influence of demographics on average ranking positions. This suggests that

²³ Note that results will depend on the choice of the reference household, why they should also at this stage be considered as illustrative. However, we check for different specifications of the reference household in Sect. 5.4.

Table 6 Average percentile positions for different sources of preference heterogeneity

	Income	Rent	RW p_{25}	RW p_{50}	RW p_{75}	Wage
(a) Source of preference heterogeneity: differences in socio-demographic composition						
	(1)	(2)	(3)	(4)	(5)	(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
(b) Source of preference heterogeneity: differences in estimated preference parameters						
	(1)	(2)	(3)	(4)	(5)	(6)
AT	43.6	41.6	50.7	55.1	57.9	59.3
BE	49.2	50.5	47.6	44.7	42.2	41.1
DK	47.2	50.4	38.4	33.3	29.8	29.3
FI	29.7	35.5	18.2	14.9	12.9	13.4
FR	34.4	36.8	33.9	32.8	32.3	32.2
DE	36.3	31.4	41.3	47.4	52.2	55.4
IE	53.1	42.7	52.9	62.2	69.9	78.6
NL	47.6	44.7	53.1	58.4	62.9	67.5
PT	19.1	21.2	16.1	14.7	13.6	13.4
SW	38.1	42.5	29.3	26.2	24.4	24.5
UK	45.0	43.4	47.1	47.9	48.6	48.6
US	63.3	64.4	61.6	59.5	57.8	56.2

Note See Table 3. Own calculations based on EUROMOD and TAXSIM

the ranking of individuals across countries in Table 3 is indeed primarily affected by estimated country-specific preferences, rather than by demographic composition.²⁴

5.4 Robustness checks

In this section, we perform robustness checks with respect to the labor supply specification, the calculation of the empirical welfare metrics and the decomposition analysis.

²⁴ There are few exceptions. For France, the trend in Table 3 is more similar to Table 6(a), 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.

We only present the main findings while detailed results are reported in [Bargain et al. \(2011\)](#).

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 which can easily be checked ex-post. Using a more flexible functional form (e.g. quadratic, see [Bargain et al. 2012](#)) 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 the countries under analysis using the same data. We find that MRS as defined in [Table 2](#) turn out to be very similar for the Box-Cox and a (constrained) quadratic specification for the labor supply model.²⁵

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. In the baseline, this level of welfare is taken as the expected value over a large number of draws for the EV-I random terms (while always taking the resulting optimal level of utility). However, alternative ways of computation can be suggested, also consistent with the random nature of the labor supply model. First, metrics for each optimal utility level of each draw are computed while averaging then follows over all calculated metric values (for each individual) rather than utilities. Second, we compute the metrics for the utility level corresponding to each discrete hours category and directly take the weighted sum (by predicted probabilities using the expected random term)—rather than artificially drawing many random terms. While these alternative procedures necessarily change the levels of the metrics, we find that they do almost not affect the resulting orderings compared to the baseline results.

Specification of the reference household. For the decomposition analysis in [Sect. 5.3](#), the reference household in the baseline scenario was specified according to the median $MRS_{c(\bar{h}),\bar{h}}$. 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 $p10$ -, ($p50$ -) and $p90$ -values in the global distribution of MRS (net income c), while checking that the reference households selected show sufficient heterogeneity, both in terms of the country they stem from and in terms of socio-demographic characteristics. Average MRS, the coefficient of variation for MRS and correlation between the metrics of course change quantitatively with the specification. Yet, our core results do not change, i.e. the finding that estimated

²⁵ Starting with 7 h 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 h/week with a step of 5 h). See also [Bargain et al. \(2012\)](#).

country-specific preference parameters (rather than socio-demographic differences) determine heterogeneity in the rankings across metrics and countries is confirmed.

6 Concluding discussion

The aim of this paper was to contribute to the ‘beyond GDP’-debate in terms of interpersonally comparing well-being in several dimensions and across different 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 differences in consumption-leisure preferences—and their normative treatment—might matter substantially when interpersonally evaluating welfare in an international context. Precisely, under criteria that tend to evaluate agents with a relatively higher willingness-to-work to be better off compared to agents with a lower willingness-to-work, households from apparently “work-loving countries” (e.g. the US or Denmark) rank higher on average. Inversely, with a metric that considers agents with a relatively lower willingness-to-work as better off, we on average attribute a higher welfare level to households from nations that appear to be more “work-averse” (e.g. Austria or Ireland). The reranking of households 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 homogeneous set of countries and since the welfare measures only add one dimension to income (“leisure”). A decomposition analysis shows that cross-country differences in consumption-leisure preferences are driving this result. The analysis is entirely performed in terms of interpersonal comparisons and our findings suggest that the respect for preference heterogeneity should precede any attempts to compare countries on the basis of a SWF or other forms of aggregated indices.

For the sake of illustration and implementation of the welfare metrics, we intended to keep the empirical framework of this paper simple. Hence, 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 \(1994\)](#)). A similar logic applies to demand-side constraints which restrict the choice set available to the individual ([Dagsvik 1994](#); [Aaberge et al. 1999](#); [Dagsvik and Strøm 2006](#)) and could also result in involuntary unemployment ([Peichl and Siegloch 2012](#)). Here, a specific and additionally demanding requirement in the present context would have been to determine country-specific choice opportunities. In addition, one limitation of the tax-transfer calculators we use is that in-kind benefits or public services more generally are not taken into account due to data limitations.

As the levels of non-cash transfers differ across countries, this has implications for cross-country differences in welfare metrics. However, it is hard to assess ex-ante how accounting for public services would affect the estimation of consumption-leisure preferences and hence the different welfare rankings. Importantly, the construction and especially interpretation of welfare metrics as used in the present paper is clearly more complicated when additionally accounting for the various factors mentioned, i.e., especially in presence of non-regular and possibly discontinuous indifference curves. We leave these considerations for further research. Finally, 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. Then, comprehensive international welfare comparisons might also involve further aspects as the respect for different population sizes or intertemporal comparisons (Fleurbaey and Tadenuma 2009).

Acknowledgements This paper uses 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. Furthermore we use EUROMOD version D16. Here, the ECHP and EU-SILC were made available by Eurostat; the Austrian version of the ECHP by Statistik Austria; the PSBH by the University of Liège and the University of Antwerp; the IDS by Statistics Finland; the EBF by INSEE; the GSOEP by DIW Berlin; the Living in Ireland Survey by the ESRI; the SEP by Statistics Netherlands; the IDS by Statistics Sweden; and the FES by the UK ONS through the Data Archive. Material from the FES is Crown Copyright and is used by permission. None of the institutions cited above bear any responsibility for the analysis or interpretation of the data reported here. Andreas Peichl is grateful for financial support by Deutsche Forschungsgemeinschaft DFG (PE1675). We would like to thank Bart Capéau, Koen Decancq, Erwin Ooghe, Andrew Oswald, two anonymous referees, the editor Marc Fleurbaey, as well as seminar and conference participants in Ann Arbor (IIPF), Bonn (IZA), Canazei (IT), Chicago (SOLE), Dublin (ISNE), Essex (ISER), Leuven (CES), Paris (New Directions in Welfare, OECD) and Stockholm (IMA) for useful comments and suggestions. 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 granting us access to TAXSIM. The usual disclaimer applies.

References

- Aaberge R, Colombino U (2011) Empirical optimal income taxation: a microeconomic application to Norway. CHILD Working Paper No. 16/2011
- Aaberge R, Dagsvik J, Strøm S (1995) Labor supply responses and welfare effects of tax reforms. *Scand J Econ* 97(4):635–659
- Aaberge R, Colombino U, Strøm S (1999) Labor supply in Italy: an empirical analysis of joint household decisions, with taxes and quantity constraints. *J Appl Econom* 14(4):403–422
- Aaberge R, Colombino U, 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. *J Popul Econ* 13(4):595–621
- Aaberge R, Colombino U, Strøm S (2004) Do more equal slices shrink the cake? An empirical investigation of tax-transfer reform proposals in Italy. *J Popul Econ* 17:767–785
- Alesina A, Glaeser E, Sacerdote B (2005) Work and leisure in the United States and Europe: why so different?. *NBER Macroecon Ann* 20:1–64
- Atkinson AB (2011) The restoration of welfare economics. *Am Econ Rev Pap Proc* 101(3):157–161
- Bargain O, Caliendo M, Haan P, Orsini K (2010) Making work pay' in a rationed labour market. *J Popul Econ* 21(1):323–351

- Bargain O, Decoster A, Dolls M, Neumann D, Peichl A, Siegloch S (2011) Welfare, labor supply and heterogeneous preferences: evidence for Europe and the US. IZA Discussion Paper No. 6102
- Bargain O, Orsini K, Peichl A (2012) Comparing labor supply elasticities in Europe and the US: new results. IZA Discussion Paper No. 6735
- Becker GS, Philipson TJ, Soares RR (2005) The quantity and quality of life and the evolution of world inequality. *Am Econ Rev* 95:277–291
- Blackorby C, Donaldson D (1988) Money metric utility: a harmless normalization? *J Econ Theory* 46:120–129
- Blackorby C, Laisney F, Schmachtenberg R (1993) Reference-price-independent welfare prescriptions. *J Public Econ* 50:63–76
- Blanchard O (2004) The economic future of Europe. *J Econ Perspectives* 18:3–26
- Blundell R, MaCurdy T (1999) Labor supply: a review of alternative approaches. In: Ashenfelter O, Card D (eds) *Handbook of labor economics*, vol 3A. Elsevier, Amsterdam, pp 1559–1695
- Blundell R, Shephard A (2012) Employment, hours of work and the optimal taxation of low income families. *Rev Econ Stud* 79(2):481–510
- Blundell R, Duncan A, McCrae J, Meghir C (2000) The labour market impact of the working families' tax credit. *Fiscal Stud* 21(1):75–104
- Brun BC, Tungodden B (2004) Non-welfaristic theories of justice: is “the intersection approach” a solution to the indexing impasse? *Soc Choice Welf* 22:49–60
- Clark AE, Oswald AJ (1994) Unhappiness and unemployment. *Econ J* 104(424):648–659
- Credry J, Hérault N (2012) Welfare-improving income tax reforms: a microsimulation analysis. *Oxford Econ Pap* 64(1):128–150
- Dagsvik J (1994) Discrete and continuous choice, max-stable processes, and independence from irrelevant attributes. *Econometrica* 62(5):1179–1205
- Dagsvik J, Strøm S (2006) Sectoral labor supply, choice restrictions and functional form. *J Appl Econom* 21(6):803–826
- Dagsvik J, Locatelli M, Strøm S (2009) Tax reform, sector-specific labor supply and welfare effects. *Scand J Econ* 111(2):299–321
- Decoster A, Haan P (2010) Empirical welfare analysis in random utility models of labour supply. *KU Leuven, CES Discussion Paper Series* 10.30
- Eissa N, Hoynes HW (2004) Taxes and the labor market participation of married couples: the earned income tax credit. *J Public Econ* 88(9–10):1931–1958
- Eissa N, Kleven HJ, Kreiner C (2008) Evaluation of four tax reforms in the United States: labor supply and welfare effects for single mothers. *J Public Econ* 92:795–816
- Ericson P, Flood L (2009) A microsimulation approach to an optimal swedish income tax. IZA Discussion Paper No. 4379
- Feenberg DR, Coutts E (1993) An introduction to the TAXSIM model. *J Policy Anal Manag* 12(1):189–194
- Fleurbaey M (2006) Social welfare, priority to the worst-off and the dimensions of individual well-being. In: Farina F, Savaglio E (eds) *Inequality and economic integration*. Routledge, London
- Fleurbaey M (2007) Social choice and the indexing dilemma. *Soc Choice Welf* 29:633–648
- Fleurbaey M (2008) Fairness, responsibility and welfare. Oxford University Press, Oxford
- Fleurbaey M (2009) Beyond GDP: the quest for a measure of social welfare. *J Econ Lit* 47:1029–1075
- Fleurbaey M (2011) Willingness-to-pay and the equivalence approach. *Revue d'économie politique* 121(1):35–58
- Fleurbaey M, Gaulier G (2009) International comparisons of living standards by equivalent incomes. *Scand J Econ* 111:597–624
- Fleurbaey M, Maniquet F (2006) Fair income tax. *Rev Econ Stud* 73(1):55–83
- Fleurbaey M, Tadenuma K (2009) Universal social orderings. *CCES Discussion Paper Series*, No.9
- Fleurbaey M, Trannoy A (2003) The impossibility of a paretian egalitarian. *Soc Choice Welf* 21:243–263
- Fuest C, Peichl A, Schaefer T (2008) Is a flat tax reform feasible in a grown-up democracy of Western Europe? A simulation study for Germany. *Int Tax Public Financ* 15(5):620–636
- Hodler R (2009) Redistribution and inequality in a heterogeneous society. *Economica* 76:704–718
- Immervoll H, Kleven H, Kreiner C, Saez E (2007) Welfare reform in European countries: a micro-simulation analysis. *Econ J* 117(516):1–44
- Jones CI, Klenow PJ (2010) Beyond GDP? Welfare across countries and time. NBER Working Paper 16352
- Kassenboehmer SC, Schmidt CM (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. *J Public Econ* 21:183–214
- McFadden D (1974) Conditional logit analysis of qualitative choice behavior. In: Zarembka P (ed) *Frontiers in econometrics*. Academic Press, New York, pp 105–142
- Ooghe E, Peichl A (2011) Fair and efficient taxation under partial control: theory and evidence. CESifo Working Paper No. 3518
- Peichl A, Siegloch S (2012) Accounting for labor demand effects in structural labor supply models. *Labour Econ* 19(1):129–138
- Pencavel J (1977) Constant-utility index numbers of real wages. *Am Econ Rev* 67(2):91–100
- Prescott EC (2004) Why do Americans work so much more than Europeans? *Federal Reserv Bank Minneap Q Rev* 28(1):2–13
- Preston I, Walker I (1999) Welfare measurement in labour supply models with nonlinear budget constraints. *J Popul Econ* 12(3):343–361
- Roberts K (1980) Price-independent welfare prescriptions. *J Public Econ* 13(3):277–298
- Slesnick DT (1991) Aggregate deadweight loss and money metric social welfare. *Int Econ Rev* 32:132–146
- Stiglitz J, Sen A, Fitoussi JP (2009) Report by the comission 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: Gupta A, Harding A (eds) *Modelling our future: population ageing, health and aged care, international symposia in economic theory and econometrics*, vol 16, Elsevier, pp 483–488
- Van Soest A (1995) Structural models of family labor supply: a discrete choice approach. *J Human Resour* 30(1):63–88
- Van Soest A, Das M, Gong X (2002) A structural labour supply model with flexible preferences. *J Econom* 107:345–374