

Working Papers

How Much Does Responsibility Matter in Fairness Measurement?

Laurence Jacquet, Zhiyang Jia, Thor O. Thoresen

Imprint:

CESifo Working Papers

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740

Email office@cesifo.de
<https://www.cesifo.org>

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge

- from the CESifo website: www.ifo.de/en/cesifo/publications/cesifo-working-papers
- from the SSRN website: www.ssrn.com/index.cfm/en/cesifo/
- from the RePEc website: <https://ideas.repec.org/s/ces/ceswps.html>

How Much Does Responsibility Matter in Fairness Measurement?*

Laurence Jacquet^{1†}, Zhiyang Jia², Thor O. Thoresen^{2,3}

¹ CY Cergy Paris Université and THEMA

² Research Department, Statistics Norway

³ Norwegian Fiscal Studies, Department of Economics, University of Oslo

Abstract. Empirical evidence suggests that social acceptance of redistribution depends on whether income differences result from preferences (of which individuals are responsible) or from circumstances. We propose a new empirical method that measures the importance of preferences in the distribution of welfare in the context of tax reforms. We compare two types of Compensating Variation: the standard CV and a new one (CV^{circ}), which is computed assuming that individuals differ only in circumstances. To obtain these metrics, we first estimate a structural job choice model that allows us to take the preferences/circumstances dyad into account. We then use the estimated parameters to compute our two metrics, leveraging a tax reform and applying a simulation approach à la McFadden (1999). Implementing our method with Norwegian data, we find that both welfare metrics display a very similar distribution, except at the very top of the households' income distribution, suggesting this is where responsibility matters.

KEYWORDS: money metric utility, fairness, tax reform, structural labor supply model

JEL CLASSIFICATION: H31, I31, J22, C25

* **Acknowledgement:** We are grateful to comments to earlier versions of the paper by Kristoffer Berg, Vidar Christiansen, John K. Dagsvik, André Decoster, Norman Gemmell, Nicholas Hèrault, Erling Holmøy, Guyonne Kalb, Christos Kotsogiannis, François Maniquet, Alari Paulus, Stéphane Robin, Dominik Sachs, Dirk Van de gaer, Eddy Zanoutene and Floris Zoutman. We also thank Jean-François Carpentier and Christelle Sapata for sharing their STATA program with us. We thank funding by CY Initiative, a program supported by the French National Research Agency (ANR) under the French government grant: “Investissements d’avenir” #France2030 (ANR-16-IDEX-0008).

† **Corresponding author:** Laurence Jacquet. Email: laurence.jacquet@cyu.fr.

1 Introduction

“It is our choices, Harry, that show what we truly are, far more than our abilities.”

(Rowling 1998, p. 358)

What makes a tax reform fair? Liberal egalitarianism offers a compelling answer: correct for inequalities due to circumstances, but respect those that arise from individual responsibility. Survey and experimental evidence support this reasoning: people are more willing to accept inequalities when they come from choices rather than chance.¹

We propose a new empirical method to measure the importance of characteristics for which individuals are responsible in the distribution of welfare, when a tax reform occurs. This method consists in comparing two types of Compensating Variation. The first one, which we denote CV, corresponds to the classic definition (Hicks 1939; McKenzie 1957; Samuelson 1974) as the amount of money needed to restore a household’s original utility after a tax reform. It does fully respect households’ preferences. The other one, which we call circumstance-CV (CV^{circ}), is the amount of money needed to restore a household’s utility assuming all households differ only in circumstances. Responsibility matters whenever the two metrics display different values.

As a first building block, our method requires a structural econometric model that allows one to distinguish between preferences and circumstances. In this paper, we focus on income tax reforms which leads us to estimate a discrete choice labor supply model known as the “job choice” model, developed by Dagsvik (1994) and Dagsvik and Jia (2016). This model intrinsically holds features that mirror the distinction between preferences and circumstances. Individuals exhibit heterogeneous preferences over leisure and labor and face a latent set of job opportunities that arises purely from circumstances.²

The second building block is the computation of our CV metrics, using the approach advocated by McFadden (1999).³ We compute CV as the monetary value that equates the maximum utility before a tax reform and the maximum utility after a tax reform, using the estimated parameters of the job choice model and applying a numerical optimization algorithm. When we compute the conventional CV, which varies according to both preferences and circumstances, we simply leave observed individual characteristics as they are. To compute CV^{circ} , we set those individual characteristics that are related to preferences for leisure (e.g., number of children) at sample median in order to simulate a situation where all households would have the same preferences.

This being done, we further relate our method to the literature on responsibility and circumstances by comparing our CV and CV^{circ} criteria to the Conditional Equality (CE) criterion, which was axiomatically developed by Fleurbaey (2008). Similar to CV^{circ} , the CE criterion involves

¹See e.g., Cappelen et al. (2007), Roemer and Trannoy (2016), Alesina, Stantcheva, and Teso (2018), Scanlon (2018), and Trautmann (2022) for overviews of this literature.

²Preferences are here thought of as tastes as opposed to for instance needs. This is a common assumption in the taxation literature, as emphasized in Kaplow (2008) and Lockwood and Weinzierl (2015). Whether it is possible to distinguish between “efforts” and “endowments”, and the extent to which individuals can control their preferences, are profound philosophical and theological issues which reach far beyond the scope of our analysis.

³This approach is valid for any random utility model, a class of models to which the job choice belongs.

defining a reference value of responsibility characteristics (preferences); see e.g., Bossert and Fleurbaey (1996). CE is determined by calculating the maximum well-being individuals can attain, with hypothetical equivalent budget lines, given the reference value of preferences. In empirical illustrations such as Carpentier and Sapata (2016), this “reference preference” is a function of (observed) variables that are considered to be at the root of differences in preferences for leisure or work, as is the case in our computation of CV^{circ} . We see CV^{circ} as a natural contender of CE, since both are responsibility-sensitive. To compare our method to the CE criterion, we compute utility levels measured by the CE criterion before and after a tax reform (CE_0 and CE_1 , respectively) and calculate $\Delta CE_i = CE_{1i} - CE_{0i}$ for household i . This yields a measure of variation in welfare that can ultimately be compared to our CV and CV^{circ} measures.

We empirically implement our methodology by addressing the welfare effects of a Norwegian tax reform. We focus on the 2013–2019 “bracket tax” reform, which increased the number of brackets in the piecewise linear tax on labor income while generally reducing marginal tax rates. We compare the welfare effects of the tax reform across the household income distribution by applying CV, CV^{circ} and ΔCE .

Comparing CV and CV^{circ} , we find that both welfare metrics display a very similar distribution, except at the very top of the households’ income distribution. There, CV^{circ} displays larger gains (of the reform) than the standard CV, whereas households in the middle of the income distribution are barely affected. The main mechanism is that females with (initial) higher preferences for leisure are assigned lower returns to leisure under CV^{circ} . In addition, the spread in welfare effects across households is slightly reduced when eliminating heterogeneity in preferences. This illustration thus suggests that preferences may not matter so much in fairness measurement, except at the very top. Of course, this finding may not hold for other countries or other reforms.

The results based on the CE criterion are remarkably close, in terms of the distribution of the welfare effects, to those obtained using CV^{circ} . This supports the view that the two methodologies for deriving empirical measures to account for fairness considerations are closely aligned. We only observe a slight divergence at the upper end of the income distribution.

The remainder of the paper is organized as follows. In the next section, we review the related literature. In Section 3, we present the theoretical framework of our study and the fairness criteria. In Section 4, we present the job choice model, explain why this particular model is useful when discussing tax change distributional effects in a preference-circumstance perspective, and demonstrate how the estimates of CV are derived through model simulations. In Section 5, we present the data, the tax reform used in the empirical illustration of our approach, and our empirical results. We conclude in a final section.

2 Related literature

A large public finance literature relies on money metric utility to derive welfare concepts such as, among others, the marginal value of public funds (MVPF), implemented with causal estimates (Hendren and Sprung-Keyser 2020), or the marginal efficiency cost of redistribution (MECR), im-

plemented thanks to a structural labor supply model (Koll et al. 2024). The MECR and MVPF rely on quantifying households' CV and accounting for fiscal externalities. Differing from their approach, we examine only the welfare impact of policies but with a fairness outlook. Since fairness is a key element in the acceptability of an ambitious public policy (Stantcheva 2026), reflecting on what a fair CV could be is potentially an interesting contribution to this dynamic literature.

Money metric utility (or equivalent income) thus plays a central role in our empirical analysis.⁴ The concept of money metric utility is convenient and functional for incorporating different views on fairness (Fleurbaey and Maniquet 2018). Although its use as a welfare measure has been subject to considerable criticism, for example by Blackorby and Donaldson (1988), it has experienced something of a revival over the recent years, notably in the fair allocation literature (Fleurbaey 2008; Fleurbaey and Maniquet 2011; Fleurbaey and Blanchet 2013; Fleurbaey and Maniquet 2018). Bosmans, Decancq, and Ooghe (2018) and Schlee and Khan (2022) more generally suggest rehabilitation of money metric utility in the measurement of welfare and welfare change.

As far as empirical applications are concerned, Preston and Walker (1999) propose a range of individual welfare metrics suitable for interpersonal comparisons of well-being. Furthermore, Decoster and Haan (2015) discuss welfare orderings of households according to different metrics, such as the wage metric and the rent metric, embodying different ethical choices. Their empirical analysis however only controls for observed heterogeneity in households' preferences. Similarly, King (1983) and Aaberge, Colombino, and Strøm (2004) impose a common utility function for all individuals. In our study, we explicitly introduce observed and unobserved heterogeneity in household's and individual's preferences – to control for responsibility versus circumstances.

CV^{circ} and ΔCE both rely on reference preferences. Using reference preferences to move away from subjectivism⁵ is not uncommon in the taxation literature, see e.g., Marchand, Pestieau, and Racionero (2003), Roemer et al. (2003),⁶ Schokkaert et al. (2004), Pestieau and Racionero (2009), Jacquet and Van de gaer (2011) and Jacquet (2014).⁷ This line of thought contrasts with the more traditional approach to welfare measurement, which does not go beyond exploiting individual-specific information (see for instance Capéau, Decoster, and De Sadeleer (2023)). Our approach is, in this sense, related to other contributions that measure unfair inequality as divergences between an observed and a “fair” welfare distribution, as for instance Almås et al. (2011), Ravallion (2017), and Hufe, Kanbur, and Peichl (2022). Decoster and Haan (2015), Carpentier and Sapata (2016), and Ooghe, Schokkaert, and Serruys (2025) are particularly relevant to the present study, as they employ labor supply models to analyze welfare measures – specifically distinguishing between “fair” and “unfair” inequalities in the context of individual heterogeneity.

⁴For comprehensive reviews, see the literature surveys in Slesnick (1998), Fleurbaey and Blanchet (2013), and Decancq, Fleurbaey, and Schokkaert (2015).

⁵The limitations of subjectivism are well recognized, as seen in criticisms such as the treatment of expensive tastes and the overly subjective and idiosyncratic nature of the individual preferences (see, e.g., Dworkin (1981))

⁶Roemer et al. (2003) assume preferences to be quasi-linear and more specifically linear in consumption. Their social objective function is not defined in terms of utilities but in terms of income which can be interpreted as a special case of reference preference.

⁷In this literature, income taxation should not eliminate income differences due to variations in individual preferences for effort, as advocated in e.g. Dworkin (1981), Arneson (1989), Cohen (1989), and Roemer (1998).

We model income and leisure as the primary sources of well-being, which is in line with the traditional consumption-leisure trade-off in labor supply. The advantages of using behavioral labor supply models to study the effects of policy changes is well documented (Blundell 2012). Here, we use a discrete choice labor supply model to obtain the distributions of CV, CV^{circ} , and ΔCE . We build on the “job choice” model developed by Dagsvik (1994) and Dagsvik and Jia (2016), arguing that it is better suited to an applied analysis focusing on the preference/circumstance distinction than the simpler conventional discrete choice labor supply model (van Soest 1995) generally used in the welfare literature (Preston and Walker 1999; Creedy and Kalb 2005; Bargain et al. 2013; Decoster and Haan 2015; Carpentier and Sapata 2016; Picchio and Valletta 2018). This is because the job choice model extends the conventional discrete labor supply model by taking into account job opportunities, which reflect heterogeneity in circumstances, while the trade-off between labor and leisure reflects preferences.⁸

Finally, a thriving literature proposes methods that aggregate money-metric utilities while incorporating fairness considerations; see Fleurbaey and Blanchet (2013), Saez and Stantcheva (2016), Piacquadio (2017), Bosmans, Decancq, and Ooghe (2018), Berg and Piacquadio (2023), and Ooghe, Schokkaert, and Serruys (2025).⁹ Differing from this literature, we focus on the welfare effects of tax changes while explicitly accounting for the preference-circumstance distinction; fairness-adjusted aggregations of money-metric utilities fall outside the scope of our analysis.

3 Fairness in descriptions of distributional effects of tax change

In this section, we present our setup and our fairness criteria. In our empirical illustrations, we focus on couples which are assumed to be unitary. They make harmonized decisions with respect to the labor supply of each spouse. In the present section, to simplify the graphical presentation of our criteria, we assume a single-worker household. In the remainder of the paper, we consider households in which two individuals supply labor, using the standard unitary family model. In the basic framework we present here, there is no childcare cost, no fixed cost of work, no benefit take-up, refinements which can be incorporated relatively easily, as detailed in Blundell and MaCurdy (1999).

Assume that household i has preferences over consumption $c_i \in \mathbb{R}^+$ and (worked hours) $h_i \in \mathbb{R}^+$ described by a twice continuously differentiable utility function $u_i(c_i, h_i; \gamma_i)$, where $\gamma_i \equiv (\gamma_i^1, \dots, \gamma_i^p)$ stands for the household’s characteristics determining its labor-leisure preferences. Households enjoy utility from consumption but endure disutility to work, so $u_{c_i} > 0$ and $u_{h_i} < 0$. Household i chooses c_i and h_i by maximizing utility u_i given its budget constraint as follows:

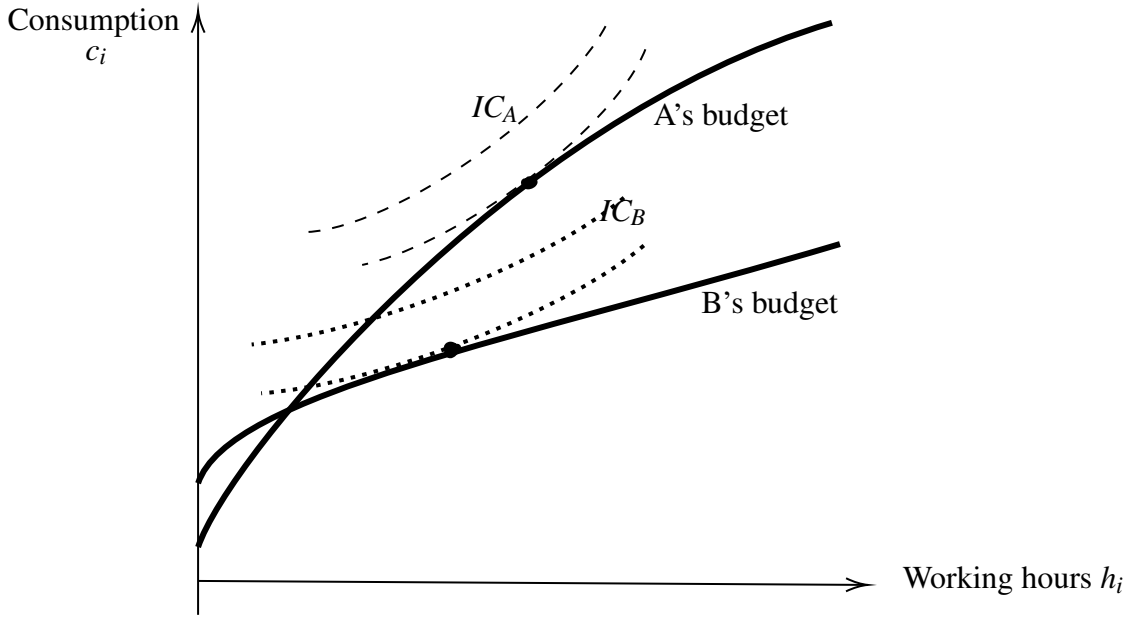
⁸There are no closed-form formulae for CV in discrete choice models that are nonlinear in income. However, Dagsvik and Karlström (2005) give formulae for the distribution and moments of CV. Previous empirical analyses using their framework include e.g. Kornstad and Thoresen (2006) and Dagsvik, Locatelli, and Strøm (2009). See also Herriges and Kling (1999) for a review of the literature on welfare measurement in discrete choice random utility models that are nonlinear in income, Dagsvik, Strøm, and Locatelli (2021) on the development of the Slutsky equation in discrete choice, and Bhattacharya (2024) on nonparametric welfare analysis in discrete choice settings.

⁹Notably, Blackorby and Donaldson (1988) argue against aggregating money-metric utility.

$$\begin{aligned}
& \max_{h_i} u_i(c_i, h_i; \gamma_i) \\
& \text{s.t. } c_i = f_i(y_i, I_i; \gamma_i) \quad \text{with } y_i = w_i h_i,
\end{aligned} \tag{3.1}$$

where I_i denotes non-labor income and y_i refers to labor income defined as the product of hourly wage w_i and hours worked h_i . These three ingredients are transformed into disposable income through the function $f_i(\cdot)$, which aggregates income sources, adds benefits and withdraws taxes and social contributions.¹⁰

Figure 1: Choices of A and B with different preferences and (nonlinear) budget sets after taxation



Notes: A has higher preferences for leisure (more work averse) as signified by steeper indifference curves. B has higher non-labor income, but lower wage than A.

Figure 1 illustrates a case in the (c_i, h_i) -space with the indifference curves (dashed and dotted lines) and budget constraints (solid lines) of two households, $i = A, B$. The slope of the indifference curve in a given bundle, (c_i, h_i) , 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) , individuals are characterized by different MRS. Preference heterogeneity, γ_i , affects the steepness of the indifference curves (Carpantier and Sapata 2016). In Figure 1, household A is relatively more work averse (i.e., has a higher preference for leisure) than household B, since, in a given bundle (c, h) , A's indifference curve IC_A is steeper than B's indifference curve IC_B , and thus, $MRS_{c_A, h_A} > MRS_{c_B, h_B}$ (the indifference curves cross at most once).

We also illustrate different budget sets (after nonlinear taxation) in Figure 1 where B has a higher non-labor income and a lower after-tax wage than A. The same figure also illustrates the incomparability of households with heterogeneous preferences. Deciding whether the high consumption, high working hours bundle of A or the low consumption, low working hours bundle of B is worse

¹⁰In the unitary multi-worker model used in the remainder of this paper, the utility function becomes $u_i(f_i(w_F h_{F_i}, w_M h_{M_i}, I_i, \gamma_i), h_{F_i}, h_{M_i}; \gamma_i)$ where F and M stands respectively for the female and male partner.

is not straightforward. Additional assumptions are required to determine whether B is better-off or worse-off than A .

CV and CV^{circ} criteria

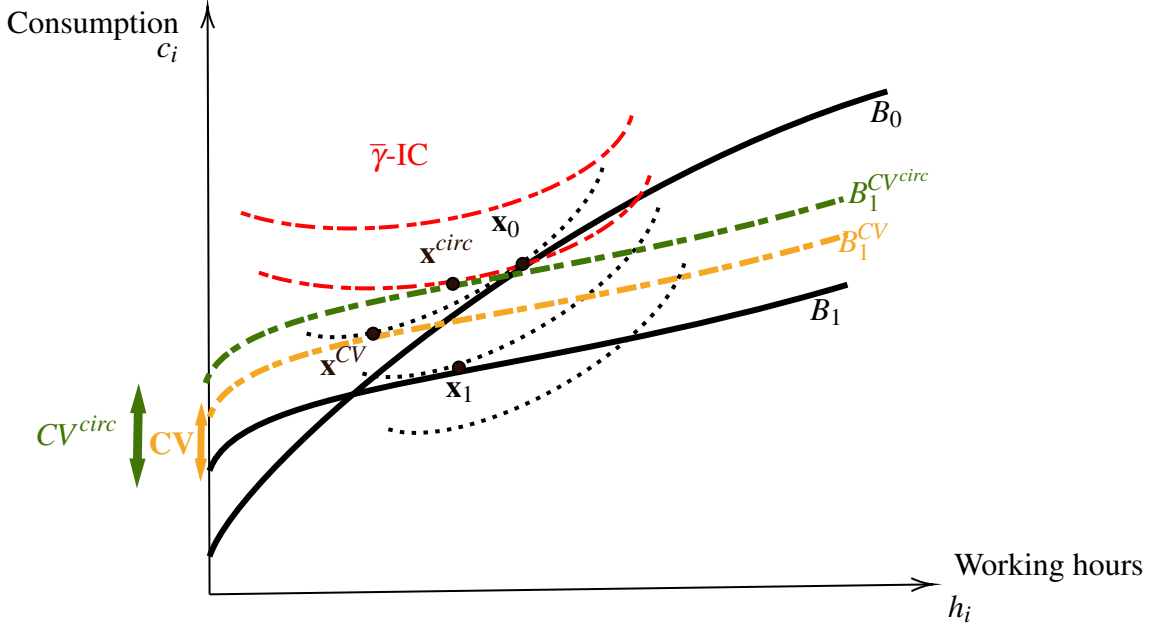
We now present CV and CV^{circ} criteria. CV is the amount of income that must be given to a household after a change in taxation to restore its original level of satisfaction. This money metric utility measures the compensation required to offset the impact of the tax change. Empirically, we will derive measures of CV directly from the estimated preferences of households (Aaberge, Dagsvik, and Strøm 1995; Creedy and Kalb 2005). We estimate CV by calculating the monetary adjustment that equalizes a household's maximized utility before and after the tax change.

The CV^{circ} criterion is a variant of CV that allows us, by comparing the two criteria, to take fairness into consideration. To emphasize the role of responsibility, CV^{circ} relies on some reference value of preferences for leisure or work, i.e., a fixed γ_i , denoted $\bar{\gamma}$. In our empirical application, we will capture this reference variable using the taste-modifying variables (i.e., the variables that determine the preferences for work or leisure in the labor supply model. With $\bar{\gamma}$, the only heterogeneity that remains is from circumstances – e.g., job opportunities, hence the CV^{circ} notation. We use CV^{circ} to examine what would happen in a fictitious, alternative world where all households have the same preference for labor/leisure. We then compare it to the standard CV. Since the latter is constructed using the real preferences of households, responsibility matters whenever the two metrics display different values.

In Figure 2, we graphically construct CV and CV^{circ} for one household. The initial budget set is denoted by B_0 , while B_1 represents the budget set after a tax reform. The bundle $\mathbf{x}_0$ maximizes utility under B_0 , and $\mathbf{x}_1$ is the utility-maximizing bundle under B_1 . CV is obtained by shifting the post-reform budget set B_1 upward until it becomes tangent to the indifference curve passing through $\mathbf{x}_0$. We denote it by the orange dashed budget set B_1^{CV} and the corresponding chosen bundle is denoted by $\mathbf{x}^{\text{CV}}$. CV appears on the vertical axis as the difference between the intercept of B_1^{CV} and that of B_1 .

The construction of CV^{circ} is based on the reference value of preferences for leisure and work, $\bar{\gamma}$. The reference indifference curves ($\bar{\gamma}$ -IC) are shown as dashed red curves. The budget set B_1 is shifted upward until it becomes tangent to the reference indifference curve passing through $\mathbf{x}_0$, yielding the green dashed budget set $B_1^{\text{CV}^{\text{circ}}}$. The corresponding chosen bundle is denoted by $\mathbf{x}^{\text{circ}}$, and the vertical size of this shift defines CV^{circ} .

Figure 2: CV and CV^{circ} for the tax reform that shifts the budget set from B_0 to B_1 .



The CE criterion and ΔCE

An issue that arises now is whether the CV-approach leads to conclusions and policy recommendations that are consistent with those obtained with a well-established criterion of fairness in the social choice literature: the CE criterion. As according to this criterion inequalities in welfare are acceptable only if they arise from conditions or factors for which households can be held responsible, like the reference preference CV approach, empirical analyses based on the CE criterion rely on a reference responsibility parameter, $\bar{\gamma}$. However, the method for measuring welfare changes under the CE framework differs substantially from that used in the reference preference CV approach.¹¹

We follow Carpentier and Sapata (2016) in computing utility levels before and after a tax reform using the CE criterion. First, we determine an equivalent budget for the household, defined as a hypothetical linear budget constraint, with a slope equal to the household's actual wage rate and a lump-sum tax T_i as intercept.¹² This tax is calibrated to maintain the household's current satisfaction as determined by the household preferences (Equation (3.1)) under a given tax system. This is illustrated in Figure A3 in Appendix A.2.

Formally, the lump-sum tax T_i is implicitly defined by

$$u(c_i, h_i; \gamma_i) = \max_h \{u(c, h; \gamma_i) \mid c \leq w_i h - T_i\}. \quad (3.2)$$

Thus, T_i is the amount of income that must be subtracted (or added) from the household's wage income to make the utility level just attainable under its own preferences. Note that T_i is implicitly

¹¹Previous empirical analyses that use the definition of CE to measure unfair inequalities for income include Devought (2008) and Almås et al. (2011).

¹²This budget has to be linear since the CE aims to preserve as far as possible the responsibility principle, which is incompatible with actual non-linear budget sets. Instead, one characterizes nested budget sets, and ranks the budget sets in an unequivocal way, identifying who the worst-off are.

dependent on the household characteristics and underlying tax system. Like CV^{circ} , CE relies on reference preferences $\bar{\gamma}$.

Second, we determine the CE criterion of each household by the highest reference preference indifference curve ($\bar{\gamma}$ -IC in Figure A3) that it can reach within its equivalent budget set. This can be expressed formally as in Carpentier and Sapata (2016):

$$CE_i = \bar{u}_i = \max_h \{ \bar{u}(c, h; \bar{\gamma}) \mid c \leq w_i h - T_i \}. \quad (3.3)$$

Empirically, we impose common variables to represent preferences for leisure and work, as explained in the case of CV^{circ} . Further illustrations of the construction of CE_i are provided in Appendix A.2.

For our purpose, we need a criterion that measures a change in welfare before and after a tax reform. We thus compute welfare levels based on the CE criterion before and after the tax reform, CE_0 and CE_1 , and then calculate $\Delta CE_i = CE_{1,i} - CE_{0,i}$ for household i . We graphically present its construction in Appendix A.3. Ultimately, as a robustness check, we will compare the empirical estimates of ΔCE_i to those of CV_i^{circ} . If the distributions of the former and the latter along the households' income distribution both differ from the distribution of CV in the same region(s), we will have extra evidence that CV^{circ} is consistent with CE, as both criteria point out that preferences matter in the same income interval(s). We present this comparison in Section 5.

4 Welfare effects obtained by structural labor supply model

4.1 Data

To illustrate the empirical content of the welfare-change measures, we use a discrete choice framework – specifically the “job choice” model of Dagsvik (1994) and Dagsvik and Jia (2016) – estimated on Norwegian data. The model is estimated for couples by employing a cross-sectional data set based on merging (using personal identification numbers) the annual Labour Force Survey (Statistics Norway 2024) and the Income and Wealth Statistics of Households (Statistics Norway 2019) for 2015. Information on both actual and formal working time for the main and secondary jobs, along with background variables, such as demographic characteristics and occupation, are included in the Labour Force Survey. Importantly, conditional on labor market participation, respondents are asked to identify themselves as either self-employed or employees. Furthermore, couples in which one of the adults has income from self-employment higher than NOK 115,000 (in 2015-prices)¹³ are excluded. To transform households' non-labor income and labor income into disposable income, we use the tax-benefit model belonging to the LOTTE-family of microsimulation models (Jia et al. 2024).

A person is defined as working if he or she works at least one hour per week. Hours of work is measured as the sum of formal hours of work in the main and the second job (if there is a second

¹³Around EUR 13,000 and USD 14,000, when average exchange rates for 2015 are used (1 USD = NOK 8.07 and 1 EUR = NOK 8.95.)

job). Nominal hourly wage rates are obtained by dividing labor income by (formal) total annual hours of work, where labor income is measured as income from the main and the second job (if the individual has a second job). Couples where one of the adults has weekly hours of work above 80 or wage rate outside the interval NOK [70, 600] (in 2015-prices) are excluded. Summary statistics on the sample are given in Section C of the Appendix, in Table C1.

4.2 Estimation of a discrete choice labor supply model

4.2.1 The job choice model

For our welfare analysis, several structural labor supply models could, in principle, be employed, including Hausman-type models (Burtless and Hausman 1978; Hausman 1979). We argue that the “job choice” model is particularly well-suited for applied fairness analysis, as it accounts for the preference–circumstance distinction and effectively captures key features of labor markets. Indeed, in Norway, as in most OECD countries, one observes considerable peaks at full-time and possibly part-time hours of work, which indicates that most jobs only offer full-time or part-time work schedules. This type of evidence is a key justification for the job choice model, which can be seen as an extension of the conventional discrete labor supply model (van Soest 1995). The job choice model relies on the assumption that the fundamental choice variable is the “job”, which is characterized by given hours of work, a wage rate and other non-pecuniary job-specific attributes. Similar to other discrete choice models, as the one in van Soest (1995), the model allows for the budget constraint to be nonlinear (piecewise linear reflecting real-world taxes and benefits) and non-convex. The job choice model can easily accommodate details of complex tax and transfer systems and simulations of alternative policies can be carried out relatively straightforwardly (Dagsvik et al. 2014).

We will now outline the job choice model and its relevance for the measurement of fairness.¹⁴ Our model and empirical application focus on married couples (with or without children). Recall that couples are assumed to be unitary, i.e., they make harmonized decisions with respect to the labor supply of each spouse, given a common budget constraint. Under this relatively common assumption, the modeling framework for two-person households is a direct extension of the case of single-person households.

4.2.2 Outline of the labor supply model

According to the job choice model, households derive utility from consumption (here set equal to household disposable income), leisure and non-pecuniary attributes of latent jobs. Let $z = 1, 2, \dots$ be an indexation of the jobs and let $z = -1, -2, \dots$ represent the non-market opportunities. The utility function of a household is assumed to have the form

$$U(C, h_F, h_M, z) = u(C, h_F, h_M) + \varepsilon(z), \quad (4.1)$$

¹⁴See also Aaberge, Dagsvik, and Strøm (1995), Creedy and Kalb (2005), Bargain et al. (2013), Decoster and Haan (2015), Carpentier and Sapata (2016), Picchio and Valletta (2018), Capéau et al. (2021), and Ooghe, Schokkaert, and Serruys (2025) on the use of discrete choice labor supply models in measurement of welfare.

where $C \geq 0$ denotes household consumption, h_F and h_M are (annual) hours of work for the female and male in the household, respectively, and $z = (z_F, z_M)$ is the vector of the indices of the jobs chosen by the couple. It is assumed that each job offers only a single given work schedule. Furthermore, $u(C, h_F, h_M)$ is a suitable deterministic function. For given hours of work, h_F and h_M , and wage rates, w_F and w_M (for females and males respectively), household consumption is given by $C = f(h_F w_F, h_M w_M, I)$ where I is non-labor income and $f(\cdot)$ is the function that transforms gross income into consumption. Finally, $\varepsilon(z)$ is a stochastic error term (taste-shifter), which is supposed to capture unobserved heterogeneity of preferences relating to job-specific non-pecuniary attributes. The reason why the index z enters the utility function is that job-specific attributes other than wage and hours of work may affect the utility of the agents.

Now, let $D \subset \mathbb{R}_{\geq 0}$ be the set of feasible hours, including zero hours and let $B(h)$ denote the set of jobs and non-market opportunities with hours of work $h \in D$ that are available to an individual. $B(h)$ is individual-specific and latent, with the number of jobs within $B(h)$, represented by $Q(h)$, which remains unobserved. The terms $\varepsilon(\mathbf{z}) = \varepsilon(z_F, z_M)$ are assumed to be i.i.d. across households and across jobs with the standard Gumbel cumulative distribution function, also known as the type 1 generalized extreme value distribution.¹⁵

From Equation (4.1), and assuming that the hours offered by the spouses in a married couple, h_F and h_M , are independent, the utility of a household choosing their jobs for any given hours (h_F, h_M) is given by:

$$\begin{aligned} V(h_F, h_M, I) &= \max_{z_F \in B(h_F), z_M \in B(h_M)} U(f(h_F w_F, h_M w_M, I), h_F, h_M, z) \\ &= u(f(h_F w_F, h_M w_M, I), h_F, h_M) + \max_{z_F \in B(h_F), z_M \in B(h_M)} \varepsilon(z_F, z_M). \end{aligned} \quad (4.2)$$

Under the Gumbel distributional assumption, as shown in Section B of the Appendix, the indirect utility function (Equation (4.2)) can be rewritten as

$$V(h_F, h_M, I) = u(f(h_F w_F, h_M w_M, I), h_F, h_M) + \log(Q_F(h_F)) + \log(Q_M(h_M)) + \eta_{h_F, h_M}, \quad (4.3)$$

where η_{h_F, h_M} are i.i.d. with standard Gumbel c.d.f. It is indeed a convenient property of the Gumbel c.d.f. that the distribution of η_{h_F, h_M} turns out to be the same as the distribution of $\varepsilon(z)$. It follows from Equation (4.3) that the indirect utility function differs from the corresponding indirect utility in the standard discrete choice model by the sum of the logarithms of the number of jobs, $\log Q_F(h_F) + \log Q_M(h_M)$. In other words, $V(h_F, h_M, I)$ can be considered as consisting of two parts: the first is $u(f(h_F w_F, h_M w_M, I), h_F, h_M) + \eta_{h_F, h_M}$, which comes from the individual preferences of the spouses, and the second part is $\log(Q_F(h_F)) + \log(Q_M(h_M))$, which represents the labor market opportunities facing the household. Thus, the job choice model holds features that mirror the distinction between preferences and circumstances of the fairness literature.

¹⁵The distributional assumptions of the random error terms $\varepsilon(\mathbf{z})$ can be justified by the notion of probabilistic rationality expressed by the Choice Axiom of Luce (Luce 1977).

The probability $\varphi(h_F, h_M)$ of choosing a job with hours of work (h_F, h_M) becomes

$$\begin{aligned}\varphi(h_F, h_M) &= P\left(V(h_F, h_M, I) = \max_{x,y} V(x, y, I)\right) \\ &= \frac{Q_F(h_F)Q_M(h_M)\exp(u(f(w_F h_F, w_M h_M, I), h_F, h_M))}{\sum_{x,y} Q_F(x)Q_M(y)\exp(u(f(w_F x, w_M y, I), x, y))}.\end{aligned}\tag{4.4}$$

Unfortunately, we do not have information on the number of jobs with hours of work h_F and h_M , so they have to be estimated simultaneously with the preference parameters. With no loss of generality we can normalize such that the number of nonworking opportunities ($h = 0$) is 1, that is, $Q_F(0) = Q_M(0) = 1$. Furthermore, for $h > 0$, we can write $Q(h) = \theta g(h)$, where $\theta = \sum_{x \in D \setminus \{0\}} Q(x)$. The term θ can be interpreted as a measure of the relative size of the job market opportunity set compared to non-working opportunities, while $g(h)$ is the share of available jobs with the specified working hours h , which can be interpreted as the opportunity density:

$$g(h) = \frac{Q(h)}{\sum_{x \in D \setminus \{0\}} Q(x)}.$$

Since our households are married couples, we define θ and $g(h)$ for both spouses. The opportunity density is assumed to be uniform for nonstandard working time requirements with a part-time peak and a full-time peak, as observed in the data. Because nearly all observed males are employed, the measure of potential job opportunities for males, θ_M , is not identified and is therefore normalized to one. For females, θ_F , we adopt the specification of Dagsvik and Jia (2016), in which the logarithm of the job availability measure (θ) is modeled as a linear function of years of schooling, as follows:

$$\log \theta_F = \gamma_{F1} + \gamma_{F2}S,\tag{4.5}$$

where S is the length of education and γ_{F1} and γ_{F2} are parameters to be estimated. This implies that we assume a higher level of education leads to greater job opportunities for an individual.

4.2.3 Further model specification and estimation results

The model specified in Equation (4.4) is estimated using Maximum Likelihood within a Conditional Logit framework, where spouses choose among 56 possible combinations of job and non-market alternatives for the couple – comprising seven discrete options for men and eight for women.¹⁶ Dagsvik and Jia (2016) discuss in detail the identification conditions and the estimation method given that cross-sectional micro data are employed. To estimate this model, we further specify $\phi(h_F, h_M | w_F, w_M, I)$, the joint density of hours of work for the spouses in the household, given wage rates and non-labor income. The empirical counterpart of this density is the fraction of couples where the husband works h_M hours and the wife works h_F hours, within the subpopulation of

¹⁶As there is no support in data for the non-market alternative for males, there are only seven alternatives for males.

couples with wage rates and non-labor income equal to (w_F, w_M, I) . We can then write:

$$\psi(h_F, h_M | w_F, w_M, I) = v(f(h_F w_F, h_M w_M, I), h_F, h_M), \quad (4.6)$$

where $\psi(\cdot) = v(\cdot)$ is now to be interpreted as the representative utility of jobs with hours of work (h_F, h_M) . We then obtain the conditional joint density of (h_F, h_M) , which, given that $h_M > 0$, equals

$$\phi(h_F, h_M | w_F, w_M, I) = \frac{\psi(h_F, h_M | w_F, w_M, I) Q_F(h_F) Q_M(h_M)}{P}, \quad (4.7)$$

for $h_F > 0, h_M > 0$, and

$$\phi(0, h_M | w_F, w_M, I) = \frac{\psi(0, h_M | w_F, w_M, I) Q_M(h_M)}{P}, \quad (4.8)$$

for $h_F = 0$ and $h_M > 0$, where

$$P = \sum_{y>0} \psi(0, y | w_F, w_M, I) Q_F(0) Q_M(y) + \sum_{x,y>0} \psi(x, y | w_F, w_M, I) Q_F(x) Q_M(y). \quad (4.9)$$

Furthermore, we specify the deterministic part of the utility (in Equation (4.3)), adopting a Box-Cox specification of the form:

$$\begin{aligned} \log u(C, h_F, h_M) = & \alpha_2 \left(\frac{[10^{-4}(C - C_0)]^{\alpha_1} - 1}{\alpha_1} \right) + \left(\frac{(L_F)^{\alpha_3} - 1}{\alpha_3} \right) \beta_F \\ & + \left(\frac{(L_M)^{\alpha_4} - 1}{\alpha_4} \right) \beta_M + \alpha_{15} \left(\frac{(L_M)^{\alpha_4} - 1}{\alpha_4} \right) \left(\frac{(L_F)^{\alpha_3} - 1}{\alpha_3} \right), \end{aligned} \quad (4.10)$$

where

$$\beta_F = \alpha_5 + \alpha_6 \log A_F + \alpha_7 (\log A_F)^2 + \alpha_8 CU6 + \alpha_9 CO6$$

and

$$\beta_M = \alpha_{10} + \alpha_{11} \log A_M + \alpha_{12} (\log A_M)^2 + \alpha_{13} CU6 + \alpha_{14} CO6.$$

In this specification, C_0 is the subsistence level, A_F and A_M are age (divided by 10), $CU6$ and $CO6$ are the number of children up to six year of age (start of school), and over six, respectively. C is given by the budget constraint in the same way as seen in Equation (4.2), i.e., $C = f(h_F w_F, h_M w_M, I)$, L_F and L_M represent leisure, and $\alpha_n (n = 1, 2, \dots, 15)$ are unknown parameters. Note that we have subtracted from total annual hours a “subsistence” level, $L_0 = 5,110$ hours, which allows for sleep and rest, corresponding to about 14 hours per day. C_0 is set to NOK 64,000, whereas disposable income, C , is measured as the sum of the post-tax annual household wage income, post-tax household capital income, and child allowances.

The contribution to the log likelihood function for couple i with observed hours of work (h_{iF}, h_{iM}) , wage rate (h_{iF}, h_{iM}) and non-labor income I_i is simply $\log \phi(h_{iF}, h_{iM} | w_{iF}, w_{iM}, I_i)$ given by Equation (4.7) or Equation (4.8) depending on whether the female is working or not. In other words, the log

likelihood function for the labor supply model is given by

$$\sum_i \log \phi(h_{iF}, h_{iM} | w_{iF}, w_{iM}, I_i). \quad (4.11)$$

Estimation results are given in Section C of the Appendix, in Table C2. In this table, we highlight that for women, the age (both the $\log(\text{age}/10)$ and $\log(\text{age}/10)^2$) and the number of children (both below and above 6 years old) variables are the significant regressors capturing the heterogeneity of preferences for leisure. Being a woman and being with children is associated with greater leisure preference.

4.3 Using the job choice model to obtain estimates of change in welfare

4.3.1 Obtaining estimates of CV

We next explain how the estimated parameters of the job choice model are used to simulate the welfare effects of a tax reform. We begin by outlining the procedures used to derive CV-based measures, before turning to the estimation of the CE-based measure in Section 4.3.3.

As explained in the Appendix, Section D, the calculation of CV in the job choice model can be done in the same way as in the conventional discrete labor supply model, but with the addition of a term representing job options, $\log(Q_F(h_F))$ and $\log(Q_M(h_M))$, as highlighted in Equation (4.3). This term may depend on individual characteristics, such as education. Thus, the job choice model accounts for heterogeneity in both preferences and circumstances, where the latter includes heterogeneity in labor market opportunities.

Consider now a policy change in the budget set and the opportunity measure where the ex-ante household consumption function, f_0 , is changed to the ex-post function, f_1 . Based on the same notation as in Equation (4.3), the conventional CV for household i is defined by

$$\begin{aligned} & \max_{h_F, h_M} (u_i(f_0(w_{iF}h_F, w_{iM}h_M, I_i), h_F, h_M) + \log(Q_{iF}(h_F)) + \log(Q_{iM}(h_M)) + \eta_i(h_F, h_M)) = \\ & \max_{h_F, h_M} (u_i(f_1(w_{iF}h_F, w_{iM}h_M, I_i) + CV_i, h_F, h_M) + \log(Q_{iF}(h_F)) + \log(Q_{iM}(h_M)) + \eta_i(h_F, h_M)), \end{aligned} \quad (4.12)$$

where index i refers to household-specific covariates that are suppressed in the notation. This definition of CV is analogous to the standard definition, see McFadden (1999). It follows that CV is stochastic because it depends on all stochastic terms, $\{\eta_i(h_F, h_M)\}$, and that it incorporates representations of labor market restrictions, $\log Q_{ik}(h_k)$. Section D.1 of the Appendix provides further details on how measures of CV can be obtained under the simulation approach of McFadden (1999).

4.3.2 Estimates of reference preference CV

As discussed in the Introduction, in order to obtain a fairness compliant welfare measure, we redefine Equation (4.12) to compute CV when controlling for individual heterogeneity in preferences,

denoted CV^{circ} . This means that we obtain estimates of CV_i when both the deterministic part, $u(C, h_F, h_M)$, and the random part, $\eta(h_F, h_M)$, of the utility function are identical across individuals, symbolized by replacing subscript i in u_i and η_i by common reference values (across individuals), indicated by $\bar{u}$ and $\bar{\eta}$.¹⁷ We obtain a measure of CV in which preference heterogeneity has been eliminated, CV^{circ} , by

$$\begin{aligned} \max_{h_F, h_M} (\bar{u}(f_0(w_{iF}h_F, w_{iM}h_M, I_i), h_F, h_M) + \log(Q_{iF}(h_F)) + \log(Q_{iM}(h_M)) + \bar{\eta}(h_F, h_M)) \\ = \max_{h_F, h_M} (\bar{u}(f_1(w_{iF}h_F, w_{iM}h_M, I_i) + CV_i^{circ}, h_F, h_M) \\ + \log(Q_{iF}(h_F)) + \log(Q_{iM}(h_M)) + \bar{\eta}(h_F, h_M)). \quad (4.13) \end{aligned}$$

In practice, a common deterministic utility function, $\bar{u}$, is obtained by using common taste-modifying variables. This means that the taste-modifying variables – gender, age and number of children – are set to their median values. Other parameters of the utility function, $\alpha_1, \alpha_2, \alpha_3$, and α_4 , do not vary across individuals (see deterministic part of the utility function in Section 4.2). An alternative approach is to neutralize differences in circumstances rather than preferences; Section D.2 of the Appendix details the implementation of this CV^{pref} alternative.

4.3.3 Estimates of welfare change based on the CE criterion

We now assess the welfare effects of tax reforms using the CE criterion, employing an empirical implementation aligned with that in Carpentier and Sapata (2016). However, our empirical specification differs slightly, reflecting the characteristics of the labor supply data used in our study.

As defined by Equation (3.3) in Section 3, CE denotes the maximum reference utility achievable within the equivalent budget set. We now explain how this criterion can be used to compute $\Delta CE_i = CE_{1i} - CE_{0i}$, employing the same labor supply model used to derive CV_i^{circ} .

In Section 3, we established that the empirical challenge for measuring welfare according to the CE criterion is to compute the maximum well-being that a household can reach, given its equivalent budget set respecting individual preferences. Carpentier and Sapata (2016) employ estimated preferences from a random utility model and leverage observed individual choices as an additional source of information to compute welfare measures that respect individual preferences.

Here, we do not exploit information about actual choices.¹⁸ Instead, we randomly draw utility error terms and assume that households select the alternative j^* (the discrete choice) that maximizes its “true” utility. For each draw, k , we first determine the consumption level of household i for each alternative j , say $\tilde{C}_{ij}^k$, by solving for the consumption level that yields the same utility as the optimal option and thereby tracing out the indifference curve corresponding to the optimal choice. We then compute the lump-sum tax, T_i^k , using Equation (3.2). Subtracting this tax from gross labor income

¹⁷See Roemer and Trannoy (2016) for a discussion on how the error term is interpreted in empirical studies of equality of opportunity – specifically, whether it is attributed to the domain of “effort” or “circumstances”.

¹⁸When applied in simulations, the model does not incorporate information on individuals’ observed choices from data.

for each alternative, $w_i h_j$, gives the equivalent consumption, $C_{ij}^k (= w_i h_j - T_i^k)$, which defines the equivalent budget set in Equation (3.3). Confer Figure A3 in Appendix A.2 for further illustration of the method.

The welfare level corresponding to the CE criterion is accounted for by introducing reference preferences in the specification symbolized by $\bar{\gamma}$ in Equation (3.3). To ensure comparability across methods, we use the same function, $\bar{u}(c, h; \bar{\gamma})$, for the deterministic part of the labor supply model under reference preference utility as in the approach to obtain CV_i^{circ} . This means that the taste-modifying variables – age, gender, and number of children – are set to their median values.

Applying the reference utility function to this equivalent budget, we identify the maximum utility value as the CE_i^k , given the draw. We average across all draws (k) to obtain CE_i . The procedure is implemented under both pre- and post-reform tax rules to derive $CE_{0,i}$ and $CE_{1,i}$ – the same random draws are used pre- and post-reform to ensure consistency. Recall that the welfare effect for household i is defined as the difference between these two: $\Delta CE_i = CE_{1,i} - CE_{0,i}$.

5 Describing distributional effects of tax reform by welfare measures reflecting fairness

5.1 The Norwegian tax reform 2013–2019

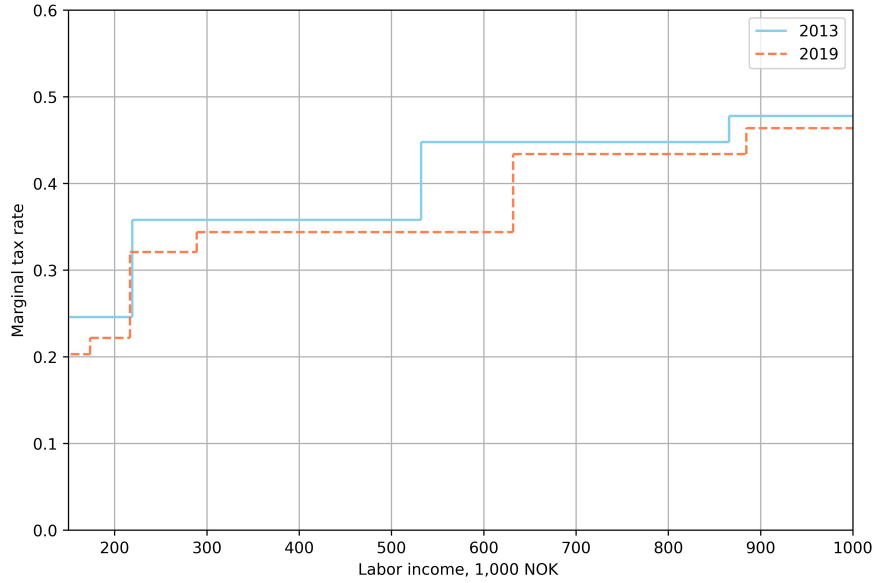
Next, we demonstrate the empirical implications of using fairness-adjusted welfare measures, rather than describing the reform solely in terms of the standard CV. To illustrate, we utilize the introduction of the so-called bracket tax from the Norwegian tax reform of 2013–2019. The Norwegian tax reform 2013–2019 followed two major reforms of the tax system in 1992 and 2006. The former introduced the dual income tax system in Norway, while the latter maintained the system with important modifications (Sørensen 2005). A dual income tax combines a low proportional tax rate on capital income and progressive tax rates on labor income.

The main element of the 2013–2019 reform was a gradual lowering of the corporate tax rate. Due to the link between the corporate tax rate and the basic tax rate on general income of persons, following from the design of the Norwegian version of the dual income tax, the immediate consequence was a cut in the lower flat-rate part of the tax on labor earnings, down from 28 percent in 2013 to 22 percent in 2019. As this tax rate cut obviously implied a significant loss of tax revenue, a major challenge was to offset (at least part of) the foregone revenue. A main move has been the introduction of additional brackets in the progressive taxation of labor income, replacing the previous two-tier surtax system. To distinguish the new schedule from the old one, the term “bracket tax” was adopted to reflect the larger number of brackets in the new piecewise linear income tax. We use the introduction of the bracket tax to illustrate how our fairness-adjusted measures of changes in welfare can be applied to assess the effects of tax reforms.

Figure 3 compares the marginal tax rates on individual income under the 2013 and 2019 tax schemes. Importantly, Norway applies individual taxation, meaning that each spouse is taxed separately and household income is not jointly assessed. The figure illustrates that there are two addi-

tional brackets in 2019, compared to 2013, and that the 2019-rates in general are lower than those of 2013. The first bracket generates revenue at lower levels by an addition of 1.9 percentage points, whereas the next step (2.3 percentage points) kicks in when the basic allowance is exhausted, which brings the marginal tax rate in the interval from approximately 300,000 Norwegian kroner (NOK)¹⁹ to around NOK 500,000 very close to the pre-reform scheme, see Figure 3. The two last steps basically replicate the two-tier surtax scheme of the pre-reform system, with somewhat lower rates.

Figure 3: Marginal tax rates on individual labor income, 2013 and 2019



5.2 Fairness considerations and distributional effects

5.2.1 Aggregate welfare effects

We apply our three money metric measures CV , CV^{circ} , and the one based on the CE criterion, ΔCE , to evaluate welfare changes along the household income distribution consecutive to the introduction of the bracket tax reform over 2013–2019.²⁰

We first present aggregate measures. Since the bracket tax reduced the marginal tax rates of labor income, as seen in Figure 3, we expect the welfare effects for most households to be positive, in the sense that the tax reductions make most taxpayers better off after the reform.²¹ This is, in a nutshell, what we find, as can be seen in Table 1, which presents the average welfare effects for all three measures. However, since the ΔCE criterion employs a different metric of welfare measurement, only the estimates derived from CV and CV^{circ} criteria are directly comparable. Nonetheless, the key insight from Table 1 is that all three methods suggest increased welfare by the reform. We will

¹⁹This corresponds to approximately 30,500 euros and 34,000 US dollars according to average exchange rates for 2019 (1 USD = NOK 8.80 and 1 EUR = NOK 9.85).

²⁰Importantly, the same framework can be used to inform policy-making in discussions of future policies.

²¹Estimates of CV become negative given the definition of compensating variation. Since we consider a price (tax) reduction, we may interpret CV as willingness to pay to avoid going back to the original level of utility. Nevertheless, we measure CV in positive values.

return later to how to make the results comparable across all three methods. The average estimates of CV and CV^{circ} are close, around NOK 18,500. Eliminating heterogeneity in preferences leads to a (slight) reduction in the spread of the welfare effects across households, as can be seen from the estimates of the standard deviations of the welfare effects in Table 1.

Table 1: Average welfare effects from the bracket reform tax under the three measurement alternatives

Empirical method	Welfare effect	Standard deviation
CV (NOK)	18,384	5,458
CV^{circ} (NOK)	18,677	5,188
ΔCE	0.356	0.108

Notes: Simulation results are derived by obtaining estimates of CV and CV^{circ} from discrete choice labor supply model simulations, with CV^{circ} based on neutralized preference heterogeneity. The ΔCE criterion is based on welfare comparisons using reference values of responsibility characteristics, calculated both before and after the reform.

Table 2: Transition matrix of household ranks for two methodologies describing welfare effects of the bracket tax reform by CV vs. CV^{circ}

CV^{circ}	CV				
	Bottom 20%	20%–40%	40%–60%	60%–80%	Top 20%
Bottom 20%	84.9	13.7	1.4	0.0	0.0
20%–40%	15.1	70.8	13.5	0.6	0.0
40%–60%	0.0	15.4	72.8	11.8	0.0
60%–80%	0.0	0.1	12.5	80.1	7.5
Top 20%	0.0	0.0	0.0	7.5	92.5

Notes: CV and CV^{circ} simulation results are derived from estimates of compensating variation obtained via discrete-choice labor supply model simulations. CV^{circ} is based on neutralized preference heterogeneity.

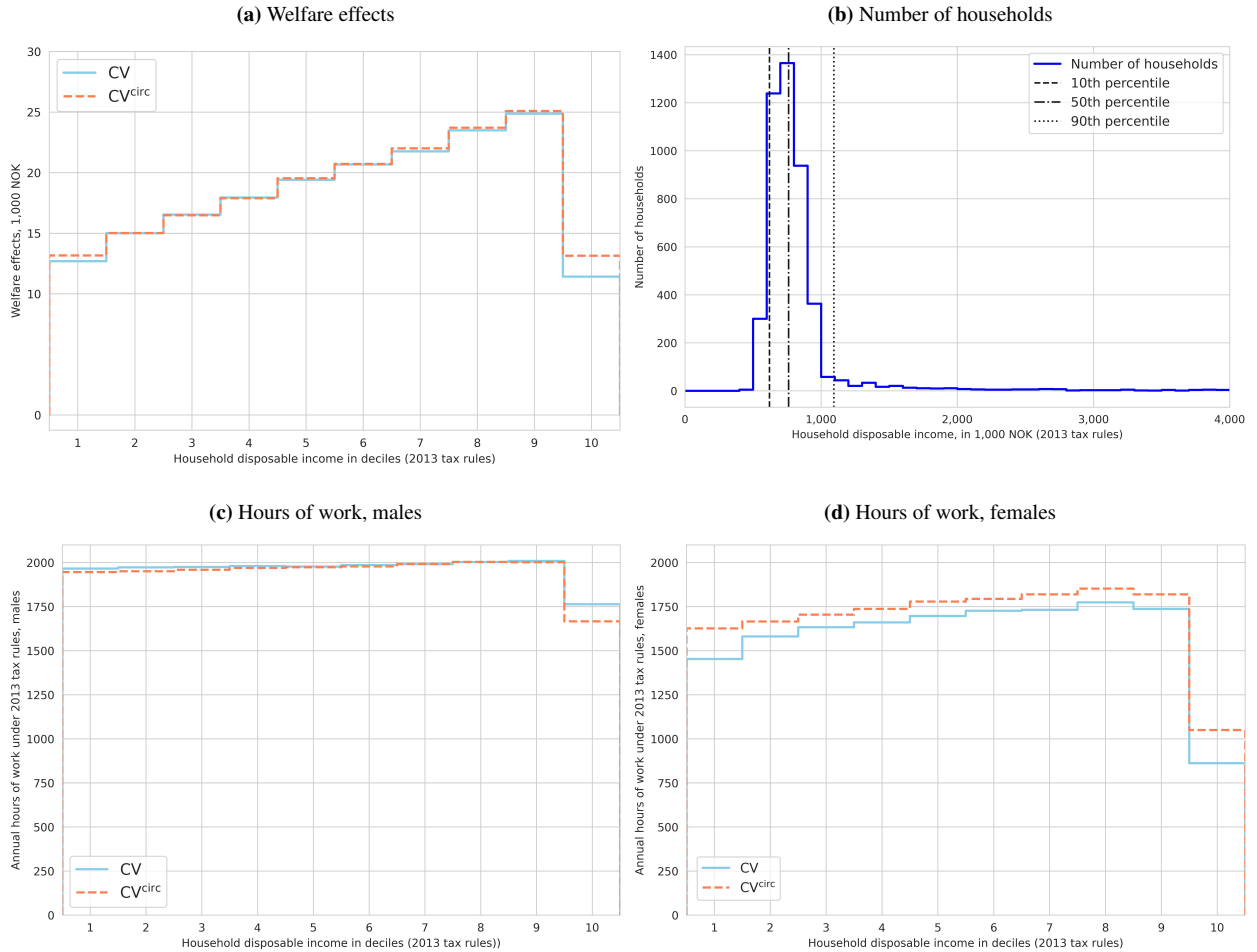
Furthermore, Table 2 suggests that the choice of measurement alternative has limited effects on household rankings due to welfare effects of the reform. Table 2 displays a transition matrix of household ranks based on quintile group categorizations according to CV vs. CV^{circ} . One can clearly see that the vast majority of households, ranging from 71 to 93 percent, remains within the same quintile group regardless of the method employed. Moreover, we have verified that the patterns in the two alternative transition matrices – CV vs. ΔCE and CV^{circ} vs. ΔCE (available upon request) – closely mirror those presented in Table 2.²²

²²Building on the CE criterion approach of Carpentier and Sapata (2016), which focuses on the 10 percent worst-off, we created a Venn diagram specifically for this group. The diagram (available upon request) reveals that about 60% of couples within this low-income group remain in this group regardless of which of the three measurement alternatives is used.

5.2.2 CV vs CV^{circ} across the income distribution of households

Now, returning to our main question, we present a graphical comparison of the results obtained using CV^{circ} and the conventional CV. The top left panel of Figure 4, Panel (A), describes the distribution of the welfare gains from the reform obtained with each metric, when households are ranked according to their disposable income (in thousands of NOK). This panel shows that CV and CV^{circ} follow the same pattern along the household income distribution. The welfare first increases up to decile group 9 (who has NOK 1 million – i.e., USD 114,000; EUR 102,000 – in disposable income), where it reaches approximately NOK 25,000 (USD 2,800; EUR 2,500). This increase is followed by a significant drop in the tenth decile group, according to both measures. Moreover, this group is the only one where the two metrics present significantly distinct values. This suggests that the upper decile of the disposable income distribution is the only one where responsibility matters in welfare measurement, in the context of this reform.

Figure 4: Welfare effects of the bracket tax reform: (individual-specific) CV vs CV^{circ}



Notes: The top left panel, Panel (A), describes the distribution of welfare effects for different levels of household disposable income. CV simulation results are obtained by a standard discrete choice labor supply model, whereas CV^{circ} simulation results are obtained by a model in which preference heterogeneity is neutralized. The top right panel, Panel (B), shows the number of households for different levels of household disposable income. The bottom panels display the distribution of annual hours worked under the two methods given the 2013 tax rules (i.e., before the tax reform): Panel (C) for males and Panel (D) for females.

To better understand these results, we report, in Panels (C) and (D) of Figure 4, the distributions of annual hours worked under CV and CV^{circ} , under the 2013 tax system. We report in these panels the average hours of work for men and women, respectively, by decile group of household disposable income. This disposable income is the one we simulate using the job choice model under the 2013 tax rules (i.e., before the tax reform). The solid blue lines plot average simulated hours of work, in each decile group, obtained using each individual's own preference parameters. These correspond to the labor supply choices made under the CV criterion. By contrast, the red dashed lines reflect the labor decisions when individual preference parameters are replaced, under CV^{circ} , by the reference ones, i.e., the median values of the taste-modifying variables and a common error term (see Appendix, Section D.2).²³

Panels (C) and (D) show that imposing reference preferences results in markedly different impacts on male and female labor supply: while female labor supply in general increases under CV^{circ} , only high-income male labor supply is affected, but decreases. This divergence is explained by differences in parameter estimates between males and females, as reported for leisure in Table C2. Although the labor supply model involves complex interactions, we attribute the relatively small effect of fairness neutralization on males – and the larger effect on females – primarily to differences in parameter estimates, specifically the gap between α_3 and its interaction with the taste-modifying variables and α_4 and its corresponding interaction.

As a result, men in all decile groups except the top one supply the same number of labor hours under both CV and CV^{circ} . Therefore, the welfare differences observed between CV and CV^{circ} – particularly in decile 10 of Panel (A) – are not driven by men's behavior or characteristics but by the ones of women. Notably, the larger welfare gain in decile group 10 under CV^{circ} compared to CV occurs despite males in the top decile group working fewer hours than those in other decile groups. Under CV^{circ} , these men further reduce their labor supply compared to CV, suggesting they benefit less from a tax reform that increases the returns to work.

Despite this, according to CV^{circ} , we observe greater welfare gains in decile group 10 following the tax reform, compared to CV (see Panel A). Unlike their male counterparts, women adopt stronger work preferences under CV^{circ} , leading to an increase in labor supply across the income distribution.²⁴ Consequently, the differences shown in Panel (A) primarily reflect how women reassess their labor-leisure trade-offs under CV^{circ} . Panel (D) demonstrates that women in decile group 10 exhibit a larger (hypothetical) increase in labor supply than those in other decile groups. CV^{circ} implies lower marginal utility to leisure for women who initially exhibited a stronger preference for it. As a result, female labor supply rises mechanically across the income distribution under CV^{circ} relative to CV. The change in labor behavior has largest welfare effects for households at the top end of the income distribution, where pre-reform female labor hours were relatively low.

We can therefore conclude that, for the vast majority of Norwegian households, the reform has

²³Recall that the taste-modifying variables include log age, log age squared, number of children below and above 6 years old. We also perform sensitivity tests using alternative median values for the above variables, such as the medians calculated for the 10th and 90th percentiles, and the results remain robust.

²⁴This implies that the adoption of median taste-modifying variables leads to an overall increase in average working hours.

resulted in significant welfare gains, regardless of whether preferences have been neutralized or not. In other words, estimating household welfare impacts using CV or CV^{circ} does not yield large differences. Consequently, controlling for preferences in the way we have done has limited effects. Of course, this conclusion depends on the particular reform used in our empirical illustration. In the following section, we confirm that this conclusion also holds under CE.

In summary, the bracket tax reform results in welfare gains for all households in Norway, and these gains are not mainly driven by characteristics related to individual responsibility. This is evident from the small differences observed between welfare effects computed under CV and those computed under CV^{circ} . Only the upper tail of the household income distribution is affected, CV^{circ} displaying there larger welfare gains than the standard CV. This divergence arises because CV is based on choices influenced by individual differences in preferences for leisure, whereas CV^{circ} uses uniform reference preferences, thereby removing this heterogeneity.

5.2.3 Robustness check: CE criterion across the income distribution of households

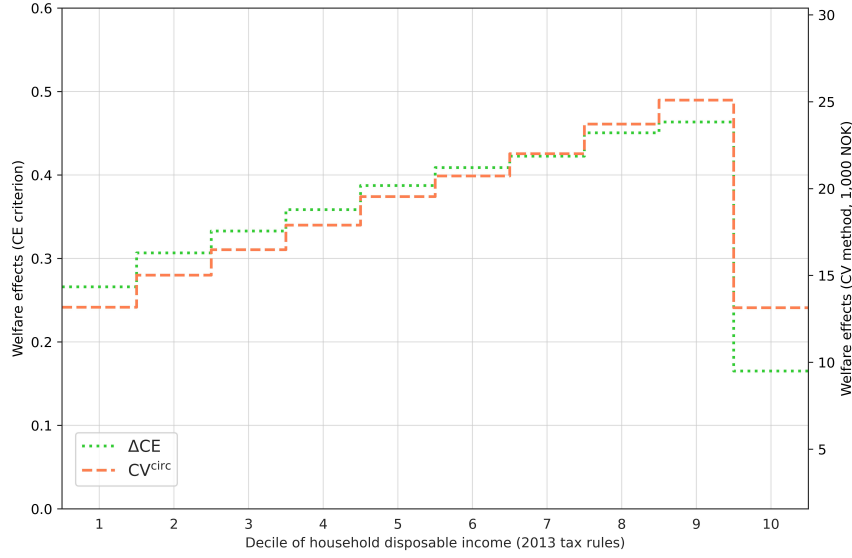
Finally, we describe the welfare effects of the bracket tax reform under ΔCE . Given that ΔCE shares key similarities with CV^{circ} , we expect the distributions of the welfare effects along the household income distribution to closely align with those observed in Figure 4, Panel (A). This is confirmed in Figure 5, where we plot the distributions of welfare effects obtained under ΔCE and CV^{circ} . As explained in Section 4.3.3, welfare effect estimates based on the CE criterion are derived by examining the difference between individual welfare measures before and after the reform. It follows that ΔCE is measured in another unit and scale than the measures of change in money metric utility like CV^{circ} . The notes under Figure 5 describe how results for ΔCE are scaled to ensure comparability across methods. The right-hand axis of the figure displays the effects in terms of ΔCE , converted to be comparable to the other measure reflecting fairness, CV^{circ} .

It is reassuring that the distributional welfare effects under ΔCE largely mirror those observed according to CV^{circ} – the alignment is remarkably close: both ΔCE and CV^{circ} yield similar conclusions across income decile groups 1 through 9. The differences are most pronounced at the top of the income distribution, where ΔCE indicates smaller welfare gains. However, we interpret this as a form of methodological noise, given the distinct estimation techniques involved. In particular, as discussed, CE-based results require a transformation procedure to align them with the CV^{circ} outcomes, as illustrated in Figure 5 (see the notes). What matters most is that both CV^{circ} and ΔCE point out that responsibility matters in the same decile as when comparing the difference between CV and CV^{circ} .

6 Conclusion

In this paper, we develop an empirical method to take into account the importance, in the distribution of household welfare, of characteristics for which individuals are responsible. This method relies on a new welfare metric, CV^{circ} , which is defined as the amount of money needed to restore a

Figure 5: Welfare effects of the introduction of the bracket tax reform: CV^{circ} vs ΔCE



Notes: This diagram illustrates the distribution of welfare effects measured by ΔCE contrasted to CV^{circ} , across decile groups of household disposable income. The left vertical axis indicates the units for ΔCE , while the right vertical axis shows the units for the measures obtained with CV^{circ} (the same as in Figure 4, Panel (A)). The scales of the two axes are adjusted so that the measures are comparable. Let μ_1 and σ_1 denote the sample mean and standard deviation of ΔCE , and μ_2 and σ_2 the corresponding values for CV^{circ} . In the diagram, the transformation applied to the right axis is given by $right = left \times \frac{\sigma_2}{\sigma_1} + \mu_2 - \frac{\sigma_2}{\sigma_1} \mu_1$.

household's utility after a tax reform, assuming households differ only in circumstances. By comparing CV^{circ} to the classic CV metric, one can conclude that responsibility matters whenever the two metrics display significantly different values. CV being well-established, its variant CV^{circ} can be seen as an intuitive and pedagogical criterion for assessing the fairness impact of a tax reform.

To build our two metrics, we first estimate a structural discrete choice labor supply model, the job choice model, which incorporates individual circumstances under the form of job opportunities. We then use the parameter estimates in a simulation approach à la McFadden (1999) to derive our measures of CV and CV^{circ} . To relate our new measure to the existing literature on fairness and the circumstances/responsibility dyad, we also compute the variation of the conditional equality criterion (CE).

We implement our empirical method by examining the distributional welfare effects of the 2013–2019 Norwegian bracket tax reform. CV^{circ} is computed like CV, with two key differences to ensure that individuals differ only in circumstances: First, the taste-modifying variables of individuals are set to their median values. Second, all households are assumed to have the same random utility component. Comparing the welfare effects obtained with CV to those obtained with CV^{circ} reveals that differences in the post-reform household welfare distribution are not primarily attributable to preferences, except at the very top of the household income distribution. Indeed, the two metrics only differ significantly at the top of the household income distribution, a finding driving by the fact that there, women show a strong preference for leisure.

As a robustness check, we compare the shape of the distribution of the welfare effects obtained with CV^{circ} to those obtained with CE. Both methods yield very similar patterns, confirming the

closeness of the underlying approaches, despite different theoretical foundations. For future research, CV^{circ} could be used with other tax reforms and to build up indicators (in the vein of the MVPF) when the policymakers do not want to compensate for preferences.

References

- Aaerge, R., J. Dagsvik, and S. Strøm (1995). Labor supply responses and welfare effects of tax reforms. *Scandinavian Journal of Economics* 97(4), 635–659.
- Aaerge, R., U. Colombino, and S. Strøm (2004). Do more equal slices shrink the cake? An empirical investigation of tax-transfer reform proposals in Italy. *Journal of Population Economics* 17(4), 767–785.
- Alesina, A., S. Stantcheva, and E. Teso (2018). Intergenerational mobility and preferences for redistribution. *American Economic Review* 108(2), 521–554.
- Almås, I., A. W. Cappelen, J. T. Lind, E. Ø. Sørensen, and B. Tungodden (2011). Measuring unfair (in) equality. *Journal of Public Economics* 95(7-8), 488–499.
- Arneson, R. (1989). Equality and equal opportunity for welfare. *Philosophical Studies* 56, 77–93.
- Bargain, O., A. Decoster, M. Dolls, D. Neumann, A. Peichl, and S. Siegloch (2013). Welfare, labor supply and heterogeneous preferences: Evidence for Europe and the US. *Social Choice and Welfare* 41(4), 789–817.
- Berg, K. and P. G. Piacquadio (2023). Fairness and Paretian social welfare functions. Manuscript, University of Oxford.
- Bhattacharya, D. (2024). Nonparametric approaches to empirical welfare analysis. *Journal of Economic Literature* 62(2), 554–593.
- Blackorby, C. and D. Donaldson (1988). Money metric utility: A harmless normalization. *Journal of Economic Theory* 46(1), 120–129.
- Blundell, R. (2012). Tax policy reform: The role of empirical evidence. *Journal of the European Economic Association* 10(1), 43–77.
- Blundell, R. and T. MaCurdy (1999). Labor supply: A review of alternative approaches. In: *Handbook of Labor Economics*. Ed. by O. Ashenfelter and D. Card. Vol. 3, Part A. Elsevier. Chap. 27, 1559–1695.
- Bosmans, K., K. Decancq, and E. Ooghe (2018). Who’s afraid of aggregating money metrics? *Theoretical Economics* 13(2), 467–484.
- Bossert, W. and M. Fleurbaey (1996). Redistribution and compensation. *Social Choice and Welfare* 13, 343–355.
- Burtless, G. and J. A. Hausman (1978). The effect of taxation on labor supply: Evaluating the Gary negative income tax experiment. *Journal of Political Economy* 86(6), 1103–1130.
- Capéau, B., A. Decoster, and L. De Sadeleer (2023). Interpersonal comparisons by means of money metric utilities: Why one should use the same reference prices for all. *Journal of Income Distribution* (3-4), 205–236.

- Capéau, B., S. Maes, L. De Sadeleer, and A. Decoster (2021). Nonparametric welfare analysis for discrete choice: Levels and differences of individual and social welfare. CESifo Working Paper No. 9071. Munich.
- Cappelen, A. W., A. D. Hole, E. Ø. Sørensen, and B. Tungodden (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review* 97(3), 818–827.
- Carpantier, J. and C. Sapata (2016). Empirical welfare analysis: When preferences matter. *Social Choice and Welfare* 46(3), 521–542.
- Cohen, G. (1989). On the currency of egalitarian justice. *Ethics* 99(4), 906–944.
- Creedy, J. and G. Kalb (2005). Discrete hours labour supply modelling: Specification, estimation and simulation. *Journal of Economic Surveys* 19(5), 697–734.
- Dagsvik, J. K. (1994). Discrete and continuous choice, max-stable processes, and independence from irrelevant attributes. *Econometrica* 62(5), 1179–1205.
- Dagsvik, J. K. and Z. Jia (2016). Labor supply as a choice among latent jobs: Unobserved heterogeneity and identification. *Journal of Applied Econometrics* 31(3), 487–506.
- Dagsvik, J. K., Z. Jia, T. Kornstad, and T. O. Thoresen (2014). Theoretical and practical arguments for modeling labor supply as a choice among latent jobs. *Journal of Economic Surveys* 28(1), 134–151.
- Dagsvik, J. K. and A. Karlström (2005). Compensating variation and Hicksian choice probabilities in random utility models that are nonlinear in income. *Review of Economic Studies* 72(1), 57–76.
- Dagsvik, J. K., M. Locatelli, and S. Strøm (2009). Tax reform, sector-specific labor supply and welfare effects. *Scandinavian Journal of Economics* 111(2), 299–321.
- Dagsvik, J. K., S. Strøm, and M. Locatelli (2021). Marginal compensated effects in discrete labor supply models. *Journal of Choice Modelling* 41.
- Decancq, K., M. Fleurbaey, and E. Schokkaert (2015). Inequality, income, and well-being. In: *Handbook of Income Distribution*. Ed. by A. B. Atkinson and F. Bourguignon. Vol. 2. Elsevier, 67–140.
- Decoster, A. and P. Haan (2015). Empirical welfare analysis with preference heterogeneity. *International Tax and Public Finance* 22(2), 224–251.
- Devooght, K. (2008). To each the same and to each his own: A proposal to measure responsibility-sensitive income inequality. *Economica* 75, 280–295.
- Dworkin, R. (1981). What is equality? Part 1: Equality of welfare. *Philosophy and Public Affairs* 10, 185–246.
- Fleurbaey, M. and D. Blanchet (2013). *Beyond GDP. Measuring Welfare and Assessing Sustainability*. Oxford University Press.
- Fleurbaey, M. (2008). *Fairness, Responsibility and Welfare*. Oxford: Oxford University Press.
- Fleurbaey, M. and F. Maniquet (2011). *A Theory of Fairness and Social Welfare*. Cambridge University Press.
- Fleurbaey, M. and F. Maniquet (2018). Optimal income taxation theory and principles of fairness. *Journal of Economic Literature* 56(3), 1029–1079.

- Hausman, J. (1979). The econometrics of labor supply on convex budget sets. *Economics Letters* 3(2), 171–174.
- Hendren, N. and B. Sprung-Keyser (2020). A unified welfare analysis of government policies. *The Quarterly Journal of Economics* 135(3), 1209–1318.
- Herriges, J. and C. Kling (1999). Nonlinear Income Effects in Random Utility Models. *The Review of Economics and Statistics* 81(1), 62–72.
- Hicks, J. (1939). The Foundations of Welfare Economics. *The Economic Journal* 49(196), 696–712.
- Hufe, P., R. Kanbur, and A. Peichl (2022). Measuring unfair inequality: Reconciling equality of opportunity and freedom from poverty. *The Review of Economic Studies* 89(6), 3345–3380.
- Jacquet, L. (2014). Tagging and redistributive taxation with imperfect disability monitoring. *Social Choice and Welfare* 42, 403–435.
- Jacquet, L. and D. Van de gaer (2011). A comparison of optimal tax policies when compensation or responsibility matter. *Journal of Public Economics* 95(11), 1248–1262.
- Jia, Z., B. Larsen, B. Lian, R. Nesbakken, O. E. Nygård, T. O. Thoresen, and T. E. Vattø (2024). The LOTTE system of tax microsimulation models. *International Journal of Microsimulation* 17(3), 73–93.
- Kaplow, L. (2008). Optimal policy with heterogeneous preferences. *B.E. Journal of Economic Analysis & Policy* 8(1). Advances, Article 40.
- King, M. (1983). Welfare analysis of tax reforms using household data. *Journal of Public Economics* 21(2), 183–214.
- Koll, D., D. Sachs, F. Stürmer-Heiber, and H. Turon (2024). Quantifying Okun’s leaky bucket: The case of progressive childcare subsidies. Mimeo, prev. CESifo Working Paper No. 10793.
- Kornstad, T. and T. O. Thoresen (2006). Effects of family policy reforms in Norway: Results from a joint labour supply and childcare choice microsimulation analysis. *Fiscal Studies* 27(3), 339–371.
- Lockwood, B. B. and M. Weinzierl (2015). De gustibus non est taxandum: Heterogeneity in preferences and optimal redistribution. *Journal of Public Economics* 124, 74–80.
- Luce, R. (1977). The choice axiom after 20 years. *Journal of Mathematical Psychology* 15, 215–233.
- Marchand, M., P. Pestieau, and M. Racionero (2003). Optimal redistribution when different workers are indistinguishable. *Canadian Journal of Economics* 36(4), 911–922.
- McFadden, D. (1999). Computing willingness-to-pay in random utility models. In: *Trade, Theory and Econometrics: Essays in Honour of John S. Chipman*. Ed. by J. R. Melvin, J. C. Moore, and R. G. Riezman. Routledge. Chap. 15, 253–274.
- McKenzie, L. (1957). Demand theory without a utility index. *The Review of Economic Studies* 24(3), 185–189.
- Ooghe, E., E. Schokkaert, and H. Serruys (2025). Fairness gaps for earnings tax design. *Journal of Political Economy: Microeconomics* 3(1), 25–58.

- Pestieau, P. and M. Racionero (2009). Optimal redistribution with unobservable disability: Welfarist versus non-welfarist social objectives. *European Economic Review* 53(6), 636–644.
- Piacquadio, P. (2017). A fairness justification of utilitarianism. *Econometrica* 85(4), 1261–1276.
- Picchio, M. and G. Valletta (2018). A welfare evaluation of the 1986 tax reform for married couples in the United States. *International Tax and Public Finance* 25(3), 757–807.
- Preston, I. and I. Walker (1999). Welfare measurement in labour supply models with nonlinear budget constraints. *Journal of Population Economics* 12, 343–361.
- Ravallion, M. (2017). Inequality and poverty when effort matters. *Econometrica* 5(4).
- Roemer, J. (1998). *Equality of Opportunity*. Cambridge and London: Harvard University Press.
- Roemer, J., R. Aaberge, U. Colombino, J. Fritzell, S. Jenkins, A. Lefranc, I. Marx, M. Page, E. Pommer, J. Ruiz–Castillo, M. San Segundo, T. Tranaes, A. Trannoy, G. Wagner, and I. Zubiri (2003). To what extent do fiscal regimes equalize opportunities for income acquisition among citizens? *Journal of Public Economics* 87, 539–565.
- Roemer, J. E. and A. Trannoy (2016). Equality of opportunity: Theory and measurement. *Journal of Economic Literature* 54(4), 1288–1332.
- Rowling, J. (1998). *Harry Potter and the Chamber of Secrets*. Bloomsbury edition. Bloomsbury.
- Saez, E. and S. Stantcheva (2016). Generalized social marginal welfare weights for optimal tax theory. *American Economic Review* 106(1), 24–45.
- Samuelson, P. A. (1974). Complementarity: An essay on the 40th anniversary of the Hicks–Allen revolution in demand theory. *Journal of Economic literature* 12(4), 1255–1289.
- Scanlon, T. (2018). *Why Does Inequality Matter?* Oxford University Press.
- Schlee, E. E. and M. A. Khan (2022). Money metrics in applied welfare analysis: A saddlepoint rehabilitation. *International Economic Review* 63(1), 189–210.
- Schokkaert, E., D. Van de gaer, F. Vandenbroucke, and R. I. Luttens (2004). Responsibility sensitive egalitarianism and optimal linear income taxation. *Mathematical Social Sciences* 48, 151–182.
- Slesnick, D. (1998). Empirical approaches to the measurement of welfare. *Journal of Economic Literature* 36(4), 2108–2165.
- Sørensen, P. B. (2005). Neutral taxation and shareholder income. *International Tax and Public Finance* 12(6), 777–801.
- Stantcheva, S. (2026). Perceptions, mindsets and beliefs shaping policy views. *Economica* forthcoming.
- Statistics Norway (2019). Income and wealth statistics for households. Statistics Norway. Statistics Norway.
- Statistics Norway (2024). Labor force survey. Statistics Norway. <https://www.ssb.no/en/arbeid-og-lonn/sysselsetting/statistikk/arbeidskraftundersokelsen>.
- Trautmann, S. T. (2022). Procedural fairness and equality of opportunity. *Journal of Economic Surveys* 37(5), 1697–1714.
- van Soest, A. (1995). Structural models of family labour supply: A discrete choice approach. *Journal of Human Resources* 30(1), 63–88.

Appendix

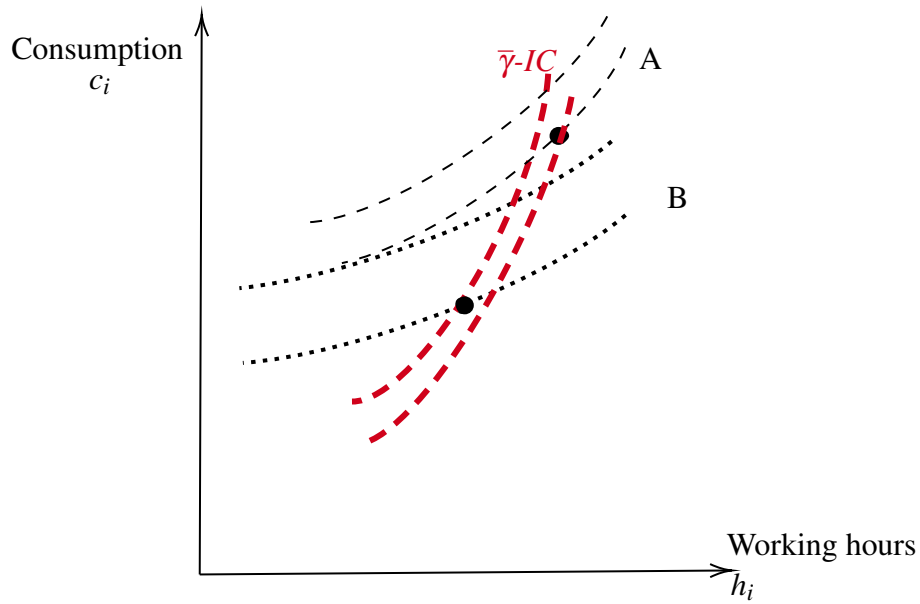
A Reference preferences and graphical construction of CV , CV^{circ} , CE and ΔCE

A.1 Reference preferences

The construction of CV^{circ} relies on the reference value of preferences for leisure/work, $\bar{\gamma}$. In Figures A1 and A2, single-worker households A and B have distinct preferences. Each household chooses a distinct (c_i, h_i) -bundle which is evaluated with $\bar{\gamma}$ reference preferences (dashed and red indifference curves $\bar{\gamma}$ -IC).

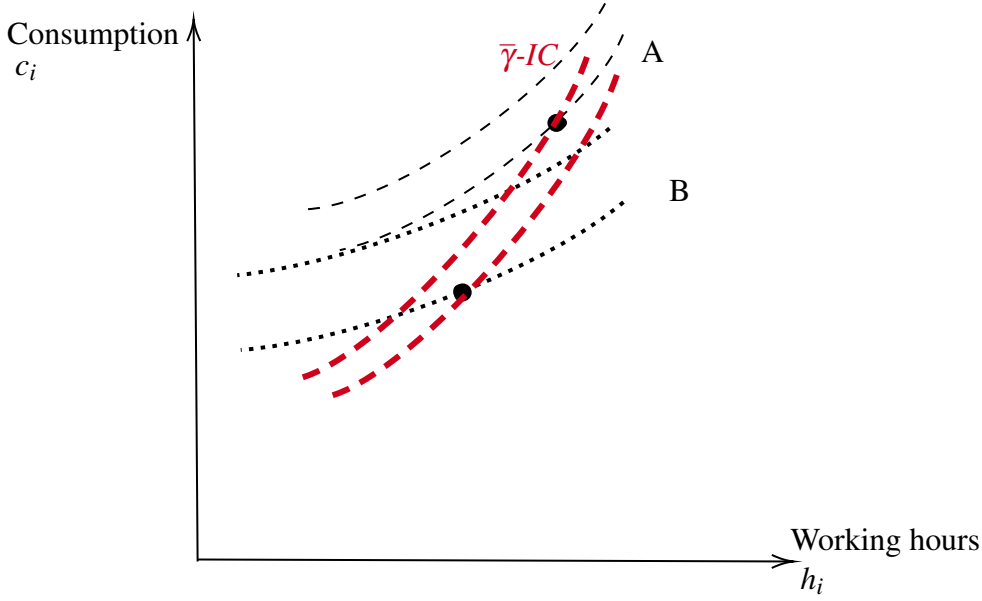
In Figure A1, household B 's bundle gives a larger satisfaction than household A 's bundle, according to $\bar{\gamma}$ reference preferences. In Figure A2, where $\bar{\gamma}$ reference indifference curves are flatter, household A 's bundle gives a larger satisfaction than household B 's bundle.

Figure A1: Choices of A and B evaluated by reference preferences ($\bar{\gamma}$)



Notes: A is worse-off than B according to reference preferences $\bar{\gamma}$ (dashed red $\bar{\gamma}$ -IC).

Figure A2: Choices of A and B evaluated according to reference preferences ($\bar{\gamma}$)



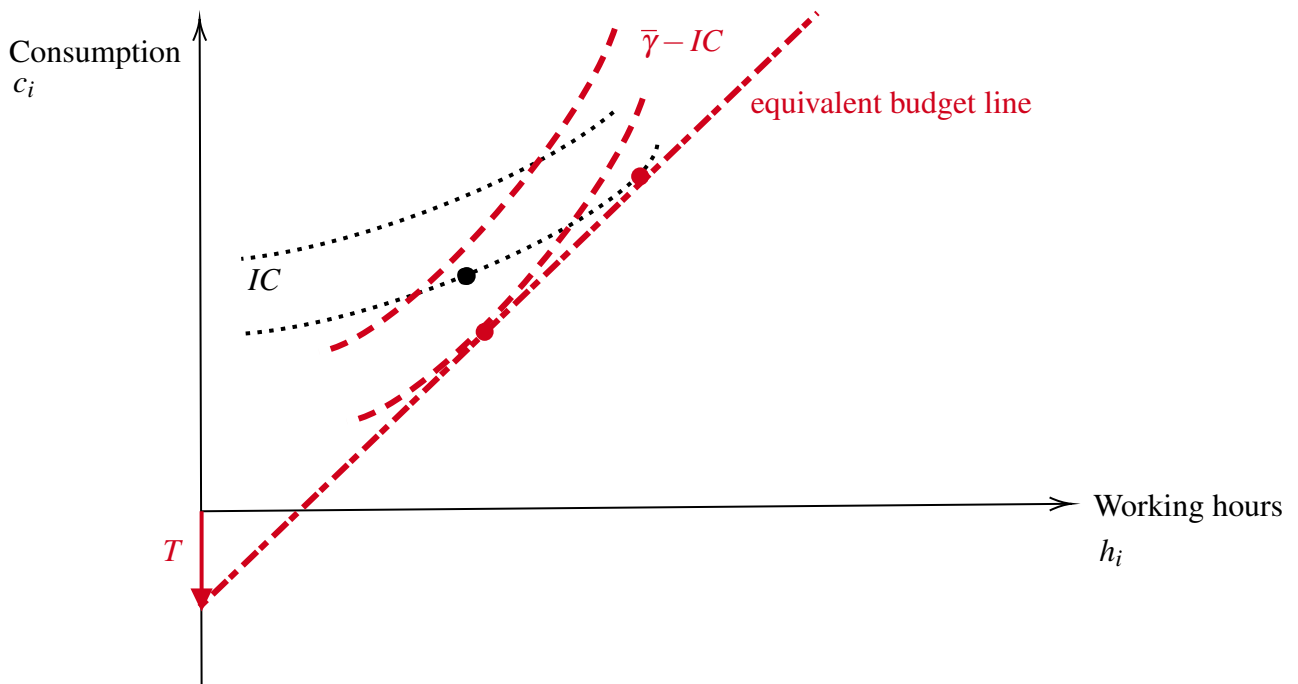
Notes: A is better-off than B according to reference preferences $\bar{\gamma}$ (dashed red $\bar{\gamma}$ -IC).

A.2 CE criterion

Under CE, households' well-being levels are ranked using the indifference curves obtained from a given reference preference for leisure, $\bar{\gamma}$, and the hypothetical equivalent budget set of each household. In Figure A3, this hypothetical linear budget line (intermittent red lines) has a slope equal to the household's wage rate and allows the household to achieve its original level of satisfaction. It is tangent to the household's subjective indifference curve, where the black dot is. At the same time, it identifies the optimal bundle that yields the household's highest level of satisfaction under the reference preferences $\bar{\gamma}$. This bundle lies on the equivalent budget line and on the highest attainable reference indifference curve (the lowest red dashed $\bar{\gamma}$ -IC drawn on Figure A3). The "CE lump-sum transfer" of this household, denoted T , is obtained from the intersection of the hypothetical budget line with the vertical axis ($h_i = 0$). In Figure A3, it corresponds to a negative transfer, i.e., a positive lump-sum tax.²⁵

²⁵Instead of fixing a reference form of the utility function, an alternative is to fix a reference value for labor time. It is equivalent with stating that all households have preferences such that they decide to work the same amount of labor hours. In this case, implementing CE involves ranking household' well-being based on the assumption that every worker works the same number of hours.

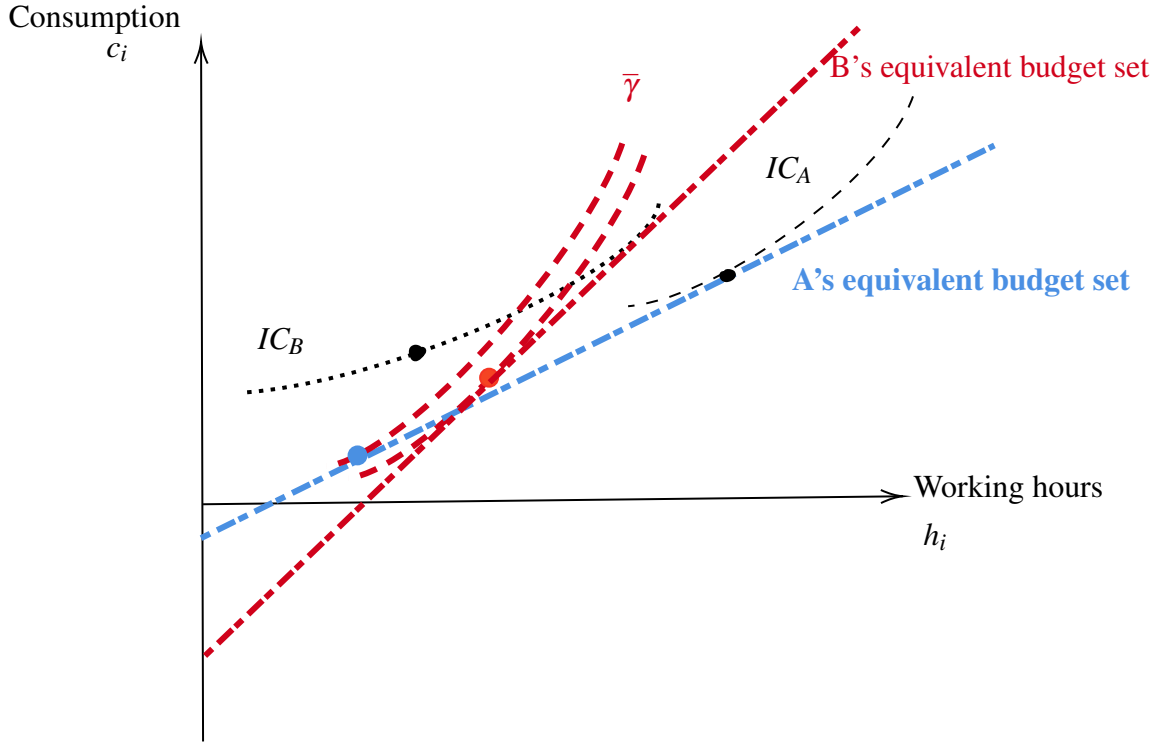
Figure A3: The conditional equality criterion (CE) obtained from a hypothetical equivalent budget line



Notes: Given the individual preferences, the CE lump-sum transfer, T , is obtained from a hypothetical linear budget line. The hypothetical linear budget line is constructed with the household's wage, w , as slope (intermittent lines) that would allow the household to achieve its original level of satisfaction under its subjective preferences. This equivalent budget line is then tangent to the indifference curve corresponding to the individual's original choice (the lowest dotted and black indifference curve in the figure). Additionally, it identifies the optimal bundle that yields the highest level of satisfaction under the reference preferences ($\bar{\gamma}-IC$) given the equivalent linear budget line. It corresponds to the tangency point between the equivalent budget line and a red dashed indifference curve.

Figure A4 presents a case where B , ranked as the worst-off individual according to CE, has a higher (simplified net) wage rate than A . Indeed household B 's "CE bundle" (in red) lies on a lower (red) reference indifference curve $\bar{\gamma}-IC$ compared to individual A 's "CE bundle" (in blue).

Figure A4: The conditional equality criterion (CE) obtained from hypothetical budget lines for A and B



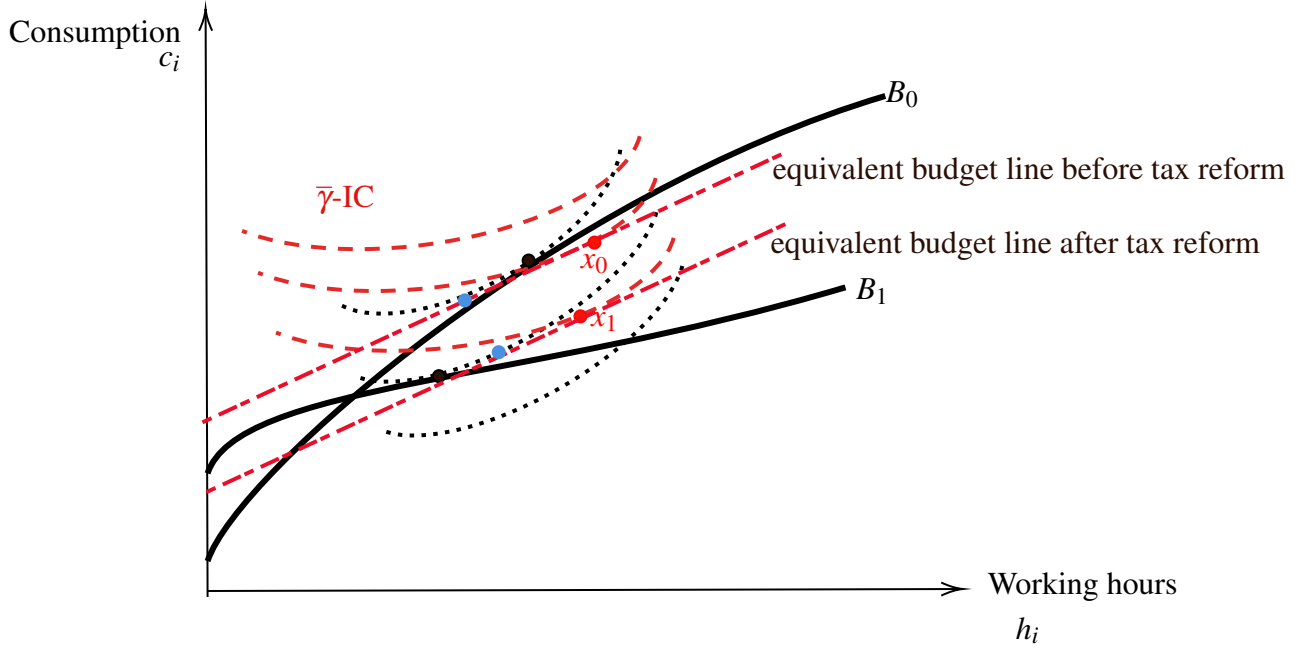
Notes: Under CE, household B, with a higher (net) wage rate than A, is ranked as the worse-off. Household B's "CE bundle" (in red) lies on a lower (red) reference indifference curve $\bar{\gamma}$ -IC compared to individual A's "CE bundle" (in blue).

A.3 ΔCE

The previous section was dedicated to the construction of CE. We now discuss how to implement it as ΔCE in the context of a tax reform. In Figure A5, we draw the hypothetical equivalent budget lines before and after a tax reform for one household (intermittent red lines). The equivalent budget line before the tax reform has a slope equal to the household wage rate and is tangent to the indifference curve passing through the bundle chosen under the pre-reform budget set B_0 . The equivalent budget line after the tax reform is defined similarly, under the post-reform budget set B_1 . With these equivalent budget lines, the household would choose other bundles highlighted in blue on Figure A5.

To compare the variation of household well-being, we need to identify two "CE-bundles", as was done in the previous subsection. They are denoted x_0 and x_1 , respectively. The welfare change under CE, ΔCE , is therefore defined as the difference between the two utility levels corresponding to x_1 and x_0 . We can also calculate the "CE lumps-sum transfers" as in the previous subsection and define the welfare change as the difference between these transfers.

Figure A5: ΔCE obtained from hypothetical equivalent budget lines (with a slope equal to the wage rate) before and after the tax reform



Notes: ΔCE is defined as the difference between levels of satisfaction under the reference preferences, for the two different tax schemes, represented by the two red dots along the (dotted red) reference preference indifference curves $\bar{\gamma}$ -IC.

B Properties of the indirect utility function

Given the distributional assumptions of the error terms the distribution of $\max_{z \in B(h_M, h_f)} \varepsilon(z)$ can easily be obtained. To realize this note that since $\varepsilon(z)$, with $z = (z_F, z_M)$, are assumed to be i.i.d. with c.d.f. $\exp(-e^{-x})$ it follows that

$$\begin{aligned} P\left(\max_{z \in B(h_F, h_M)} \varepsilon(z) \leq x\right) &= P\left(\bigcap_{z \in B(h_F, h_M)} (\varepsilon(z) \leq x)\right) = \prod_{z \in B(h_F, h_M)} P(\varepsilon(z) \leq x) \\ &= \prod_{z \in B(h_F, h_M)} \exp(-e^{-x}) = \exp\left(-\sum_{z \in B(h_F, h_M)} e^{-x}\right) = \exp(-Q_F(h_F)Q_M(h_M)e^{-x}). \end{aligned}$$

Accordingly, if we let $\eta_{h_F, h_M} = \max_{z \in B(h_F, h_M)} \varepsilon(z) - \log(Q_F(h_F)) - \log(Q_M(h_M))$, we obtain from the equation above that

$$\begin{aligned} P(\eta_{h_M, h_F} \leq x) &= P\left(\max_{z \in B(h_M, h_f)} \varepsilon(z) \leq x + \log(Q_F(h_F)) + \log(Q_M(h_M))\right) \\ &= \exp\{-Q_F(h_F)Q_M(h_M)\exp(-x - \log(Q_F(h_F)) - \log(Q_M(h_M)))\} = \exp(-e^{-x}), \end{aligned}$$

which means that the error term η_{h_F, h_M} has the same distribution as $\varepsilon(z)$. We can therefore write the conditional indirect utility function in Equation (4.2) as

$$\begin{aligned} V(h_F, h_M, I) &= \max_{z_F \in B(h_F), z_M \in B(h_M)} U(f(h_F w_F, h_M w_M, y), h_F, h_M, z) \\ &= u(f(h_F w_F, h_M w_M, I), h_F, h_M) + \max_{z_F \in B(h_F), z_M \in B(h_M)} \varepsilon(z_F, z_M). \end{aligned}$$

C Empirical labor supply model for married couples: summary statistics and estimation results

In Table C1, we present the summary statistics for the labor supply model for married couples and our estimation results in Table C2.

We note that for women, the age (both the $\log(\text{age}/10)$ and $\log(\text{age}/10)^2$) and the number of children (both below and above 6 years old) variables are the significant regressors capturing the heterogeneity of preferences for leisure. Being a woman and being with children is associated with greater preference for leisure.

Table C1: Summary statistics for the estimation sample of married couples, data from the Labor Force Survey 2015

	Mean	Std deviation
Age, male	4.66	0.87
Age, female	4.42	0.86
Non-labor income, male	10,798	20,437
Non-labor income, female	22,755	22,029
No. children under or equal to 6 years	0.29	0.59
No. children over 6 years	0.86	0.99
Years of schooling, male	1.32	0.27
Years of schooling, female	1.34	0.28

Table C2: Estimates of utility function parameters for couples, data from the Labor Force Survey 2015

	Parameter	Estimate	Std Error
Preferences			
<i>Consumption</i>			
Exponent	α_1	0.6694	0.075
Scale 10^4	α_2	1.0713	0.333
Consumption subsistence	C_0	57,000	
<i>Female leisure</i>			
Exponent	α_3	-1.1490	0.359
Constant	α_5	12.4108	4.758
Log(age/10)	α_6	-13.8036	5.425
Log(age/10) squared	α_7	5.2036	1.959
No. children under or equal to 6 years	α_8	0.4805	0.157
No. children over 6 years	α_9	0.2556	0.095
<i>Male leisure</i>			
Exponent	α_4	0.2309	0.308
Constant	α_{10}	4.3945	11.614
Log(age/10)	α_{11}	7.0184	16.451
Log(age/10) squared	α_{12}	-3.1912	5.632
No. children under or equal to 6 years	α_{13}	-0.1976	0.434
No. children over 6 years	α_{14}	-0.5784	0.272
<i>Leisure interaction</i>			
Interaction term	α_{15}	1.2111	0.863
Leisure subsistence	L_0	5,110	
The parameters of θ_F; $\log \theta_F = \gamma_{F1} + \gamma_{F2}\mathcal{S}$			
Constant	γ_{F1}	-2.9199	0.505
Education	γ_{F2}	0.1653	0.389
Opportunity density of hours offered			
Male full-time peak		2.8936	0.129
Female full-time peak		1.5027	0.096
Male part-time peak		-0.1512	0.301
Female part-time peak		-0.0451	0.122
Number of observations		1,594	
Log likelihood		-3070.9	
McFadden's ρ^2		0.52	

D Calculating the welfare effects of tax reforms

D.1 Estimation using McFadden's simulation approach

In the random utility framework, the model formulation implies that CV become random variables, and in general no closed-form formulae for welfare effects exist. Dagsvik and Karlström (2005) provide a short review of the approach, giving formulae for the distribution and moments for CV. McFadden (1999) describes a practical procedure for obtaining average estimates of CV based on simulations. In this paper, we adopt the simulation approach to calculate the CV of the tax reform

we study. In this appendix, for the sake of simplicity, we follow McFadden (1999) and present our approach for the individual-specific case. Recall that we, in the main part of the analysis, let the household be the unit of analysis.

The pre-reform tax rules are represented by $f_0(\cdot)$ while $f_1(\cdot)$ denotes the post-reform structure. We now need to introduce observed individual characteristics into the notation. Specifically, let the subscript i index the individual and let X_i be a vector of observed individual characteristics of individual i . It follows from Equation (4.12) that CV_i is determined by:

$$\begin{aligned} & \max_{h \in D} (u(f_0(w_i h, I_i), h; X_i) + \log(Q(h; X_i)) + \eta_i(h)) = \\ & \max_{h \in D} (u(f_1(w_i h, I_i) + CV_i, h; X_i) + \log(Q(h; X_i)) + \eta_i(h)). \end{aligned} \quad (\text{D.1})$$

A key assumption is that the error terms do not change from the pre-reform to the post-reform situation. Recall that $u(f_0(w_i h, I_i), h; X_i) + \log(Q(h; X_i)) + \eta_i(h)$ is a conditional indirect utility function in the sense that it is the utility of the most preferred job within the latent choice set $B(h)$. This means that our notion of compensating variation takes into account that the choices of the individual agents are constrained choices. Specifically, conditional on X_i , CV is a random variable that in general depends on all stochastic terms $\{\eta_i(h), h \in D\}$. This is because the maximum of the left-hand side of Equation (D.1) is not necessarily attained at the same discrete alternative as the maximum of the right hand side, except in special cases. Consequently, the random terms on each side do not cancel.

To obtain estimates of CV for household i by the simulation approach of McFadden (1999), we draw K set of random terms $\{\eta_i^k(h), h \in D, i \in S, k \in \{1, 2, \dots, K\}\}$ from the standard Gumbel distribution, where k indexes the draw. For each random draw, we numerically solve Equation (D.1) to obtain, for each individual, K estimates of CV_i^k . At first glance, directly solving Equation (D.1) numerically might seem time-consuming since it requires a maximization algorithm. This may be the case when random error terms are correlated, but not when they are independent. In this case, a standard one-dimensional search algorithm can be applied.²⁶

From the estimated $\{CV_i^k, k \in \{1, 2, \dots, K\}\}$ we obtain the simulated distribution of CV. This can be done in two ways. One method, which is the one we use, is based on computing the mean CV for each individual by taking the mean across all individual-specific CV_i^k and, subsequently calculating the empirical distributions of these individual-specific means. An alternative procedure estimates the distribution by directly utilizing all simulated CV values, accounting for unobserved heterogeneity within observationally identical groups of individuals.

²⁶As proposed by Creedy and Kalb (2005), one could constrain the error term distributions based on observed choices. However, we anticipate that this alternative methodology would not significantly alter our results.

D.2 Two alternative measures of CV accounting for fairness

The CV^{circ} alternative

To calculate CV adjusted for preference heterogeneity, we utilize a simulation approach similar to the one described in Section D.1. This method is based on Equation (4.13). It involves defining a common reference preference value, denoted as $\bar{u}(f(wh, y), h)$, and a common error term, $\bar{\eta}(h)$. We establish these reference values by setting common household characteristics, represented as $\bar{X}$, using median values of the corresponding variables in the sample (gender, age, presence and age of children).²⁷

With reference to Equation (4.13), we let $\bar{u}(f_0(w_{iF}h_F, w_{iM}h_M, I_i), h_F, h_M)$ be represented by

$$u(f_0(w_{iF}h_F, w_{iM}h_M, I_i), h_F, h_M; \bar{X}),$$

which illuminates that heterogeneity in preferences comes from the taste-modifying variables $\bar{X}$, see Table C2. Keeping in mind that $\bar{\eta}(h_F, h_M)$ signifies that a common error term is used too, we reformulate Equation (4.13) as follows:

$$\begin{aligned} \max_{h_F \in D, h_M \in D} & \left(\bar{u}(f_0(w_{iF}h_F, w_{iM}h_M, I_i), h_F, h_M; X_{ref}) + \log(Q(h_F, h_M; X_i)) + \bar{\eta}(h_F, h_M) \right) \\ &= \max_{h_F, h_M \in D} \left(\bar{u}(f_1(w_{iF}h_F, w_{iM}h_M, I_i) + CV_i^{circ}, h_F, h_M; X_{ref}) \right. \\ & \quad \left. + \log(Q_{iF}(h_F; X_i)) + \log(Q_{iM}(h_M; X_i)) + \bar{\eta}(h_F, h_M) \right). \end{aligned} \quad (D.2)$$

In a manner akin to the individual specific case (see Section D.1), we employ numerical methods to obtain CV^{circ} for each household for each draw of the reference vector $\bar{\eta}(h_F, h_M)$ within the domain D .

The CV^{pref} alternative

We may also calculate CV_i using an alternative neutralization, i.e., removing individual differences in circumstances. This can be done by introducing a (gender-specific) reference wage rate $\bar{w}$ and reference opportunity measure, $\bar{Q}(h_F, h_M)$. Using a common reference wage contributes to reducing inequalities arising from inequalities in ability, as it can be argued that abilities predominantly result from circumstances beyond the individual's control.²⁸ Importantly, given the job choice labor supply model, we are also able to control for differences in opportunities in the labor market,

²⁷Sensitivity tests, using other common values, such as the 10th and 90th percentiles, suggest that results are not much influenced by this choice.

²⁸Of course, it can be argued that wages are also determined by factors that can be viewed as being under individual's control, such as effort. In other words, fairness is achieved when individuals are treated the same, unless there are more morally relevant reasons for differential treatment (Berg and Piacquadio 2023). As already mentioned, it is challenging to empirically distinguish between “preferences” and “circumstances” or “efforts” and “endowments”.

represented by $\bar{M}(h)$. The corresponding measure of CV, referred to as CV_i^{pref} , is then derived by

$$\begin{aligned} \max_{h_F, h_M} & \left(u_i(f_0(\bar{w}_F h_F, \bar{w}_M h_M, I_i), h_F, h_M) + \log(\bar{Q}_F(h_F)) + \log(\bar{Q}_M(h_M)) + \eta_i(h_F, h_M) \right) \\ & \max_{h_F, h_M} \left(u_i(f_1(\bar{w}_F h_F, \bar{w}_M h_M, I_i) + CV_i^{pref}, h_F, h_M) \right. \\ & \quad \left. + \log(\bar{Q}_F(h_F)) + \log(\bar{Q}_M(h_M)) + \eta_i(h_F, h_M) \right). \quad (D.3) \end{aligned}$$

It follows that simulation results obtained by both Equation (D.2) and Equation (D.3) can be used to provide information on “fair” distributional effects of a tax change. In this paper, we focus on results when using Equation (D.2) and comparison with evidence obtained with standard CV and according to the CE criterion. Implications of using Equation (D.3) in empirical work is left for future research.