NBER WORKING PAPER SERIES

GENERALIZED SOCIAL MARGINAL WELFARE WEIGHTS FOR OPTIMAL TAX THEORY

Emmanuel Saez
Stefanie Stantcheva

Working Paper 18835
http://www.nber.org/papers/w18835

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2013

We thank Steve Coate, Florian Ederer, Marc Fleurbaey, Bas Jacobs, Louis Kaplow, Henrik Kleven, Etienne Lehmann, Ben Lockwood, Thomas Piketty, Larry Samuelson, Maxim Troshkin, Aleh Tsyvinski, Matthew Weinzierl, Nicolas Werquin, three anonymous referees, and numerous conference participants for useful discussions and comments. We acknowledge financial support from NSF Grant SES-1156240, the MacArthur Foundation, and the Center for Equitable Growth at UC Berkeley. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Generalized Social Marginal Welfare Weights for Optimal Tax Theory
Emmanuel Saez and Stefanie Stantcheva
NBER Working Paper No. 18835
February 2013, Revised July 2015
JEL No. H21

ABSTRACT

This paper proposes a new way to evaluate tax reforms, by aggregating losses and gains of different individuals using "generalized social marginal welfare weights." A tax system is optimal if no budget neutral small reform can increase the weighted sum of (money metric) gains and losses across individuals. Optimum tax formulas take the same form as standard welfarist tax formulas by simply substituting standard marginal social welfare weights with those generalized marginal social welfare weights. Weights directly capture society's concerns for fairness allowing us to cleanly separate individual utilities from social weights. Suitable weights can help reconcile discrepancies between the welfarist approach and actual tax practice, as well as unify in an operational way the most prominent alternatives to utilitarianism such as Libertarianism, Equality of Opportunity, or Poverty alleviation.

Emmanuel Saez
Department of Economics
University of California, Berkeley
530 Evans Hall #3880
Berkeley, CA 94720
and NBER
saez@econ.berkeley.edu

Stefanie Stantcheva
Department of Economics
Littauer Center 232
Harvard University
Cambridge, MA 02138
and NBER
sstantcheva@fas.harvard.edu

This paper proposes a novel approach to optimal tax theory using generalized social marginal welfare weights. In our approach, presented in Section I, there is no social welfare objective primitive that the government maximizes. Instead, our primitives are *generalized social marginal welfare weights* which represent the value that society puts on providing an additional dollar of consumption to any given individual.[1] These weights directly reflect society's concerns for fairness. Equipped with such weights, we can evaluate small budget neutral tax reforms by simply aggregating money metric utility gains and losses across individuals using the weights. If the net aggregate gain is positive, the reform is desirable (and conversely). We define a tax system as locally optimal if no small reform is desirable.[2] This has four implications.

First, optimal tax formulas in our theory take the same form as optimal tax formulas in the standard approach by simply substituting standard social welfare weights with our generalized weights. Hence, our theory remains as tractable as the standard approach and can easily be operationalized for any specification of generalized welfare weights. Second, our theory nests the standard welfarist approach where the government maximizes a social welfare function that depends solely on individual utilities if we define generalized social welfare weights to be equal to the standard marginal social welfare weights.[3] Third, if the weights are non-negative, then our theory respects the Pareto principle in the sense that, around the local optimum, there is no Pareto improving small reform. Fourth, our theory is by nature local as we can only evaluate small changes around a given tax/transfer system but we have no way of systematically comparing two local optima. In most applications however, the local optimum will be unique.

The key advantage of our approach is that weights can be defined very generally allowing us to capture a broader set of concepts of justice than the standard approach. Weights can depend on individual and aggregate characteristics, some of which are endogenous to the tax and transfer system. The characteristics which enter the welfare weights determine the dimensions along which society considers redistribution to be fair. These characteristics could be part of individuals' utilities (in the welfarist spirit). Importantly, though, social welfare weights can also depend on individual or aggregate characteristics which do not enter individuals' utilities. Conversely, the welfare weights can omit some characteristics which enter individuals' utility functions, but for which society does not deem it fair to compensate individuals.

[1] We take society's preferences as given and do not analyze how they could arise through the political process.

[2] With the tax reform approach, we can evaluate beneficial reforms even starting from non optimal tax system.

[3] To be precise, our theory nests the first order approach of the standard social welfare function maximization. As is well known (see e.g. Piketty and Saez, 2013, for a survey), the optimum of the standard approach is also such that no small reform can increase social welfare and the welfare effect of the reform is the aggregate of gains and losses using the standard social marginal welfare weights. One advantage of the standard approach relative to ours is that local optima (if they are multiple) can be ranked using the primitive social welfare function.

In Section II, we contrast our approach with the standard approach through several examples. First, we show that making generalized social marginal weights depend not only on net disposable income but also on net taxes paid produces a non-degenerate optimal tax theory even absent behavioral responses. Second, generalized social weights can depend on what individuals would have done absent taxes and transfers. Hence, we can capture the idea that society dislikes marginal transfers toward "free loaders" who would work absent means-tested transfers. Third, our approach can capture horizontal equity concerns. A reasonable criterion is that introducing horizontal inequities is acceptable only if it benefits the group discriminated against. This dramatically limits the scope for using non-income based tags.

In Section III, we show how the most prominent alternatives to welfarism can be re-cast within our theory, i.e., we can derive the generalized social welfare weights implied by those alternative theories. First, the equality of opportunity principle developed by Roemer (1998) and Roemer et al. (2003) concentrates weights uniformly on those coming from a disadvantaged background as, conditional on earnings, they have more merit (have worked harder) than those coming from an advantaged background. As the likelihood of coming from a disadvantaged background decreases with income, social weights decrease with income for a reason completely orthogonal to the decreasing marginal utility of income in utilitarianism. It also provides a rationale for less progressive taxes when there is high social or intergenerational mobility. Second, a poverty alleviation objective that respects the Pareto principle can be captured by social welfare weights concentrated on those below the poverty threshold.[4]

All formal proofs, several theoretical extensions, as well as results from a simple online survey designed to elicit social preferences of subjects are presented in the online Appendix.

I Presentation of our Approach

Individual utilities and taxes. We present our approach using income taxation.[5] Consider a population with a continuum of individuals indexed by i. Population size is normalized to one. Individual i derives utility from consumption c_i and incurs disutility from earning income $z_i \geq 0$, with a utility representation:

$$u_i = u(c_i - v(z_i; x_i^u, x_i^b))$$

[4] In online Appendix B, we consider additional alternatives to welfarism, such as Libertarianism, Rawlsianism, or the Fair Income Tax of Fleurbaey and Maniquet (2011) and show how they map into social weights.

[5] The same approach can be applied to other forms of taxation or transfer policies.

where x_i^u and x_i^b are sets of characteristics. The functions u and v are common to all individuals. u is increasing and concave and v is increasing and convex in z. The quasilinear functional forms rules out income effects on earnings which greatly simplifies optimal tax formulas (see below). u_i is a cardinal utility representation for individual i as viewed by the Planner (see below). x^u are characteristics that exclusively enter the utility function, while x^b will be characteristics that also affect the generalized social welfare weights introduced below.[6] Two important individual characteristics are considered throughout the paper. First, we consider a person's productivity per unit of effort $w_i \equiv z_i/l_i$ where l_i is labor supply. w_i is distributed in the population with a density $f(w)$ on $[w_{\min}, w_{\max}]$. Second, we consider the cost of work θ_i that affects the disutility from producing any unit of effort, distributed according to a distribution $p(\theta)$ on $[\theta_{\min}, \theta_{\max}]$. For instance, the disutility from labor could take the form $\theta_i \tilde{v}(z_i/w_i)$, so that any unit of effort is more costly for high θ_i agents.

The government sets an income tax $T(z)$ as a function of earnings only so that $c_i = z_i - T(z_i)$.[7] Individual i chooses z_i to maximize $u(z_i - T(z_i) - v(z_i; x_i^u, x_i^b))$.

Generalized social marginal welfare weights. We propose a novel theory of taxation that starts from the social welfare weights. For any individual, we define a *generalized social marginal welfare weight* g_i which measures how much society values the marginal consumption of individual i. We will assess the welfare gains from any reform by weighting the money metric welfare gains or losses to each individual using these weights and characterize optimal tax systems as systems around which no small reform can yield a welfare gain.

Definition 1 *The generalized social marginal welfare weight on individual i is $g_i = g(c_i, z_i; x_i^s, x_i^b)$ where g is a function, x_i^s is a set of characteristics which only affect the social welfare weight, while x_i^b is a set of characteristics which also affect utility.*

Naturally, the generalized weights are only defined up to a multiplicative constant as they measure only the *relative* value of consumption of individual i. Importantly, they are allowed to depend on individual characteristics x_i^s and x_i^b. These characteristics may be unobservable to the government or it may be impossible or unacceptable to condition the tax system on them. They nevertheless enter the social welfare weights because they affect how deserving a person is deemed by society. For instance, x_i^s might include family background (see our treatment of

[6] It is possible to consider a more general utility with $u_i = u(x_i^c \cdot c_i - v(z; x_i^u, x_i^b))$ where x_i^c would be a shifter parameter for the marginal utility from consumption. For simplicity, we abstract from this heterogeneity as none of our examples require it.

[7] Section II.C considers tax systems that can also depend on other observable individual characteristics ("tags").

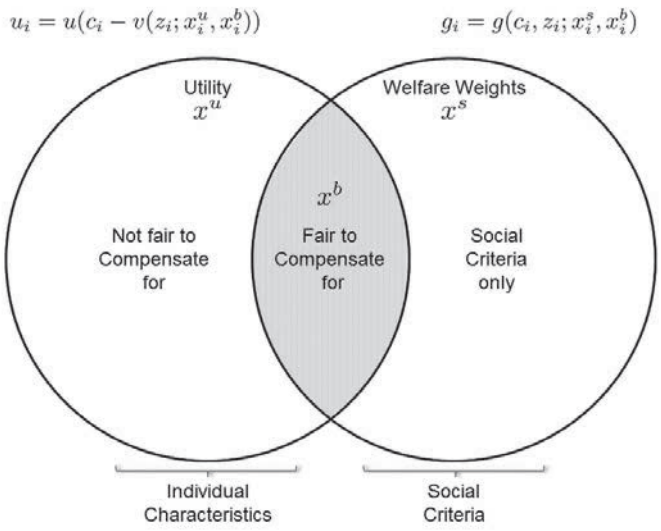


Figure 1: Generalized Social Welfare Weights Approach

Notes: The figure depicts the three sets of individual characteristics x^b, x^u, and x^s. Characteristics x^u enter solely the utility function (i.e, they affect individual utilities and choices). Characteristics x^s enter solely the generalized social welfare weights (i.e., they affect how society values marginal transfers to each individual). Characteristics x^b enter both the utility function and social weights.

"Equality of Opportunity" in section III.A), a characteristic that typically does not affect one's taxes directly but affects perceptions of deservedness.

As depicted on Figure 1, there is an important conceptual distinction between the sets of characteristics x^b, x^u, and x^s. Characteristics which enter the social welfare weight are dimensions that society considers potentially *fair to redistribute across* and to compensate for. For instance, if the disutility of work θ_i is due mostly to differences in health status or disability, then it might be fair to include it in the social welfare weight (see the "Free Loaders" example in section II.B). Conversely, if differences in disutility of work are mostly based on preferences for leisure, then they might not enter the social welfare weight (see the "Fair Income Tax" example in online Appendix B.5). These value judgments are directly embodied in the specification of the social welfare weights.

We keep individual preferences standard and individuals' utility maximization intact. Although it is possible to modify individual preferences to directly incorporate justice and fairness criteria (see Alesina and Angeletos, 2005, Fehr and Schmidt, 1999), that would still leave open the question of how to aggregate individual preferences.

Tax reform analysis. We now turn to defining desirable tax reforms around $T(z)$ and an optimal tax criterion. For a given ε, we define the perturbed tax system in the direction ΔT as

the tax system $z \to T(z) + \varepsilon \Delta T(z)$. We will derive the behavioral responses, revenue changes, and social welfare changes when $\varepsilon \to 0$, i.e., according to the Gâteaux differential.[8] In the main text, we consider only the first order approach because of its simplicity and its applicability to all cases we consider later on. We show in online Appendix A.1 how to formally consider second order conditions and a fully rigorous local optimization approach.

Denote by $z_i(T)$ the earning choice of agent i under tax system T. The behavioral response induced by the tax reform of system T in the direction ΔT is $\Delta z_i \equiv \lim_{\varepsilon \to 0} \frac{z_i(T+\varepsilon \Delta T) - z_i(T)}{\varepsilon}$. Using the standard envelope argument from the individual's optimization of z_i, the small tax reform $\varepsilon \Delta T$ mechanically changes individual disposable income by $-\varepsilon \Delta T(z_i)$, but the induced behavioral response in earnings $\varepsilon \Delta z_i$ has no first-order effect on utility. We denote by $U_i(T)$ the indirect utility of agent i under T, the utility first order change is $\Delta U_i \equiv \lim_{\varepsilon \to 0} \frac{U_i(T+\varepsilon \Delta T) - U_i(T)}{\varepsilon} = -u_{c_i} \cdot \Delta T(z_i)$.[9] Therefore, $\Delta T(z_i)$ measures the money-metric first order welfare impact of the tax reform in direction ΔT on individual i.[10] Hence, the first order effect on welfare is $-\int_i g_i \Delta T(z_i) di$ where di denotes the integral measure over all agents i in the population.

Next, we define a small tax reform as being budget neutral if it has no first order effect on net tax revenue (i.e., the Gâteaux differential of tax revenue is zero).

Definition 2 ***Budget neutral tax reform.*** *A small tax reform in direction $\Delta T(z)$ is budget neutral if and only if* $\lim_{\varepsilon \to 0} \{\int_i [T(z_i(T+\varepsilon \Delta T)) + \varepsilon \Delta T(z_i(T+\varepsilon \Delta T))] di - \int_i T(z_i(T)) di\}/\varepsilon = 0$.

In Definition 2, $\varepsilon \to 0$ both from the right ($\varepsilon > 0$) and left ($\varepsilon < 0$). Hence, if a small reform in direction $\Delta T(z)$ is budget neutral then a reform in direction $-\Delta T(z)$ is also budget neutral.

Next, we define a tax reform as being desirable if its first order effect on welfare is positive.

Definition 3 ***Tax reform desirability criterion.*** *A small budget neutral tax reform around $T(z)$ in direction $\Delta T(z)$ is desirable if and only if $\int_i g_i \Delta T(z_i) di < 0$, with g_i the generalized social marginal welfare weight on individual i evaluated at $(c_i = z_i - T(z_i), z_i, x_i^s, x_i^b)$.*

Last, we can define a locally optimal tax system as a tax system around which no budget neutral tax reform can improve welfare. Note that if $\int_i g_i \Delta T(z_i) di > 0$ then $\int_i g_i \cdot (-\Delta T(z_i)) di < 0$.

[8] Effectively, this amounts to considering first order effects in ε. We always consider interior tax systems so that if a small reform in the direction ΔT is feasible, a small reform in direction $-\Delta T$ is also feasible.

[9] To see this, denoting by $u_i(c, z)$ individual i utility function, we have the following first order expansion: $U_i(T+\varepsilon \Delta T) = u_i(z_i(T+\varepsilon \Delta T) - T(z_i(T+\varepsilon \Delta T)) - \varepsilon \Delta T(z_i(T+\varepsilon \Delta T)), z_i(T+\varepsilon \Delta T)) = U_i(T) - u_{c_i} \cdot \varepsilon \Delta T(z_i) + [u_{c_i} \cdot (1 - T'(z_i)) + u_{z_i}] \cdot \varepsilon \Delta z_i + o(\varepsilon) = -u_{c_i} \cdot \varepsilon \Delta T(z_i) + o(\varepsilon)$ as $u_{c_i} \cdot (1 - T'(z_i)) + u_{z_i} = 0$ by individual maximization.

[10] The incentive constraints are reflected in the behavioral response to the tax change. There are revenue effects from the behavioral responses. The welfare effect calculus requires that there are no externalities of individuals' actions and that the social welfare function is not paternalistic (i.e., respects individuals' optimization).

This immediately gives necessary conditions for a local optimum:[11]

Proposition 1 ***Local optimal tax criterion.*** *If a tax system $T(z)$ is a local optimum then for any small budget neutral tax reform in direction $\Delta T(z)$, $\int_i g_i \Delta T(z_i) di = 0$, with g_i the generalized social marginal welfare weight on individual i evaluated at $(z_i - T(z_i), z_i, x_i^s, x_i^b)$.*

As in the standard theory (see e.g. Kleven and Kreiner 2006, Chetty 2009, Hendren, 2013), the tax reform approach requires knowing only the weights g_i and the behavioral responses around the current system while the optimal tax criterion requires knowing the weights g_i and the behavioral responses *at the optimum*. Hence, the tax reform is more readily applicable.

From individual weights to applicable weights in the tax formula. The weights on each individual g_i are not immediately applicable. As noted, they can depend on unobservable characteristics or elements that the tax system cannot condition upon. They merely embody society's judgment of fairness, without taking into account observability or feasibility. To apply the weights to the evaluation of tax systems, the individual weights need to be "aggregated" up to only those characteristics that the tax system can be conditioned on. For a tax system purely based on income, $T(z)$, the weights need to be aggregated at each income level z.

Local optimal tax formulas. With Proposition 1, we can obtain an optimal tax formula that is particularly simple in the case with no income effects on labor supply that we consider here (Diamond, 1998). The proof is the same as in the standard approach (see Piketty and Saez, 2013) and is presented in online Appendix A.2.[12] Let $H(z)$ denote the cumulative earnings distribution function and $h(z)$ the earnings density. Both naturally depend on the tax system.

Definition 4 *Let $\bar{G}(z)$ be the (relative) average social marginal welfare weight for individuals who earn more than z:*

$$\bar{G}(z) \equiv \frac{\int_{\{i:z_i \geq z\}} g_i di}{Prob(z_i \geq z) \cdot \int_i g_i di} \tag{1}$$

Let $\bar{g}(z)$ be the corresponding average social marginal welfare weight at earnings level z, with $\bar{G}(z)[1 - H(z)] = \int_z^\infty \bar{g}(z')dH(z')$, or, equivalently $\bar{g}(z) = -\frac{1}{h(z)} \frac{d(\bar{G}(z)\cdot[1-H(z)])}{dz}$.

From definition 4, we have $\bar{G}(0) = \int_0^\infty \bar{g}(z)dH(z) = 1$, so that the weights $\bar{g}(z)$ average to one.

[11] We discuss second order conditions in online Appendix A.1. The extra complexity required is not needed in the applications we consider later on and hence relegated to the appendix.

[12] Using small tax reforms around the optimum, Saez (2001) informally derived an optimal income tax formula that generalizes the formulas of Mirrlees (1971) to situations with heterogeneous populations, where individuals differ not only in skills but also possibly in preferences. Jacquet and Lehmann (2015) provide a fully rigorous proof of the formula with heterogeneous populations and give conditions under which it applies. We always assume here that these conditions hold.

Proposition 2 *The optimal marginal tax rate at income level z satisfies the formula:*

$$T'(z) = \frac{1 - \bar{G}(z)}{1 - \bar{G}(z) + \alpha(z) \cdot e(z)} \tag{2}$$

with $e(z)$ the average elasticity of earnings z_i with respect to the retention rate $1 - T'$ for individuals earning $z_i = z$, $\alpha(z)$ the local Pareto parameter defined as $zh(z)/[1 - H(z)]$.[13]

The optimal tax formula looks exactly as in Saez (2001), replacing the standard weights by the generalized social marginal welfare weights, averaged at each income level. Accordingly, some standard results would hold under the same conditions. For instance, $T'(z) \geq 0$ if average social marginal welfare weights $\bar{g}(z)$ are decreasing in income.[14] The optimal marginal tax rate at the very top is zero if the income distribution is bounded.

We can similarly express the optimal *linear* tax rate τ as a function of the welfare weights:[15]

Proposition 3 *The optimal linear income tax rate satisfies the formula:*

$$\tau = \frac{1 - \bar{g}}{1 - \bar{g} + e} \quad \text{with} \quad \bar{g} \equiv \frac{\int_i g_i \cdot z_i di}{\int_i g_i di \cdot \int_i z_i di} \tag{3}$$

and e the elasticity of aggregate income $\int_i z_i di$ with respect to the retention rate $(1 - \tau)$.

$\bar{g}$ can be interpreted as the average g_i weighted by income z_i (relative to the population average g_i) or symmetrically as the average z_i weighted by g_i (relative to $\int_i z_i di$).

As in standard optimal tax theory, formulas (2) and (3) hold at an optimum but are not necessarily sufficient. Importantly, there may be several tax systems that can satisfy the formulas. Our theory is fundamentally local and hence cannot rank optima when they are multiple. This is a disadvantage relative to any approach that maximizes a social objective and hence provides a complete ordering of all tax systems.

Advantages of the generalized approach. First, our approach allows us to re-use existing tax formulas such as equations (2) and (3) once they are written in terms of the welfare weights at each income level. Hence, it also nests the standard approach. Second, it can easily ensure that any tax optimum is (locally) constrained Pareto efficient as long as the generalized weights g_i are all non-negative. In that case, our theory respects individual preferences. Third and

[13] The local Pareto parameter $\alpha(z)$ is constant and equal to the Pareto parameter for Paretian distributions. To be precise, when defining $\alpha(z)$, $h(z)$ is defined as the virtual density that would hold at z if the income tax system were linearized at z. See Saez (2001) and Piketty and Saez (2013) for details.

[14] This condition is satisfied with standard utilitarian weights. See Appendix A.2 for details on these results.

[15] See again online Appendix A.2 for a short derivation and Piketty and Saez (2013) for details.

most importantly, as we will outline throughout the paper, our approach grants great flexibility in the choice of the welfare weights g_i. The fact that social welfare weights can depend on characteristics outside of individuals' utilities, as well as ignore characteristics from individuals' utilities–as illustrated in Figure 1–allows us to incorporate elements that matter in actual tax policy debates but that cannot be captured with the standard welfarist approach.[16]

Implicit Pareto weights. As is well known, under standard regularity conditions, any second-best allocation can be obtained maximizing a linear social welfare function $SWF = \int_i (\omega_i \cdot u_i) di$ with suitable *Pareto weights* $\omega_i \geq 0$. The following proposition extends this result to our setting.

Proposition 4 *For any non-negative generalized weights function and a local optimum $T(z)$, there exist Pareto weights $\omega_i \geq 0$ such that $T(z)$ satisfies the first order condition formula (2) of the maximization of $SWF = \int_i (\omega_i \cdot u_i) di$. The Pareto weights are such that $\omega_i = g_i / u_{c_i} \geq 0$ where g_i and u_{c_i} are evaluated at the optimum allocation.*

Hence, our approach can be reverse-engineered to obtain a set of Pareto weights ω_i and a corresponding standard social welfare function $\int_i (\omega_i \cdot u_i) di$. Importantly, because g_i / u_{c_i} is evaluated at the optimum allocation, it is taken as fixed in the maximization of SWF. In practice as we shall see, it is impossible to posit the correct weights ω_i without *first* having solved for the optimum using our approach that starts with the social marginal weights g_i. In addition, it is typically impossible to translate the generalized social welfare weights into Pareto weights that are fixed and hence cannot be modified with changes in the environment. A social welfare function with endogenous Pareto weights such as $\int_i \omega_i(T) u_i di$ would not deliver the same solution as our approach and could lead to Pareto dominated outcomes.[17]

II Enriching the Welfarist Approach

In this section, we show how the use of suitable generalized social marginal welfare weights can enrich the welfarist approach by reconciling some of the discrepancies between the welfarist approach (and particularly the widely used utilitarian case) and actual tax practice. Table 1 summarizes these results by contrasting actual tax practice (column 1), the standard welfarist

[16] The tax reform approach, which is *per se* not specific to the generalized welfare weights, has the advantage that it can evaluate reforms even starting from non-optimal systems.

[17] A closely related point has also been made by Fleurbaey and Maniquet (2015) who show that their fair income tax social objective maximization can also be obtained as the maximization of a weighted sum of utilities but the weights "would have to be computed for each new problem, that is, as a function of the set of allocations among which the choice has to be made,..., and could only be ascertained after the optimal allocation is identified."

approach (column 2), and our generalized social marginal welfare weights approach (column 3) in various situations. In each situation, column 3 indicates what property of social marginal welfare weights is required to make this approach fit with actual tax policy practice.

II.A Optimal Tax Theory with Fixed Incomes

We start with the simple case in which pre-tax incomes are completely inelastic to taxes and transfers. This puts the focus solely on redistributive issues. It is a useful illustration of our approach, especially as contrasted with the standard welfarist approach. We specialize our general framework with a disutility of work $v(z; z_i^0) = 0$ if $z \leq z_i^0$ and $v(z; z_i^0) = \infty$ if $z > z_i^0$. Thus, z_i^0 is an exogenous characteristic of individual i, contained in x_i^u, and choosing $z_i = z_i^0$ is always optimal for individual i, so that the distribution of incomes is exogenous to the tax system. In equilibrium, utility is $u_i = u(c_i)$. We review first the standard utilitarian setting.

Standard utilitarian approach. The government chooses $T(z)$ to maximize $\int_i u(z_i - T(z_i))di$ subject to the resource constraint $\int_i T(z_i)di \geq 0$ (with multiplier p). A point-wise maximization with respect to $T(z)$ yields $u'(z - T(z)) = p$ so that $c = z - T(z)$ is constant across z. Hence, utilitarianism with inelastic earnings and a concave individual utility function, homogeneous across individuals, leads to complete redistribution of incomes. The government taxes 100 percent of earnings and redistributes income equally across individuals (Edgeworth, 1897).

This simple case highlights three drawbacks of utilitarianism. First, complete redistribution seems too strong a result. In reality, even absent behavioral responses, many and perhaps even most people would still object to full taxation on the grounds that it is unfair to fully confiscate individual incomes. Second, the outcome is extremely sensitive to the specification of individual utilities, as linear utility calls for no taxes at all, while introducing just a bit of concavity leads to complete redistribution. Third, the utilitarian approach cannot handle well heterogeneity in individual utility functions, a problem known as inter-personal utility comparisons. To see this, suppose that utilities are heterogeneous of the form $u_i = x_i \cdot u(c_i)$. The optimum is such that $x_i \cdot u'(c_i)$ are equal for all i. Hence, consumption should be higher for individuals with higher x_i, i.e., individuals more able to enjoy consumption. In reality, society would be reluctant to redistribute based on preferences, as confirmed by our online survey in online Appendix C.[18]

Generalized Social Marginal Welfare Weights. The simplest way to illustrate the power of our approach with fixed incomes is to use the generalized weights from definition 1 without

[18]Redistribution based on marginal utility is socially acceptable if there are objective reasons a person has higher needs, such as having a medical condition requiring high expenses, or a large family with many dependents.

using any additional characteristics ($x_i^u = z_i^0$, for all i, and x^b and x^s are empty).

Definition 5 Simple generalized weights: *Let* $g_i = g(c_i, z_i) = \tilde{g}(c_i, z_i - c_i)$ *with* $\tilde{g}_c \equiv \frac{\partial \tilde{g}}{\partial c} \leq 0$, $\tilde{g}_{z-c} \equiv \frac{\partial \tilde{g}}{\partial (z-c)} \geq 0$. *There are two polar cases of interest:*
i) Utilitarian weights: $g_i = g(c_i, z_i) = \tilde{g}(c_i)$ *for all* z_i, *with* $\tilde{g}(\cdot)$ *decreasing.*
ii) Libertarian weights: $g_i = g(c_i, z_i) = \tilde{g}(z_i - c_i)$ *with* $\tilde{g}(\cdot)$ *increasing.*

Weights depend not only negatively on c but also positively on net taxes paid $z - c$. $\tilde{g}_c \leq 0$ can reflect decreasing marginal utility of consumption as under utilitarianism or more generally the old notion of "ability to pay" (those with higher disposable income can afford paying \$1 extra in taxes more easily). $\tilde{g}_{z-c} \geq 0$ reflects the fact that taxpayers contribute more to society and hence are more deserving. Alternatively, individuals are entitled to their income and hence become more deserving as the government taxes away their income.[19,20]

The optimal tax system, according to Proposition 1 is such that no reform can increase social welfare at the margin, where transfers are evaluated using the g weights. Since, $T(z) = z - c$, $\tilde{g}(c, z - c) = \tilde{g}(z - T(z), T(z))$. With no behavioral responses, the optimal rule is simple: social welfare weights $\tilde{g}(z - T(z), T(z))$ need to be equalized across all incomes z. Intuitively, if we had non equalized weights with $\tilde{g}(z_1 - T(z_1), z_1) > \tilde{g}(z_2 - T(z_2), z_2)$, transferring a dollar from those earning z_2 toward those earning z_1 (by adjusting $T(z_1)$ and $T(z_2)$ correspondingly and in a budget balanced manner) would be desirable (the formal proof is in online Appendix A.4).

Proposition 5 *The optimal tax schedule with no behavioral responses is such that:*

$$T'(z) = \frac{1}{1 - \tilde{g}_{z-c}/\tilde{g}_c} \quad \text{so that} \quad 0 \leq T'(z) \leq 1. \tag{4}$$

Corollary 1 *In the standard utilitarian case,* $T'(z) \equiv 1$. *In the libertarian case,* $T'(z) \equiv 0$.

We present in online Appendix C results from a survey asking subjects to rank taxpayers with various incomes and tax burdens in terms of deservedness of a tax break. Respondents put weight on both disposable income and gross income, showing that social preferences are in between the polar utilitarian and libertarian cases. Such data can be used to recover social preferences $g(c, z)$. For instance, the specification $g(c, z) = \tilde{g}(c - \alpha(z - c)) = \tilde{g}(z - (1 + \alpha)T(z))$ with $\tilde{g}$ decreasing and where α is a constant parameter delivers an optimal tax with a constant marginal tax rate $T'(z) = 1/(1 + \alpha)$ and is empirically calibrated in online Appendix C.2.

[19] This is akin to an "Equal Sacrifice" principle in money metric utility terms. Weinzierl (2014) also incorporates a libertarian element in an optimal tax model by considering a mixed objective.

[20] We assume away government funded public goods in our setup for simplicity.

II.B Transfers and Free Loaders

The public policy debate on transfers often focuses on whether non-workers are deserving of support or not. Transfer beneficiaries are deemed deserving if they are truly unable to work, that is, if absent any transfers, they would still not work and live in great poverty without resources. Conversely, they are considered non-deserving, or "free loaders" if they could work and would do so absent more generous transfers. The presence of such "free loaders", perceived to take undue advantage of a generous transfer system, is precisely why many oppose welfare (see e.g., Ellwood (1988)). It is also the reason why many welfare programs try to target populations which are deemed more vulnerable and less prone to taking advantage of the system. Historically, disabled people, widows, and later on single parents have been most likely to receive public transfers.[21] Our online survey in online Appendix C confirms that people have strong views on who, among those out-of-work, is deserving of support. We consider a basic model to explain how generalized weights can be used to capture the concept of free loaders.

Model. Starting from our general model, assume that individuals can either work and earn $z = 1$, or not work and earn zero $z = 0$. We take the special functional utility form: $u(c - \theta \cdot z)$ where $z \in \{0, 1\}$. Consumption c is equal to c_0 if an individual is out of work and to $c_1 = (1 - \tau) + c_0$ if she works, so that τ is the net tax rate on earnings.[22] Taxes fund the transfer c_0. The cost of work θ is private information and distributed according to a cdf $P(\theta)$. An individual with cost of work θ works if and only if $\theta \leq c_1 - c_0 = 1 - \tau$. Hence, the fraction of people working is $P(1 - \tau)$. Let e be the elasticity of total earnings $P(1 - \tau)$ with respect to $1 - \tau$.

This model is a special case of the optimal linear tax model discussed in Section I. Hence, we can immediately apply formula (3) and the optimal tax is $\tau^* = (1-\bar{g})/(1-\bar{g}+e)$. As non-workers have $z_i = 0$ and workers have $z_i = 1$, we have $\bar{g} = \int_i g_i z_i di / (\int_i g_i di \cdot \int_i z_i di) = \bar{g}_1/[P \cdot \bar{g}_1 + (1-P) \cdot \bar{g}_0]$ where $P = P(1 - \tau)$, $\bar{g}_1$ is the average g_i on workers, and $\bar{g}_0$ is the average g_i on non-workers.

Standard utilitarian approach. Under the utilitarian objective, we have $g_i = u'(c_0)$ for all non-workers so that $\bar{g}_0 = u'(c_0)$.[23] Hence, the utilitarian approach cannot distinguish between the deserving poor and free loaders. The utilitarian social welfare weight placed on non-workers depends only on c_0 and is completely independent of whether the person would have worked absent taxes and transfers: The standard approach cannot take into account counterfactuals.

[21] The origins of the US welfare system since 1935, starting with the Aid to Families with Dependent Children and continuing with the Temporary Assistance to Needy Families assistance programs highlights that logic.

[22] As there are only two earnings outcomes, 0 and 1, this tax system is fully general.

[23] For workers, $g_i = u'(c_1 - \theta_i)$ so that $\bar{g}_1 = \int_{\theta \leq 1-\tau} u'(c_0 + (1 - \tau) - \theta) dP(\theta)/P < \bar{g}_0$.

Generalized social welfare weights. The generalized weights allow us to treat the deserving poor differently from the free loaders. The deserving poor are those with $\theta > 1$, (who would not work, even absent any transfer). The free loaders are those with $1 \geq \theta > 1-\tau$ (who do not work because of the existence of transfers). Denoting by $P_0 = P(1)$ the fraction working when $\tau = 0$, there are $P = P(1-\tau)$ workers, $1 - P_0$ deserving poor, and $P_0 - P$ free loaders.

Let us assume that society sets social marginal welfare weights as follows. Workers obtain a standard weight $g_i = u'(c_1 - \theta_i)$ if $z_i = 1$. The deserving poor obtain a standard weight $g_i = u'(c_0)$ if $z_i = 0$ and $\theta_i \geq 1$. Finally, the free-loaders obtain a weight $g_i = 0$ if $z_i = 0$ and $\theta_i < 1$. The weight function is $g_i = g(c_i, z_i; \theta_i, \tau)$. In our general notation, $x_i^b = (\theta_i, \tau)$.[24]

With such weights, we have $\bar{g}_0 = u'(c_0) \cdot (1-P_0)/(1-P)$, as only a fraction $(1-P_0)/(1-P)$ of the non-workers are deserving. Hence, $\bar{g}_0$ is lower relative to the utilitarian case. $\bar{g}_1$ is unchanged relative to the utilitarian case. Therefore, $\bar{g} = \bar{g}_1/[P \cdot \bar{g}_1 + (1-P) \cdot \bar{g}_0]$ is now *higher* than in the utilitarian case and τ^* is correspondingly lower (keeping e and P constant). As expected, the presence of free loaders reduces the optimal tax rate τ^* relative to the standard case.

In the extreme case in which all non-workers are free-loaders, the optimal transfer (and hence the tax rate financing it) is zero. This corresponds to the (extreme) view that all non employment is created by an overly generous transfer system. As long as there are some deserving poor though, the tax rate and transfers will be positive.

For a given τ, when e is larger, $(1 - P_0)/(1 - P)$ is smaller as more people become free loaders. Hence a higher e reduces τ^* not only through the standard efficiency effect e but also through the channel $\bar{g}$ as it negatively affects society's view on how deserving the poor are.

This example also illustrates that *ex post*, it is possible to find suitable Pareto weights for the welfarist approach that can rationalize the tax rate τ^* obtained with generalized welfare weights. In this case, maximizing $\int_\theta \omega(\theta) \cdot u(c - \theta \cdot z) \cdot dP(\theta)$ with $\omega(\theta) = 1$ for $\theta \leq 1-\tau^*$ (workers) and $\theta \geq 1$ (deserving poor) and $\omega(\theta) = 0$ for $1 - \tau^* < \theta < 1$ (free loaders). However, the Pareto weights ω depend on the optimum tax rate τ^* and, hence, cannot be specified *ex ante*. Note that our approach is not equivalent to making the Pareto weights endogenous. Indeed, a social welfare function of the form: $\int_\theta \omega(\theta, \tau) \cdot u(c - \theta \cdot z) \cdot dP(\theta)$ with $\omega(\theta, \tau) = 1$ for $\theta \leq 1 - \tau$ and $\theta \geq 1$ and $\omega(\theta, \tau) = 0$ for $1 - \tau < \theta < 1$ would not yield the same solution because we would have to take into account the change in social welfare stemming from changes in the endogenous Pareto weights. What is more, such a social welfare objective with endogenous Pareto weights

[24] Note that the cost of work θ_i enters into the social welfare weights so that it is fair to compensate people for their differential cost of work. If instead θ reflected pure laziness, then this might not be the case (as in the Fair Income Tax theory of Fleurbaey and Maniquet, 2011, also covered in online Appendix B.5).

could yield Pareto dominated outcomes.[25] Instead, our approach consists in weighting utility changes using locally fixed weights, which respects the local Pareto principle.

Application 1: Desirability of in-work benefits. As shown in Piketty and Saez (2013), in the nonlinear tax model of Section I, when some individual do not work, the optimal marginal tax rate at the bottom with zero earnings takes the simple form $T'(0) = (\bar{g}_0 - 1)/(\bar{g}_0 - 1 + e_0)$ where $\bar{g}_0$ is the average social marginal welfare weight on those out-of-work and e_0 is (minus) the elasticity of the fraction of individuals out-of-work with respect to $1 - T'(0)$. Unlike in the standard utilitarian case, in which $\bar{g}_0 > 1$ and hence $T'(0) > 0$, with free loaders, $\bar{g}_0$ is lower and could be lower than one, in which case $T'(0) < 0$, i.e., in-work benefits are optimal.

Application 2: Transfers over the business cycle. Individuals are less likely to be responsible for their unemployment status in a recession than in an expansion, so that the composition of those out of work changes over the business cycle. Hence, this is a force pushing toward expanding benefits in bad times. Results from the online survey show indeed that support for the unemployed depends strongly on whether they can or cannot find jobs (see online Appendix C).

II.C Tagging and Horizontal Equity Concerns

Under welfarism, if agents can be separated into different groups, based on attributes, so-called "tags," which are correlated with earnings ability or behavioral elasticities, then an optimal tax system should generally have differentiated taxes across those groups.[26] Mankiw and Weinzierl (2010) explore a tax schedule differentiated by height and use this stark example as a critique of utilitarianism. In practice, society seems to oppose taxation based on such characteristics, probably because it is deemed unfair to tax differently people with the same ability to pay. These 'horizontal equity' concerns, or the wish to treat 'equals as equals' seem important in practice and a realistic framework for optimal tax policy needs to be able to include them.[27]

It is possible to capture horizontal equity concerns using generalized social welfare weights if we extend our basic framework to allow the weights to be dependent on the *reform considered* (instead of depending solely on the current level of taxes and transfers as we have done so far). To save space, we will not develop a general theory but will solely focus on a simple example.

[25] Indeed, reforms that increase the weights could be desirable even if all agents' utilities decline slightly.

[26] Some attributes can be perfect tags in the sense of being impossible to influence by the agent. An example would be height, which has been shown to be positively correlated with earnings (see Mankiw and Weinzierl, 2010), or gender. Others are potentially elastic to taxes (such as the number of children or marital status).

[27] Kaplow (2001) criticizes the concept of horizontal equity and highlights that it will conflict with the Pareto principle in some cases. Our non-negative social welfare weights guarantee that this cannot occur in our setup.

To illustrate this in the simplest way possible, we consider a Ramsey tax problem where linear taxation is designed to raise a given (non-transfer) revenue E. There are two groups which differ according to an observable and perfectly inelastic attribute $m \in \{1, 2\}$. The two groups differ according to their taxable income elasticities (respectively denoted e_1 and e_2). The tax rate on group m is denoted by τ_m. The budget constraint, with multiplier p, is:

$$\tau_1 \cdot \int_{i \in 1} z_i di + \tau_2 \cdot \int_{i \in 2} z_i di \geq E. \tag{5}$$

We abstract entirely from vertical equity issues by assuming that utilities are linear in consumption so that $u_{c_i} = 1$. As a result, under the utilitarian criterion, all individuals have the same social marginal welfare weight. The standard utilitarian approach would naturally lead to different tax rates for each group, following standard Ramsey considerations, such that $\tau_m = (1 - 1/p)/(1 - 1/p + e_m)$ where $p > 0$ is determined so that the tax system raises exactly E to meet the budget constraint (5) (see online Appendix A.5 for a formal proof). Hence, the standard tax system would generate horizontal inequities, i.e., some individuals are taxed more than others based on a tag and conditional on income. Without loss of generality, we assume $e_2 > e_1$ so that group 2 is more elastic and should optimally be taxed less.

How can we specify generalized social marginal welfare weights to reflect the view that horizontal inequities are unfair? First, as in the utilitarian case, absent any horizontal inequity (i.e., if $\tau_1 = \tau_2$), we assume that weights are equal across the population (and normalized to one). Second, if the tax system creates horizontal inequity (i.e., $\tau_1 \neq \tau_2$), then social marginal welfare weights are equal to one for individuals in the group facing the highest tax rate and zero for those facing the lowest tax rate. However, such weights are not sufficient to ensure that the no tagging tax system with $\tau_1 = \tau_2$ is a local optimum. Indeed, starting from this tax system, lowering τ_2 and increasing τ_1 would be desirable as everybody has the same weights (no equilibrium would exist). Hence and last, we need to expand our framework and make weights depend on the direction of reform as well when we start from an equitable tax system and the small reform introduces a horizontal inequity.[28] In our simple example with linear taxation, starting from identical tax rates for the two groups, a small reform that introduces inequity can lead to (a) $\tau_1 > \tau_2$, or (b) $\tau_1 < \tau_2$. In case (a), group 1 suffers from a horizontal inequity because of the reform and the weights are set equal to one on group 1 and zero on group 2. In case (b) conversely, the weights are set equal to zero on group 1 and one on group 2. Such weights are

[28] If the system is inequitable to start with (i.e., $\tau_1 \neq \tau_2$), then a small reform will not affect the sign of the initial inequity and hence the weights do not need to depend on the reform.

designed to avoid horizontal inequities unless they benefit the group discriminated against.

Regularity assumptions. We assume that there is a uniform tax rate $\tau_1 = \tau_2 = \tau^*$ that can raise E. We assume that the tax functions $\tau_1 \to \tau_1 \cdot \int_{i\in 1} z_i di$ and $\tau_2 \to \tau_2 \cdot \int_{i\in 2} z_i di$, and the uniform rate tax function $\tau \to \tau \cdot (\int_{i\in 1} z_i di + \int_{i\in 2} z_i di)$ are all single peaked (Laffer curves).

Naturally, the peaks are, respectively, at $\tau_1 = 1/(1+e_1)$, $\tau_2 = 1/(1+e_2)$, and $\tau = 1/(1+e)$, where e is the elasticity of total income with respect to $1-\tau$.

Proposition 6 *Let τ^* be the (smallest) uniform rate that raises E: $\tau^*(\int_{i\in 1} z_i di + \int_{i\in 2} z_i di) = E$.*
i) If $1/(1+e_2) \geq \tau^*$ then the only local optimum has horizontal equity with $\tau_1 = \tau_2 = \tau^*$.
ii) If $1/(1+e_2) < \tau^*$ then the only local optimum has horizontal inequity with $\tau_2^* = 1/(1+e_2) < \tau^*$ (revenue maximizing rate on group 2) and $\tau_1^* < \tau^*$ is the smallest τ_1 such that $\tau_1 \cdot \int_{i\in 1} z_i di + \tau_2^* \cdot \int_{i\in 2} z_i di = E$. This optimum Pareto dominates the uniform tax system $\tau_1 = \tau_2 = \tau^*$.

Therefore, a tax system with horizontal inequity can be an optimum only if it helps the group discriminated against, i.e., no reform can improve the welfare of those discriminated against. This can happen only when tagging creates a Pareto improvement, which dramatically reduces the scope for using tagging in practice. In our example, if the government wants to set τ_2 at a lower level than τ_1, then τ_2 must necessarily be set at the revenue maximizing rate.

This is reminiscent of a Rawlsian setup, in which society only cares about the least well-off. Here, the set of people whom society cares about is endogenous to the tax system. Namely, they are the ones discriminated against because of tagging. In other words, we can rephrase the Rawlsian criterion as follows: "It is permissible to discriminate against a group using taxes and transfers not based on ability to pay only in the case where such discrimination allows to improve the welfare of the group discriminated against."

III Link with Alternative Justice Principles

In this section, we illustrate how our framework can be connected to justice principles that are not captured by the standard welfarist approach but have been discussed in the normative tax policy literature. We use formula (2) to show how social welfare weights derived from alternative justice principles map into optimal tax formulas.

III.A Equality of Opportunity

To capture the concept of equality of opportunity, Roemer (1998) and Roemer *et al.* (2003) consider models where individuals differ in their ability to earn, but part of the ability is due to family background (which individuals are not responsible for) and to merit (which individuals are responsible for). Conditional on family background, merit is measured by the percentile of the earnings distribution the individual is in. Society is willing to redistribute across family backgrounds but not across earnings conditional on family background.

Formally, individual i's utility is $u_i = u(c_i - v(z_i/w_i))$. w_i is productivity stemming from two sources. First, individual i is born into a low or high family background denoted by $B_i = 0, 1$. Second, through merit (e.g., hard work at school), individual i reaches rank r_i conditional on background and hence obtains a productivity w_i at the r_i-th percentile of the distribution of productivity conditional on B_i, $F(w|B_i)$. For any rank r, productivity is higher when coming from the high background.

Once w_i is realized, individuals have identical utility (and hence earnings behavior) regardless of background. Let $\bar{c}(r) \equiv (\int_{(i:r_i=r)} c_i di)/Prob(i : r_i = r)$ be the average consumption of those at conditional rank r. Equality of opportunity can be captured by weights of the form $g_i = g(c_i; \bar{c}(r_i)) = 1(c_i \leq \bar{c}(r_i))$, with $x_i^s = \bar{c}(r_i)$, $x_i^u = w_i$ and x_i^b empty. Weights are concentrated on those who consume less than average conditional on within background rank (i.e. merit).

We assume that the government cannot observe family background and hence has to base taxes and transfers on earnings z only.[29] In that case, with any $T(z)$ such that $T'(z) < 1$, conditional on rank r, individuals from an advantaged background earn more and consume more than those from a disadvantaged background. Hence, $g_i = 1$ for all individuals from the disadvantaged background and $g_i = 0$ for all individuals from the advantaged background. Next, we aggregate the weights at each z level. They imply that $\bar{G}(z)$ is the fraction of individuals from disadvantaged background earning at least z relative to the population wide fraction of individuals from disadvantaged background. This is also known as the representation index. This leads to the social welfare function proposed by Roemer et al. (2003) which counts only the utility of those from a disadvantaged background. Naturally, we expect $\bar{G}(z)$ to decrease with z as it is harder for those coming from a disadvantaged background to reach upper incomes. If the representation of individuals from a disadvantaged background is zero at the top, the top

[29] In reality, family background is observable but the real advantage derived from family background for each individual is not. As a result, family background is an imperfect tag for ability advantage derived from family background. Society could be reluctant to use such a tag because of horizontal equity concerns as discussed above. A person might come from a privileged family background and yet might not have received much parental support. Taxing such a person more (conditional on earnings) would be unfair.

tax rate should be set to maximize tax revenue. Hence, equality of opportunity provides a justification for having social welfare weights decreasing with income, which is orthogonal to the utilitarian mechanism of decreasing marginal utility of consumption.

Calibrating the weights to US intergenerational mobility: The US intergenerational income mobility statistics produced by Chetty *et al.* (2014) can be used to illustrate this discussion. Suppose we define low background as having parents coming from the bottom half of the income distribution of parents. Column (1) in Table 2 displays the fraction of individuals with parents below median income above various percentiles of the income distribution.[30] As individuals with parents below median are by definition half of the population, $\bar{G}(z)$ is simply half of column (1) and is reported in column (2). $\bar{G}(z)$ falls from 1 at percentile 0 (by definition) to .333 at the 99.9th percentile. Hence, in contrast to the standard utilitarian case where $\bar{G}(z)$ converges to zero for large z (with a concave utility function with marginal utility converging to zero), in the equality of opportunity case, $\bar{G}(z)$ converges to a positive value of 1/3 because a substantial fraction of high earners come from a disadvantaged background. $\bar{G}(z)$ appears stable within the 99th percentile as $\bar{G}(z)$ is virtually the same at the 99th percentile and the 99.9th percentile. Hence, under this equality of opportunity criterion, individuals at the 99.9th percentile are deemed no less deserving than individuals at the lower 99th percentile because they are equally likely to come from a disadvantaged background.

This has two important optimal tax consequences for top earners. First, with a Pareto parameter $a = 1.5$ and an elasticity $e = .5$, the optimal top asymptotic tax rate is $\tau = 1/(1 + a \cdot e) = .57$ in the utilitarian case and $\tau = (1 - 1/3)/(1 - 1/3 + a \cdot e) = .47$ in the equality of opportunity case.[31] Second, a society which values individuals coming from a low background would use progressive income taxation but the top tax rate would be stable within the top one percent because the representation of individuals from disadvantaged background is stable above the 99th percentile.

To illustrate these properties, Table 2, column (3) presents optimal marginal tax rates at various income levels using formula (2). The weights $\bar{G}(z)$ are from column (2). We calibrate $\alpha(z)$ using the actual distribution of income based on 2008 income tax return data, the latest year available. We use a constant elasticity $e = 0.5$ which is a mid to upper range estimate based on the literature (see Saez, Slemrod, and Giertz, 2012). Because of uncertainty in the level of

[30] These estimates are based on all US individuals born in 1980-1 with their income measured at age 30-31. In this simulation, we take a short-cut and assume these estimates hold more broadly in the full population.

[31] Naturally, in a less meritocratic society than the United States at present, $\bar{G}(z)$ for large z could possibly be smaller and the optimal top tax rate correspondingly closer to the optimal utilitarian tax rate.

e, the simulations should be considered as illustrative at best. Column (3) shows that optimal marginal tax rates are U-shaped but about constant above the 99th percentile. For comparison, columns (4) and (5) present the utilitarian weights $\bar{G}(z)$ and optimal marginal tax rates $T'(z)$ assuming a log-utility so that the welfare weight $\bar{g}(z)$ at income level z, is proportional to $1/(z - T(z))$. The utilitarian case delivers optimal tax rates that are about 10 points higher than the equality of opportunity case and significantly more progressive.

III.B Poverty Alleviation

The poverty rate, defined as the fraction of households below a given disposable income threshold (the poverty threshold) attracts substantial attention in the public debate. Hence, it is conceivable that governments aim to either reduce the poverty gap (defined as the amount of money needed to lift all households out of poverty) or reduce the poverty rate (the number of households below the poverty threshold). A few studies have considered government objectives incorporating such poverty concerns. Besley and Coate (1992) and Kanbur, Keen, and Tuomala (1994) show how adopting poverty minimization indexes affects optimal tax analysis. Importantly, their findings imply that the outcomes can be Pareto dominated. In this section, we show how generalized welfare weights can incorporate poverty alleviation considerations in the traditional optimal tax analysis while maintaining the Pareto principle.

Let us denote the poverty threshold by $\bar{c}$. Anybody with disposable income $c < \bar{c}$ is poor. Utility is taken to be a special case of our general formulation: $u_i = u(c_i - v(z_i/w_i))$.

Poverty gap minimization and implicit negative weights. Kanbur, Keen, and Tuomala (1994) derive the tax system minimizing the poverty gap. If the lowest ability agent exerts positive labor supply, the authors find that the bottom marginal tax rate should be negative as illustrated on Figure 2(a) in a (pre-tax income, post-tax income) plane. It is well-known, however, that in the welfarist case, the optimal tax rate at the bottom cannot be negative if everybody works and the lowest earnings are strictly positive (Seade, 1977). Indeed, starting from a negative bottom rate at minimum earnings level $z_b > 0$, slightly increasing the bottom marginal tax rate below z_b would be Pareto improving without violating incentive constraints: it allows the lowest productivity agent to work less, which is welfare improving, and also raises more revenue, as the marginal tax rate is negative to start with. This is a Pareto improvement.[32]

The discrepancy between the poverty gap minimization and the welfarist objective arises because the former does not take into account the disutility from work for the lowest productivity

[32]The proof is exactly symmetrical to the proof of the famous zero marginal tax rate top result.

agents. Pushing them to work more is always desirable under a poverty gap objective.

Generalized Social Welfare Weights. Let us instead consider our generalized weights approach. If the demogrant can be made bigger than $\bar{c}$, then the optimum way to fight poverty is to raise enough taxes to set the demogrant equal to $\bar{c}$. Once the poverty threshold has been attained, there is no reason to have differences in social welfare weights and hence the weights would all be equal to a fixed g for those with positive earnings so that $\bar{G}(z) = g$ for $z > 0$. Using formula (2), we would have $T'(z) = (1-g)/(1-g+\alpha(z)\cdot e(z))$, where g is set so that total taxes collected raise enough revenue to fund the demogrant $\bar{c}$. The less trivial case is when even with $g = 0$, tax revenue cannot fund a demogrant as large as $\bar{c}$. Let us denote by $\bar{z}$ the (endogenous) earnings level such that $\bar{c} = \bar{z} - T(\bar{z})$, i.e., that defines the pre-tax poverty threshold.

A natural way to capture a "poverty gap alleviation" objective is to assume that social welfare weights are concentrated among individuals with disposable income c below the poverty threshold $\bar{c}$. We can therefore specify the generalized welfare weights as follows: $g(c, z; \bar{c}) = 1$ if $c < \bar{c}$ and $g(c, z; \bar{c}) = 0$ if $c \geq \bar{c}$.[33] In this case, $x_i^u = w_i$, $x_i^s = \bar{c}$, and x^b is empty. We have $\bar{g}(z) = 0$ for $z \geq \bar{z}$ and $\bar{g}(z) = 1/H(\bar{z})$ for $z < \bar{z}$ so that $\int_0^\infty \bar{g}(z)dH(z) = 1$. Hence, we have $\bar{G}(z) = 0$ for $z \geq \bar{z}$. We obtain $\bar{G}(z) = [1 - H(z)/H(\bar{z})]/[1 - H(z)]$ for $z < \bar{z}$ (as $\bar{G}(z)[1 - H(z)] = \int_z^{\bar{z}} \bar{g}(z')dH(z')$). Applying formula (2) yields the following proposition.

Proposition 7 *The optimal tax schedule that minimizes the poverty gap is:*

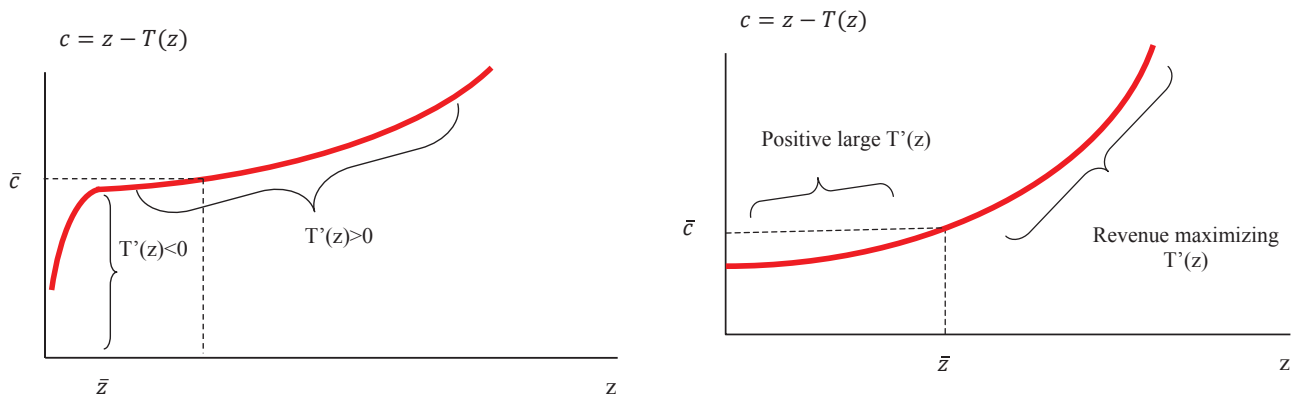
$$T'(z) = \frac{1}{1 + \alpha(z) \cdot e(z)} \quad \text{if} \quad z > \bar{z}$$

$$T'(z) = \frac{[1/H(\bar{z}) - 1]H(z)}{[1/H(\bar{z}) - 1]H(z) + \alpha(z)[1 - H(z)] \cdot e(z)} \quad \text{if} \quad z \leq \bar{z}$$

As $[1/H(\bar{z})-1]H(\bar{z}) = 1-H(\bar{z})$, the marginal tax rate is continuous at the poverty threshold $\bar{z}$. The marginal tax rate maximizes revenue above $\bar{z}$ and is positive (and typically large) below $\bar{z}$. The shape of optimal tax rates is quite similar to the standard utilitarian case and is illustrated on Figure 2(b) in a (pre-tax income, post-tax income) plane. Our approach can be viewed as minimizing the poverty gap while respecting the Pareto principle.[34]

[33] A less extreme (but still tractable) version of this assumption would set $g(c, z; \bar{c}) = \underline{g}$ if $c \geq \bar{c}$ with $0 < \underline{g} < 1$.

[34] "Poverty rate minimization," where the government attempts to minimize the number of people living below the poverty line by concentrating weights on those at the poverty threshold, is treated in online Appendix B.4.



(a) Direct poverty gap minimization

(b) Generalized weights approach

Figure 2: Optimal policies for poverty gap minimization

Notes: The figure displays the optimal tax schedule for poverty gap alleviation in a (pre-tax income z, post-tax income $c = z - T(z)$) plane. Panel (a) plots the schedule for the approach that consists in directly minimizing the poverty gap (Kanbur, Keen, and Tuomala,1994). The marginal tax rate is negative below the poverty threshold $\bar{z}$. Panel (b) plots the schedule derived using generalized welfare weights concentrated on those in poverty. The optimal tax schedule is similar to the standard utilitarian case with high marginal tax rates at the bottom.

IV Conclusion

This paper has shown that the concept of *generalized marginal social welfare weights* is a fruitful way to extend the standard welfarist theory of optimal taxation. The use of suitable generalized social welfare weights can help enrich the traditional welfarist approach and account for existing tax policy debates and structures while retaining (local) Pareto constrained efficiency. Our theory brings back social preferences as a critical element for optimal tax theory analysis. Naturally, this flexibility of generalized social weights begs the question of what social welfare weights ought to be and how they are formed.

Generalized welfare weights can be derived from social justice principles, leading to a normative theory of taxation. The most famous example is the Rawlsian theory where the generalized social marginal welfare weights are concentrated solely on the most disadvantaged members of society. As we have discussed, binary weights (equal to one for those deserving more support and zero otherwise) have normative appeal and can be used in a broad range of cases. The Rawlsian case can also be extended to a discrete number of groups, ranked according to deservedness, such that society has redistributive preferences across groups but libertarian preferences within groups. Naturally, who is deserving might itself be endogenous to the tax system. Such weights can also prioritize justice principles in a lexicographic form.

First, injustices created by tax policy (such as horizontal inequities) may have the highest

priority. In that case, those deserving of support are those discriminated against whenever horizontal inequities arise. Hence, horizontal inequities are allowed only if they help the group discriminated against, dramatically lowering the scope for such policies (such as tagging) that are recommended by the standard welfarist approach but typically not observed in practice.

Second, deserving individuals will be those who face difficult economic situations through no fault of their own. This captures the principle of compensation. Health expenses come to mind, explaining why virtually all advanced countries adopt generous public health insurance that effectively compensate individuals for the bad luck of facing high health expenses. Once disparities in health care costs have been compensated by public health insurance provision, this element naturally drops out of social welfare weights. Family background is obviously another element that affects outcomes and that individuals do not choose. This explains why equality of opportunity has wide normative appeal both among liberals and conservatives. Policies aiming directly to curb such inequities such as public education or inheritance taxation can therefore be justified on such grounds.[35] Naturally, public education or inheritance taxation cannot fully erase inequalities due to family background. This leaves a role for taxes and transfers based on income that aim at correcting remaining inequities in opportunity as in the theory of Roemer et al. (1993) which can be implemented using intergenerational mobility statistics.

Third, even conditional on background, there remains substantial inequality in incomes. Part of this inequality is due to choices (preferences for leisure vs. consumption) but part is due to luck (ability and temperament are often not based on choice). Naturally, there is a debate on the relative importance of choices vs. luck, which impacts the resulting social welfare weights.[36] As in the fair income tax theory, the generalized social welfare weights have the advantage of highlighting which differences society considers unfair (for example, due to intrinsic skill differences) and which it considers fair (for example, due to different preferences for work).[37]

Finally, there might be scope for redistribution based on more standard utilitarian principles, i.e., the fact that an additional dollar of consumption matters more for lower income individuals than for higher income individuals. In the public debate, this principle seems relevant at the low income end to justify the use of anti-poverty programs but less widely invoked to justify progressive taxation at the upper end.

Social preferences are indeed shaped by beliefs about what drives disparities in individual economic outcomes (effort, luck, background, etc.) as in the model of Piketty (1995). We show

[35] Piketty and Saez (2013b) consider optimal inheritance taxation with such "meritocratic" weights concentrated on those receiving no inheritance.

[36] The case of luck vs. deserved income is treated in online Appendix B.2.

[37] See the treatment of the fair income tax theory in online Appendix B.5.

in online Appendix C that online surveys can be used to empirically estimate social preferences, leading to a positive theory of taxation. Alternatively, the inverse optimum tax literature estimates the weights that would make the current tax system optimal. More ambitiously, economists can also cast light on those mechanisms and hence enlighten the public perceptions so as to move the debate up to the higher level of normative principles.

References

Alesina, Alberto, and George-Marios Angeletos. 2005. "Fairness and Redistribution." *American Economic Review* 95 (3): 960–980.

Besley, Timothy, and Steven Coate. 1992. "Workfare versus Welfare: Incentive Arguments for Work Requirements in Poverty-Alleviation Programs." *American Economic Review* 82(1): 249–261.

Chetty, Raj. 2009. "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods." *Annual Review of Economics* 1: 451–488.

Chetty, Raj, Nathan Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." *Quarterly Journal of Economics* 129 (4): 1553–1623.

Diamond, Peter. 1998. "Optimal Income Taxation: An Example with a U-Shaped Pattern of Optimal Marginal Tax Rates." *American Economic Review* 88 (1): 83–95.

Edgeworth, Francis. 1897. "The Pure Theory of Taxation." *Economic Journal* 7: 46–70.

Ellwood, David. 1988. *Poor Support: Poverty and the American Family.* New York: Basic Books.

Fehr, Ernst, and Klaus Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114 (3): 817–868.

Fleurbaey, Marc, and François Maniquet. 2011. *A Theory of Fairness and Social Welfare.* Cambridge: Cambridge University Press.

Fleurbaey, Marc, and François Maniquet. 2015. "Optimal Taxation Theory and Principles of Fairness." unpublished.

Hendren, Nathan. 2013. "The Policy Elasticity." National Bureau of Economic Research Working Paper 19177.

Jacquet, Laurence, and Etienne Lehmann. 2015. "Optimal Income Taxation when Skills and Behavioral Elasticities are Heterogeneous." CESifo Working Paper 5265.

Kanbur, Ravi, Michael Keen, and Matti Tuomala. 1994. "Optimal Nonlinear Income Taxation for the Alleviation of Income-Poverty." *European Economic Review* 38 (8): 1613–1632.

Kaplow, Louis. 2001. "Horizontal Equity: New Measures, Unclear Principles (Commentary)."In *Inequality and Tax Policy,* edited by Kevin A. Hassett and R. Glenn Hubbard, 75–97. Washington DC: American Enterprise Institute Press.

Kleven, Henrik, and Claus Kreiner. 2006. "The Marginal Cost of Public Funds: Hours of Work Versus Labor Force Participation." *Journal of Public Economics* 90: 1955–1973.

Mankiw, Greg, and Matthew Weinzierl. 2010. "The Optimal Taxation of Height: A Case Study of Utilitarian Income Redistribution." *American Economic Journal: Economic Policy* 2 (1): 155–76.

Mirrlees, James. 1971. "An Exploration in the Theory of Optimal Income Taxation." *Review of Economic Studies* 38: 175–208.

Piketty, Thomas. 1995. "Social Mobility and Redistributive Politics," *Quarterly Journal of Economics* 110 (3): 551–584.

Piketty, Thomas, and Emmanuel Saez. 2013. "Optimal Labor Income Taxation." In *Handbook of Public Economics,* Vol. 5, edited by Alan Auerbach, Raj Chetty, Martin Feldstein, and Emmanuel

Saez, 391-474. Amsterdam: Elsevier Science, North-Holland.
Piketty, Thomas, and Emmanuel Saez. 2013b. "A Theory of Optimal Inheritance Taxation," *Econometrica* 81 (5): 1851–1886.
Roemer, John. 1998. *Equality of Opportunity*, Cambridge: Harvard University Press.
Roemer, John et al., 2003. "To What Extent Do Fiscal Systems Equalize Opportunities for Income Acquisition Among Citizens?" *Journal of Public Economics* 87: 539–565.
Saez, Emmanuel. 2001. "Using Elasticities to Derive Optimal Income Tax Rates." *Review of Economic Studies* 68: 205–229.
Saez, Emmanuel, Joel Slemrod, and Seth Giertz. 2012. "The Elasticity of Taxable Income with Respect to Marginal Tax Rates: A Critical Review." *Journal of Economic Literature* 50 (1): 3–50.
Seade, Jesus. 1977. "On the Shape of Optimal Tax Schedules." *Journal of Public Economics* 7 (2): 203–235.
Weinzierl, Matthew. 2014. "The Promise of Positive Optimal Taxation: Normative Diversity and a Role for Equal Sacrifice." *Journal of Public Economics* 118: 128–142.

Table 1: Generalized Social Marginal Welfare Weights

	Actual Practice	Standard Welfare Criterion	Generalized Social Marginal Welfare Weights
	(1)	(2)	(3)
Pareto Efficiency	Desirable	Yes	Yes (local) if g non negative
Optimal taxes with fixed incomes	Non-degenerate	Degenerate (Full redistribution desirable)	g must depend on taxes paid in addition to consumption
Free Loaders	Important	Cannot capture counterfactuals	Can be captured if g depends on counterfactual (whether work absent transfers)
Tagging	Used minimally	Highly desirable	Can justify non desirability if g depends on horizontal equity
Poverty Alleviation	Relevant criterion	Direct poverty minimization leads to Pareto inefficiency	Can maintain Pareto efficiency

Notes: This table contrasts actual practice (column 1), the standard welfarist approach (column 2), and our generalized social marginal welfare weights approach (column 3) in various situations listed on the left-hand-side of the table. In each situation, column 3 indicates what property of social marginal welfare weights (denoted by g) is required to make this approach fit with actual tax policy practice.

Table 2: Equality of Opportunity vs. Utilitarian Optimal Tax Rates

	Equality of Opportunity			Utilitarian (log-utility)	
	Fraction from low background (=parents below median) above each percentile	Implied social welfare weight $\bar{G}(z)$ above each percentile	Implied optimal marginal tax rate at each percentile (in percent)	Utilitarian social welfare weight $\bar{G}(z)$ above each percentile	Utilitarian optimal marginal tax rate at each percentile (in percent)
	(1)	(2)	(3)	(4)	(5)
Income Percentile					
$z =$ 25th percentile	0.443	0.886	53	0.793	67
$z =$ 50th percentile	0.373	0.746	45	0.574	58
$z =$ 75th percentile	0.303	0.606	40	0.385	51
$z =$ 90th percentile	0.236	0.472	34	0.255	42
$z =$ 99th percentile	0.170	0.340	46	0.077	54
$z =$ 99.9th percentile	0.165	0.330	47	0.016	56

Notes: This table compares optimal marginal tax rates at various percentiles of the distribution (listed by row) using an equality of opportunity criterion (in column (3)) and a standard utilitarian criterion (in column (5)). Both columns use the optimal tax formula $T'(z) = [1 - \bar{G}(z)]/[1 - \bar{G}(z) + \alpha(z) \cdot e]$ discussed in the text where $\bar{G}(z)$ is the average social marginal welfare weight above income level z, $\alpha(z) = (zh(z))/(1 - H(z))$ is the local Pareto parameter (with $h(z)$ the density of income at z, and $H(z)$ the cumulative distribution), and e the elasticity of reported income with respect to $1 - T'(z)$. We assume $e = 0.5$. We calibrate $\alpha(z)$ using the actual distribution of income based on 2008 income tax return data (and ignoring the effects of changing taxes on $\alpha(z)$). For the equality of opportunity criterion, $\bar{G}(z)$ is the representation index of individuals with income above z who come from a disadvantaged background (defined as having a parent with income below the median). This representation index is estimated using the national intergenerational mobility statistics of Chetty et al. (2014) based on all US individuals born in 1980-1 with their income measured at age 30-31. For the utilitarian criterion, we assume a log-utility so that the social welfare weight $\bar{g}(z)$ at income level z is proportional to $1/(z - T(z))$.

Online Appendix for "Generalized Social Marginal Welfare Weights for Optimal Tax Theory"

By Emmanuel Saez and Stefanie Stantcheva

A Proofs of Results in the Text

A.1 Local Optimum Approach

In this section, we provide definitions of local optima and derive first and second order conditions.

We focus on money-metric utilities $u_i = c_i - v(z_i; x_i^u, x_i^b)$ (i.e., removing the common concave transformation $u(.)$) so as to express all utility gains or losses in dollar terms.

We start from an initial budget neutral tax schedule T. As in the main text, we denote by $z_i(T)$ the earnings of agent i and by $U_i(T) = z_i(T) - T(z_i(T)) - v(z_i(T); x_i^u, x_i^b)$ the indirect money-metric utility of agent i under the tax system T. We denote by $g_i(T) = g(z_i(T) - T(z_i(T)), z_i(T), x_i^s, x_i^b)$ the generalized social welfare weight under the tax system T.

For any tax schedule $\tilde{T}$, we define the budget function $B(\tilde{T}) = \int_i \tilde{T}(z_i(\tilde{T}))di$ where $\tilde{T}(z_i(\tilde{T}))$ is tax paid by individual i when the tax schedule is $\tilde{T}$. The fact that the initial tax system T is budget neutral implies that $B(T) = 0$. We define the social welfare function $W(\tilde{T}|T)$ from tax system $\tilde{T}$ using the social welfare weights evaluated at T as follows:

$$W(\tilde{T}|T) = \int_i g_i(T)U_i(\tilde{T})di. \tag{A1}$$

Since the weights $g_i(T)$ are evaluated at the initial tax system T, they are held fixed in this definition, and hence $\tilde{T} \to W(\tilde{T}|T)$ is a standard social welfare function for evaluating tax systems. Because all individual utilities are money-metric, the social marginal welfare weight on individual i is indeed $g_i(T) \cdot u_{c_i} = g_i(T)$. We can define a local tax optimum as follows:

Definition A1 ***Local Optimum.*** *The initial tax system T is a local optimum if and only if there exists a neighborhood of T such that for any budget neutral tax system $\tilde{T}$ (i.e., a tax system such that $B(\tilde{T}) = 0$) in this neighborhood, we have $W(\tilde{T}|T) \leq W(T|T)$.*

Using this formal definition, we can easily prove Proposition 1 from the main text.

As in Proposition 1, consider a perturbed tax system $T + \varepsilon\Delta T$ in direction ΔT that is budget neutral to a first order (according to Definition 2). This implies that the budget function

$\varepsilon \to R(\varepsilon) = B(T + \varepsilon\Delta T)$ is such that (a) $R(0) = 0$ (as the initial T is budget neutral, $R(0) = B(T) = 0$), (b) $R'(0) = 0$ (as the direction of reform ΔT is first order neutral budget, we have $[R(\varepsilon) - R(0)]/\varepsilon = [B(T + \varepsilon\Delta T) - B(T)]/\varepsilon) \to 0$ when $\varepsilon \to 0$). Hence the tax system $\tilde{T}_\varepsilon(z) = T(z) + \varepsilon\Delta T(z) - R(\varepsilon)$ (where $-R(\varepsilon)$ is a lumpsum adjustment) is exactly budget neutral for all ε. Indeed, as there are no income effects, earnings decisions are the same under tax system $T + \varepsilon\Delta T$ and tax system $\tilde{T}_\varepsilon$ so that $B(\tilde{T}_\varepsilon) = B(T + \varepsilon\Delta T) - R(\varepsilon) \equiv 0$.

For ε small enough, the tax system $\tilde{T}_\varepsilon = T + \varepsilon\Delta T - R(\varepsilon)$ is in the neighborhood of T from Definition A1. Because T is a local optimum, we have $W(\tilde{T}_\varepsilon|T) \leq W(T|T)$ for ε small enough and $W(\tilde{T}_{\varepsilon=0}|T) = W(T|T)$. Hence, the function $\varepsilon \to W(T + \varepsilon\Delta T - R(\varepsilon)|T)$ has a local maximum at $\varepsilon = 0$. This implies that the first and second order conditions hold. The first order condition $\frac{dW(T+\varepsilon\Delta T-R(\varepsilon)|T)}{d\varepsilon}|_{\varepsilon=0} = 0$ is exactly equivalent to the condition $\int_i g_i \Delta T(z_i) di = 0$ in the proposition.

The second order condition is $\frac{d^2W(T+\varepsilon\Delta T-R(\varepsilon)|T)}{d\varepsilon^2}|_{\varepsilon=0} \leq 0$. This second order condition does not have a simple expression but it can be checked on a case by case basis, like in standard optimal tax theory.

A.2 Derivation of the Optimal Tax Formulas using Weights

We show how to derive the optimal nonlinear tax formula (2) and the optimal linear tax formula (3) using the generalized welfare weights approach. In each case, we consider a small budget neutral tax reform. At the optimum, the net welfare effect has to be zero.

Proof of Proposition 2. Optimal non-linear tax. Consider a small reform $\delta T(z)$ in which the marginal tax rate is increased by $\delta\tau$ in a small band from z to $z + dz$, but left unchanged anywhere else. The reform mechanically collects extra taxes $dz\delta\tau$ from each taxpayer above z. As there are $1 - H(z)$ individuals above z, $dz\delta\tau[1 - H(z)]$ is collected. With no income effects on labor supply, there is no behavioral response above the small band.

Those in the income range from z to $z + dz$ have a behavioral response to the higher marginal tax rate. A taxpayer in the small band reduces her income by $\delta z = -ez\delta\tau/(1 - T'(z))$ where e is the elasticity of earnings z with respect to the net-of-tax rate $1 - T'$. As there are $h(z)dz$ taxpayers in the band, those behavioral responses lead to a tax loss equal to $-dz\delta\tau \cdot h(z)e(z)zT'(z)/(1 - T'(z))$ with $e(z)$ the average elasticity in the small band. Hence, the net revenue collected by the reform is

$$dR = dz\delta\tau \cdot \left[1 - H(z) - h(z) \cdot e(z) \cdot z \cdot \frac{T'(z)}{1 - T'(z)}\right] \tag{A2}$$

This revenue is rebated lumpsum so that the reform is budget neutral. With no income effects, this lumpsum rebate has no labor supply effect on earnings.

What is the effect of the reform on welfare using the generalized welfare weights g_i? The

welfare effect is $\int_i g_i dR di$ for $z_i \leq z$ and $-\int_i g_i(\delta\tau dz - dR)di$ for $z_i > z$. Hence, the net effect on welfare is $dR \cdot \int_i g_i di - \delta\tau dz \int_{\{i:z_i \geq z\}} g_i di$. At the optimum, the net welfare effect is zero. Using the expression for dR above and the fact that $(1 - H(z))\bar{G}(z) = \int_{\{i:z_i \geq z\}} g_i di / \int_i g_i di$, the net welfare effect can be rewritten as

$$dz\delta\tau \cdot \int_i g_i di \cdot \left[1 - H(z) - h(z) \cdot e(z) \cdot z \cdot \frac{T'(z)}{1 - T'(z)}\right] - dz\delta\tau \cdot \int_i g_i di \cdot (1 - H(z)) \cdot \bar{G}(z) = 0 \quad \text{(A3)}$$

Dividing by $dz\delta\tau \cdot \int_i g_i di$ and re-arranging, we get

$$\frac{T'(z)}{1 - T'(z)} = \frac{1}{e(z)} \cdot \frac{1 - H(z)}{z \cdot h(z)} \cdot (1 - \bar{G}(z)).$$

Using the local Pareto parameter $\alpha(z) = zh(z)/(1 - H(z))$, we obtain formula (2).

Discussion of the optimal non-linear tax formula: If average social marginal welfare weights $\bar{g}(z)$ are decreasing in income, $T'(z)$ is always non-negative. To see this, note that by definition $\bar{G}(z) = 1$ when $z = z_{\text{bottom}}$, the bottom of the earnings distribution. Next, $\bar{g}(z)$ decreasing implies that $\bar{G}(z) < 1$ for $z > z_{\text{bottom}}$. Hence, the optimal tax formula then implies that $T'(z) > 0$. The condition that social welfare weights are decreasing in income is always satisfied with standard utilitarian weights ($g_i = u_{c_i}$), or using a standard concave social welfare function $\int_i G(u_i)di$ (with $g_i = G'(u_i)u_{c_i}$). We do not impose this condition *a priori* for all primitive weights g_i in our generalized framework.

Similarly, the zero top tax rate result continues to hold if the income distribution is bounded and average social marginal welfare weights $\bar{g}(z)$ are decreasing in income. The argument is the same as in the Mirrlees model: The top tax rate cannot be positive because otherwise, we could reduce it, which would induce top tax earners to work more (and would not have an adverse effect on revenues above the top earner, since there is no agent earning more). Next, the top tax rate cannot be negative by the argument in the previous paragraph.

Finally, the zero bottom tax rate result continues to hold under the same conditions as in the Mirrlees model, namely, that the bottom earner has strictly positive earnings, in which case the result directly follows from the formula with $\bar{G}(z) = 1$ when $z = z_{\text{bottom}}$.

Proof of Proposition 3. Optimal linear tax. Consider a small reform $\delta\tau$. This increases mechanically tax revenue by $\delta\tau \cdot \int_i z_i di$. By definition of the aggregate elasticity e of $\int_i z_i di$ with respect to $1 - \tau$, this reduces tax revenue through behavioral responses by $-e \cdot \frac{\tau}{1-\tau} \cdot \delta\tau \cdot \int_i z_i di$. Hence, the net effect on revenue is $dR = [1 - e \cdot \frac{\tau}{1-\tau}] \cdot \delta\tau \cdot \int_i z_i di$. This revenue is rebated lumpsum to individuals so that the reform is budget neutral. With no income effects on labor supply, this rebate has no further impact on earnings.

What is the effect of the reform on welfare using the generalized welfare weights g_i? The welfare effect $-\int_i g_i(-dR + z_i \cdot \delta\tau)di$, or, rearranged, $dR \cdot \int_i g_i di - \delta\tau \int_i (z_i \cdot g_i)di$. At the optimum,

this is zero. Using the expression for dR above, this implies:

$$\left[1 - e \cdot \frac{\tau}{1-\tau}\right] \cdot \delta\tau \cdot \int_i z_i di \cdot \int_i g_i di = \delta\tau \cdot \int_i (z_i \cdot g_i) di$$

or equivalently

$$1 - e \cdot \frac{\tau}{1-\tau} = \bar{g} \quad \text{with} \quad \bar{g} = \frac{\int_i (z_i \cdot g_i) di}{\int_i z_i di \cdot \int_i g_i di}$$

which can easily be re-expressed as the optimal formula (3).

A.3 Proof of Proposition 4

Suppose that we have a non-negative generalized weights function and a local optimum $\tilde{T}(z)$. Consider maximizing the social welfare function SWF:

$$SWF = \int_i (\omega_i \cdot u_i) di \tag{A4}$$

with the Pareto weights such that $\omega_i = g_i / u_{c_i} \geq 0$ where g_i and u_{c_i} are evaluated at the optimum allocation (held fixed in the maximization). For any tax function, consumption is given by $c_i = z_i - T(z_i)$.

We can solve this maximization problem using again a variation approach as in Section A.2 to obtain the same optimal tax formula as in the case for generalized social welfare weights with the individual weights $\omega_i u_{c_i}$. Hence, it is clear that $\tilde{T}(z)$ satisfies the optimal tax formula (2) coming from the first order condition of the maximization of SWF in (A4). ■

A.4 Taxation with fixed incomes

Proof of Proposition 5: $\tilde{g}(z - T(z), T(z))$ has to be constant with z. Hence, setting the derivative of $\tilde{g}(z - T(z), T(z))$ with respect to z to zero, yields $\tilde{g}_c \cdot (1 - T'(z)) + \tilde{g}_{z-c} \cdot T'(z) = 0$ and the optimal tax formula (4). $0 \leq T'(z) \leq 1$ since $\tilde{g}_c \leq 0$ and $\tilde{g}_{z-c} \geq 0$. Note that this is a first-order ordinary nonautonomous differential equation of the form

$$T'(z) = f(z, T(z))$$

with initial condition on $T(0)$ given by the government budget constraint. If $\tilde{g}$ is continuous in both its arguments, so is $f(z, T(z))$ for $z \in [0, \infty)$. Then, by the Cauchy-Peano theorem, a solution $T(z)$ exists, with continuous derivative on $[0, \infty)$. If both $f(z, T(z))$ and $\frac{\partial f(z,T(z))}{\partial z}$ are continuous, then, by the uniqueness theorem of the initial value problem, the solution is unique. ■

A.5 Horizontal Equity

Derivation of the optimal differentiated tax rates:

Individual i belonging to group $m \in \{0,1\}$ chooses z_i to maximize $u_i = z_i \cdot (1-\tau_m) - v(z_i; x_i^u, x_i^b)$ so that $1-\tau_m = v_z(z_i; x_i^u, x_i^b)$ and $du_i/d\tau_m = -z_i$ (using the envelope theorem). The government maximizes $\int_{i\in 1} u_i di + \int_{i\in 2} u_i di$ subject to $\tau_1 \cdot \int_{i\in 1} z_i di + \tau_2 \cdot \int_{i\in 2} z_i di \geq E$. Denoting by $Z_m(1-\tau_m) = \int_{i\in m} z_i di$, the aggregate income in group m as a function of the net-of-tax rate $1-\tau_m$, we can form the Lagrangian (p is the multiplier):

$$L = \int_{i\in 1} u_i di + \int_{i\in 2} u_i di + p \cdot [\tau_1 \cdot Z_1(1-\tau_1) + \tau_2 \cdot Z_2(1-\tau_2) - E]$$

The first order condition in τ_m is:

$$0 = \frac{dL}{d\tau_m} = -\int_{i\in m} z_i di + p \cdot \left[Z_m - \tau_m \frac{dZ_m}{d(1-\tau_m)} \right] = -Z_m + p \cdot Z_m \left[1 - \frac{\tau_m}{1-\tau_m} \frac{1-\tau_m}{Z_m} \frac{dZ_m}{d(1-\tau_m)} \right],$$

Hence, introducing the elasticity $e_m = \frac{1-\tau_m}{Z_m} \frac{dZ_m}{d(1-\tau_m)}$, we have

$$1 = p \cdot \left[1 - \frac{\tau_m}{1-\tau_m} e_m \right] \quad \text{i.e.,} \quad \frac{\tau_m}{1-\tau_m} = \frac{1-1/p}{e_m},$$

so that, re-arranging, we obtain $\tau_m = (1-1/p)/(1-1/p+e_m)$ as in the main text. The multiplier p is set so that the government budget constraint is met. Naturally, this requires E to be below the revenue maximizing level that is obtained with $p = \infty$ and the standard revenue maximizing tax rates $\tau_m = 1/(1+e_m)$.

Proof of Proposition 6: Suppose $1/(1+e_2) \geq \tau^*$. Start with the tax system $\tau_1 = \tau_2 = \tau^*$ with τ^* below the revenue maximizing rate $1/(1+e_2)$ for group 2. Hence, any budget neutral reform with $\delta\tau_2 < 0$ requires $\delta\tau_1 > 0$. Given the structure of our weights (that load fully on group 1 which becomes discriminated against), this cannot be desirable either. Naturally, as $e_1 < e_2$, τ^* is also below the revenue maximizing rate $1/(1+e_1)$ for group 1 so that symmetrical reforms $\delta\tau_2 > 0$ and $\delta\tau_1 < 0$ are not desirable. Hence, $\tau_1 = \tau_2 = \tau^*$ is an optimum.

Let us prove that this optimum is unique. Suppose (τ_1, τ_2) is another optimum. If $\tau_1 = \tau_2$ then $\tau_1 = \tau_2 > \tau^*$ as τ^* is the smallest uniform rate raising E. Then $\delta\tau_1 = \delta\tau_2 < 0$ will typically raise revenue and benefit everybody (as the Laffer curve $\tau \to \tau \cdot (Z_1 + Z_2)$ is single peaked in τ). Hence, we can assume without loss of generality that $\tau_2 < \tau^* < \tau_1$:[1] The optimum has horizontal inequity and τ_2, τ_1 bracket τ^*. If not and $\tau_2 < \tau_1 < \tau^*$, then τ^* would not be the smallest uniform τ raising E. If $\tau^* < \tau_2 < \tau_1$ then by singlepeakedness of the Laffer curve in τ_2, decreasing τ_2 (which is above its revenue maximizing rate) would raise revenue and improve everybody's welfare. With $\tau_2 < \tau^* < \tau_1$, it must be the case that $\delta\tau_2 > 0$ does not raise revenue.

[1] The proof in the other case $\tau_2 > \tau^* > \tau_1$ proceeds the same way.

If it did, that reform with $\delta\tau_1 < 0$ would benefit group 1 where all the weight is loaded. Hence, τ_2 is above the revenue maximizing rate $1/(1+e_2)$ but this contradicts $1/(1+e_2) \geq \tau^*$.

Suppose $1/(1+e_2) < \tau^*$ and consider the tax system $\tau_2 = 1/(1+e_2)$ and $\tau_1 < \tau^*$ the smallest tax rate such that $\tau_1 \cdot \int_{i\in 1} z_i di + \tau_2 \cdot \int_{i\in 2} z_i di = E$. τ_2 maximizes tax revenue on group 2. So $\delta\tau_2 > 0$ requires $\delta\tau_1 > 0$ to balance budget and is not desirable. $\delta\tau_2 < 0$ requires $\delta\tau_1 > 0$ to budget balance and is not desirable as all the weight is loaded on group 1. $\delta\tau_1 < 0$ with $\delta\tau_2 = 0$ lowers revenue (as τ_1 is the *smallest* tax rate raising E). Hence, this is an optimum. Note that $\tau_2 = 1/(1+e_2)$ raises more revenue than $\tau_2 = \tau^*$. Hence, τ_1 does not need to be as high at τ^* to raise (combined with $\tau_2 = 1/(1+e_2)$), total revenue E so that $\tau_1 < \tau^*$.

We can prove that it is unique. First, the equitable tax system $\tau_1 = \tau_2 = \tau^*$ is not an optimum because $\delta\tau_2 < 0$ raises revenue and hence allows $\delta\tau_1 < 0$ which benefits everybody. Suppose $\tau_2 < \tau_1$ is another optimum. Then τ_2 must be revenue maximizing (if not moving in that direction while lowering τ_1 is desirable), then τ_1 must be set as in the proposition. ■

B Additional Results

B.1 Mapping Pareto weights to generalized welfare weights

Proposition B1 *For any social welfare function of the form $SWF = \int_i (\omega_i \cdot u_i) di$ with $\omega_i \geq 0$ exogenous Pareto weights, there exist generalized social welfare weights with function $g(c, z_i; x_i^s, x_i^b) = \omega_i \cdot u_{c_i}(c_i - v(z_i; x_i^b))$ with $x_i^s = i$, such that the tax system maximizing SWF is an optimum for generalized social welfare weights given the function g.*

The proof is immediate by comparing the first-order conditions for the social welfare maximization to the condition characterizing the optimum with generalized weights in (A3) in Section A.3. Note here that individual identities directly enter the welfare weights, so that, in the notation from the text, $x_i^s = i$. Alternatively, suppose the Pareto weights ω depended on i only through a set of characteristics x_i^s, so that $\omega_i = \omega(x_i^s)$. Then, again, the corresponding generalized weights are directly functions $g(c_i, z_i; x_i^s, x_i^b) = \omega(x_i^s) \cdot u_{c_i}(c_i - v(z_i; x_i^b))$.

B.2 Luck Income vs. Deserved Income

A widely held view is that it is fairer to tax income due to "luck" than income earned through effort and that it is fairer to insure against income losses beyond individuals' control.[2] Our framework can capture in a tractable way such social preferences, which differentiate income streams according to their source.[3] These preferences could, under some conditions on the

[2] See e.g., Fong (2001) and Devooght and Shokkaert (2003) for how the notion of control over one's income is crucial to identify what is deserved income and Cowell and Shokkaert (2001) for how perceptions of risk and luck inform redistributive preferences.

[3] The problem of luck vs. deserved income is also discussed in Fleurbaey (2008), chapter 3.

income processes, also provide a micro-foundation for generalized social welfare weights $\tilde{g}(c, z-c)$ increasing in $T = z - c$, as presented in Section II.A in the main text.

Suppose there are two sources of income: y^d is deserved income, due to one's own effort, and y^l is luck income, due purely to one's luck. Total income is $z = y^d + y^l$. Let us denote by Ey^l average luck income in the economy.

Consider a society with the following preferences for redistribution: Ideally, all luck income y^l should be fully redistributed, but individuals are fully entitled to their deserved income y^d. These social preferences can be captured by the following binary set of weights:

$$g_i = 1(c_i \leq z_i - y_i^l + Ey^l) \tag{A5}$$

In our notation, $x_i^s = (y_i^l, Ey^l)$, with Ey^l being an aggregate characteristic common to all agents. A person is "deserving" and has a weight of one if her tax confiscates more than the excess of her luck income relative to average luck income. Otherwise, the person receives a zero weight.[4]

Observable luck income: Suppose first that the government is able to observe luck income and condition the tax system on it, with $T_i = T(z_i, y_i^l)$. In this case, as discussed in Section I, it is necessary to aggregate the individual g_i weights in (A5) at each (z, y^l) pair. The aggregated weights are given by: $\bar{g}(z, y^l) = 1(z - T(z, y^l) \leq z - y^l + Ey^l)$, where Ey^l is a known constant, independent of the tax system. Hence, the optimum is to, first, ensure everybody's luck income is just equal to Ey^l with $T(z, y^l) = y^l - Ey^l + T(z)$ where $T(z)$ is now a standard income tax set according to formula (4), which leads to $T(z) = 0$, as society does not want to redistribute deserved income. A real-world example of luck income are health costs. Health costs are effectively negative luck income and the desire to compensate people for them leads to universal health insurance in all advanced economies.

No behavioral responses and unobservable luck income. We assume first that y^l and y^d are exogenously distributed in the population and independent of taxes (we consider below the case with elastic effort). With unobservable luck income, we make sufficient assumptions that guarantee that any change in total income is partially driven by luck income and partially by deserved income.

Assumption: $y_i^l = \alpha \bar{y}_i + \varepsilon_i$ and $y_i^d = (1 - \alpha)\bar{y}_i - \varepsilon_i$. $\bar{y}_i$ is distributed iid across agents with a density $f_{\bar{y}}(.)$, ε_i is distributed iid across agents with a density $f_\varepsilon(.)$ on $[\underline{\varepsilon}, \bar{\varepsilon}]$, and $0 < \alpha < 1$.

$\bar{y}_i$ is an individual-specific income effect that affects total income: Individuals with high $\bar{y}_i$ have both higher deserved income and higher luck income. On the other hand, α is an economy-wide share factor that determines how much luck income an individual has relative to deserved income for a given total income (say, as a function of the institutional features of the economy). ε_i is a random shock to the split between luck income and deserved income.

[4] In this illustration, we have considered the special case of binary individual weights. More generally, we could specify weights in a continuous fashion based on the difference between $y_i^l - Ey^l$ and $z_i - c_i$. Such alternative weights would also provide a micro-foundation for the function $\tilde{g}(c, z - c)$.

If luck income is not observable, taxes can only depend on total income, with $T_i = T(z_i)$. This model can provide a micro-foundation for the generalized weights $\tilde{g}(c, z-c)$ introduced in Definition 5.

If we aggregate the individual weights at each (c, z), we obtain $\tilde{g}(c, z-c) = Prob(y_i^l - Ey^l \leq z_i - c_i | c_i = c, z_i = z)$. Using the expression for luck income and the fact that $z_i = y_i^l + y_i^d = \bar{y}_i$, this can be rewritten as $\tilde{g}(c, z-c) = Prob(\varepsilon_i \leq Ey^l + (1-\alpha)z_i - c_i | c_i = c, z_i = z)$. By the independence assumption, the distribution of ε conditional on c and z is equal to the unconditional distribution. Hence,

$$\tilde{g}(c, z-c) = \int_{\underline{\varepsilon}}^{(1-\alpha)z-c+Ey^l} f_\varepsilon(x)dx = \int_{\underline{\varepsilon}}^{(1-\alpha)(z-c)-\alpha\cdot c+Ey^l} f_\varepsilon(x)dx$$

.

From the expression above, we obtain: $\frac{\partial \tilde{g}(c,z-c)}{\partial c}|_{(z-c)} = -\alpha f_\varepsilon((1-\alpha)z - c + Ey^l) < 0$. Hence, $\tilde{g}(c, z-c)$ is decreasing in its first argument c.

Next, $\frac{\partial \tilde{g}(c,z-c)}{\partial (z-c)}|_c = (1-\alpha) f_\varepsilon((1-\alpha)z - c + Ey^l) > 0$. Hence, despite the absence of behavioral effects here, the social weights depend positively on $z-c$, even controlling for c.

As in Proposition 5, the optimal tax system $T(z)$ equalizes $\tilde{g}(z - T(z), T(z))$ across all z. The presence of indistinguishable deserved income and luck income is enough to generate a non-trivial theory of optimal taxation, even in the absence of behavioral responses.

Beliefs about what constitutes luck income versus deserved income will naturally play a large role in the level of optimal redistribution with two polar cases. If all income is deserved, as libertarians believe in a well-functioning free market economy, the optimal tax is zero. Conversely, if all income were due to luck, the optimal tax is 100% redistribution. If social beliefs are such that high incomes are primarily due to luck while lower incomes are deserved, then the optimal tax system will be progressive.

Behavioral responses and unobservable luck income. If we assume that deserved income responds to taxes and transfers (for example through labor supply responses), while luck income does not, we can obtain multiple equilibria. The discussion here is heuristic. Individuals are allowed to differ in their productivity. Utility is $u_i = u(c_i - v(z_i - y_i^l, w_i))$ where w_i is productivity. In this case, $x_i^u = w_i$, $x_i^b = y_i^l$ and $x_i^s = Ey^l$, again common to all agents. We consider the linear tax case and the rest of the notation is as in Proposition 3. We also assume that individual know their luck income before they make labor supply decisions so that no individual decisions are taken under uncertainty.

Intuitively, there can be multiple locally optimal tax rates if the elasticity e of deserved income with respect to $(1-\tau)$ is sufficiently high at low tax rates and sufficiently low at high tax rates. This is expected to happen because luck income is inelastic while deserved income is elastic.

To see this, recall that the optimal linear tax is given by formula (3): $\tau = (1-\bar{g})/(1-\bar{g}+e)$.

Note that with a linear tax redistributed lumpsum, we have $c_i = (1-\tau)\cdot z_i + \tau \cdot Ez$ where Ez is the average of z in the population. Hence, $y_i^l - Ey^l \leq z_i - c_i$ is equivalent to $(Ez - z_i)\tau \leq Ey^l - y_i^l$. Therefore, we can rewrite $\bar{g}$ as:

$$\bar{g} = \frac{\int_i 1((Ez - z_i)\tau \leq Ey^l - y_i^l) \cdot z_i}{\int_i 1((Ez - z_i)\tau \leq Ey^l - y_i^l) \cdot \int_i z_i}$$

At $\tau = 0$, $\bar{g} = \frac{\int_i 1(0 \leq Ey^l - y_i^l) \cdot z_i}{\int_i 1(0 \leq Ey^l - y_i^l) \cdot \int_i z_i}$. If higher luck income is on average correlated with a higher total income, $Cov(1(0 \leq Ey^l - y_i^l), z_i) < 0$. Then, at $\tau = 0$, $\bar{g} < 1$ and hence the right-hand-side of the optimal tax formula (3) is positive so that society would like a tax rate τ higher than zero. Suppose that at $\tau = 0.5$, $e > 1$ (i.e, deserved income is very elastic and the fraction of deserved income in total income is large at medium tax rate levels such as $\tau = 0.5$). As $\bar{g} \geq 0$, the right hand side of (3) is below 0.5 so that society would like a tax rate τ below 0.5. Consequently, by continuity, there is a tax rate in the interval $[0, 0.5]$ that satisfies the optimal tax formula (3) and also satisfies the second order condition. We call this equilibrium the "low tax optimum."

Similarly, at $\tau = 0.9$, as long as $e < (1 - \bar{g}_{0.9})/9$ where $\bar{g}_{0.9}$ is the average welfare weight from formula (3) evaluated at $\tau = 0.9$ (i.e., at high tax levels, deserved income is small relative to luck income and hence total income is fairly inelastic), then we know that at $\tau = 0.9$, the right hand side of the optimal tax formula in (3) is above 0.9. Hence, by continuity, there is a point in $[0.5, 0.9]$ where the two sides are equated. Note that this equilibrium does not satisfy the second order condition: Just below this equilibrium, decreasing the tax rate is desirable while just above this equilibrium, increasing the tax rate is desirable. Hence, it is not a local optimum.

Furthermore, by the assumption that luck income is exogenous to taxes, we know that at $\tau = 1$, nobody supplies any deserved income and therefore $e = 0$. Hence, $\tau = 1$ is also an equilibrium. Whether this equilibrium is a local maximum or not depends on whether there is an additional equilibrium in $[0.9, 1)$. For instance, suppose that at $\tau = 0.95$, we have $e > (1 - \bar{g}_{0.95})\frac{5}{95}$, where $\bar{g}_{0.95}$ is the average welfare weight from formula (3) evaluated at $\tau = 0.95$. In this case, there is another equilibrium in $[0.9, 0.95]$ which is stable (i.e., satisfies the second order conditions) and, if there are no more equilibria in $[0.95, 1)$, the equilibrium at $\tau = 1$ is unstable (i.e., does not satisfy the second order condition). On the other hand, if there is no additional equilibrium in $[0.9, 1)$, then the equilibrium at $\tau = 1$ satisfies the second order conditions. These additional equilibria are also "high tax optima."

In either case, this heuristic example illustrates the possibility of having multiple equilibria with generalized social welfare weights.

Economies with social preferences favoring hard-earned income over luck income could hence end up in two possible situations. In the low tax optimum, people work hard, luck income makes up a small portion of total income and hence, in a self-fulfilling manner, social preferences tend to favor low taxes. In the alternative optimum, high taxes lead people to work less, which

implies that luck income represents a larger fraction of total income. This in turn pushes social preferences to favor higher taxes, to redistribute away that unfair luck income (itself favored by the high taxes in the first place). Thus, our framework can encompass the important multiple equilibria outcomes of Alesina and Angeletos (2005) without departing as drastically from optimal income tax techniques.[5]

Note that although each of the equilibria is locally Pareto efficient, the low tax can well Pareto dominate the high tax optima. The tax reform approach is inherently local.

B.3 Libertarianism, Rawlsianism, and Political Economy

Libertarian case. From the libertarian point of view, any individual is fully entitled to his pre-tax income and society should not be responsible for those with lower earnings. This view could for example be justified in a world where individuals differ solely in their preferences for work but not in their earning ability. In that case, there is no good normative reason to redistribute from consumption lovers to leisure lovers (exactly as there would be no reason to redistribute from apple lovers to orange lovers in an exchange economy where everybody starts with the same endowment). This can be modeled in our framework by assuming that $g_i = g(c_i, z_i) = \tilde{g}(c_i - z_i)$ is increasing in its only argument. Hence, x_i^s and x_i^b are empty. Formula (2) immediately delivers $T'(z_i) \equiv 0$ at the optimum since then $\bar{g}(z) \equiv 1$ and hence $\bar{G}(z) \equiv 1$ when marginal taxes are zero. In the standard framework, the way to obtain a zero tax at the optimum is to either assume that utility is linear in consumption or to specify a *convex* transformation of $u(.)$ in the social welfare function which undoes the concavity of $u(.)$.

Rawlsian case. The Rawlsian case is the polar opposite of the Libertarian one. Society cares most about those with the lowest earnings and hence sets the tax rate to maximize their welfare. With a social welfare function, this can be captured by a maximin criterion.[6] In our framework, it can be done instead by assuming that social welfare weights are concentrated on the least advantaged: $g_i = g(u_i - \min_j u_j) = 1(u_i - \min_j u_j = 0)$ so that neither z_i nor c_i (directly) enter the welfare weight and $x_i^s = u_i - \min_j u_j$, while x^b is empty (there could still be heterogeneity in individual characteristics as captured in x_i^u.) If the least advantaged people have zero earnings, independently of taxes, then $\bar{G}(z) = 0$ for all $z > 0$. Formula (2) then implies $T'(z) = 1/[1 + \alpha(z) \cdot e(z)]$ at the optimum. Marginal tax rates are set to maximize tax revenue so as to make the demogrant $-T(0)$ as large as possible.

Political Economy. Political economy considerations can be naturally incorporated. The

[5] In Alesina and Angeletos (2005), the preferences of the agents are directly specified so as to include a taste for "fairness," while social preferences are standard. We leave individual preferences unaffected and load the concern for fairness exclusively onto the social preferences. We find this more appealing because, first, this allows a separation between private and social preferences that do not always coincide in reality and, second, because it leaves individual preferences fully standard.

[6] Atkinson (1975) derives formally the Ralwsian optimal income tax using the maxi-min approach.

most popular model for policy decisions among economists is the median-voter model. Consider one specialization of our general model, with $u_i = u((1-\tau)z_i + \tau \int_i z_i - v(z_i; x_i^u))$. These are single peaked preferences in τ, so that the preferred tax rate of agent i is: $\tau_i = (1 - z_i/\int_i z_i)/(1 - z_i/\int_i z_i + e)$. Hence, the median voter is the voter with median income, denoted by z_m and hence the optimum has: $\tau = \frac{1-z_m/\int_i z_i}{1-z_m/\int_i z_i+e}$. Note that $\tau > 0$ when $z_m < \int_i z_i$, which is the standard case with empirical income distributions. This case can be seen as a particular case of generalized weights where all the weight is concentrated at the median voter.

B.4 Poverty Alleviation – Poverty Rate Minimization

Suppose the government cares only about the number of people living in poverty, that is the poverty rate. In that case, the government puts more value in lifting people above the poverty line than helping those substantially below the poverty line. Yet, let us assume that, in contrast to the analysis of Kanbur, Keen, and Tuomala (1994), the government also wants to respect the Pareto principle.

We can capture such an objective by considering generalized social marginal welfare weights concentrated solely at the poverty threshold $\bar{c}$. Hence $g(c, z; \bar{c}) = 0$ for c below $\bar{c}$ and above $\bar{c}$, and $g(c, z; \bar{c}) = \bar{g}$ for $c = \bar{c}$ ($\bar{g}$ is finite if a positive fraction of individuals bunch at the poverty threshold as we shall see, otherwise $g(c, z; \bar{c})$ would be a Dirac distribution concentrated at $c = \bar{c}$). This implies that $\bar{G}(z) = 0$ for $z \geq \bar{z}$ and $\bar{G}(z) = 1/[1 - H(z)]$ for $z < \bar{z}$.

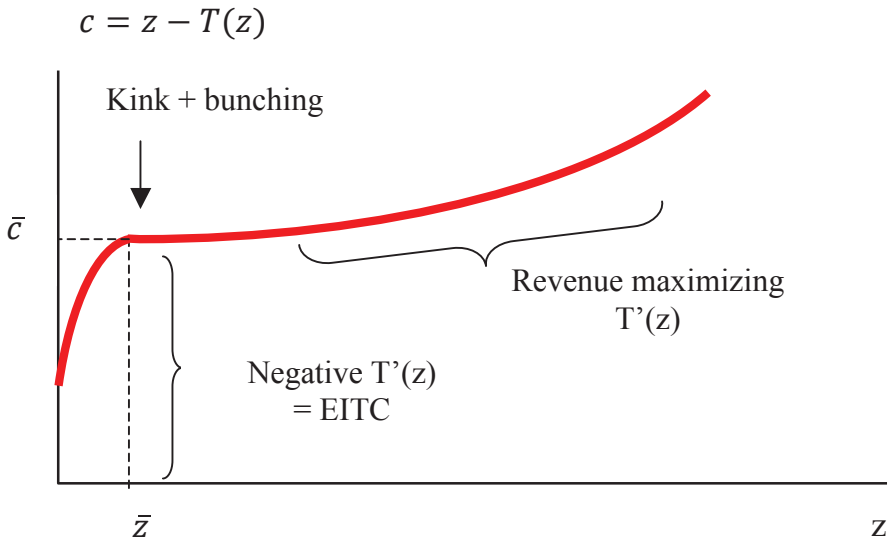
Proposition B2 *The optimal tax schedule that minimizes the poverty rate is:*

$$T'(z) = \frac{1}{1 + \alpha(z) \cdot e(z)} \quad \text{if} \quad z > \bar{z}$$

$$T'(z) = \frac{-H(z)}{-H(z) + \alpha(z)[1 - H(z)] \cdot e(z)} \quad \text{if} \quad z \leq \bar{z}$$

Hence, there is a kink in the optimal tax schedule with bunching at the poverty threshold $\bar{c}$. The marginal tax rate is Rawlsian above the poverty threshold and is negative below the poverty threshold so as to push as many people as possible just above poverty. Hence, the optimum would take the form of an EITC designed so that at the EITC maximum, earnings plus EITC equal the poverty threshold as illustrated in Figure A1. This schedule is indeed closer to the schedule obtained by Kanbur, Keen, and Tuomala (1994) than the poverty gap minimization we considered in the main text. However, in contrast to Kanbur, Keen, and Tuomala (1994), our schedule remains constrained Pareto efficient.

Figure A1: Poverty Rate Minimization



The figure displays the optimal tax schedule in a (pre-tax income z, post-tax income $c = z - T(z)$) plane for poverty rate minimization. The optimal tax schedule resembles an EITC schedule with negative marginal tax rates at the bottom.

B.5 Fair Income Taxation

The fair income taxation theory developed by Fleurbaey and Maniquet considers optimal income tax models where individuals differ in skills and in preferences for work.[7] Based on the "Compensation objective" (Fleurbaey, 1994) and the "Responsibility objective", the theory develops social objective criteria that trade-off the "Equal Preferences Transfer Principle" (at the same preferences, redistribution across unequal skills is desirable) and the "Equal Skills Transfer Principle" (at a given level of skill, redistribution across different preferences is not desirable). A trade-off arises because it is impossible to satisfy both principles simultaneously. Intuitively, the government wants to favor the hard working low skilled but cannot tell them apart from the "lazy" high skilled. In this section, we outline how one criterion of fair income tax theory (the $w_{\min}$-equivalent leximin criterion) translates into a profile of social marginal welfare weights. Our outline does not provide complete technical details. We simply reverse engineer the weights using the optimal fair income tax formula. Fleurbaey and Maniquet (2015) provide (independently) a more rigorous and complete connection between the axioms of fair income tax theory and standard optimal income taxation.[8]

[7] Fleurbaey (2008) and Fleurbaey and Maniquet (2011), chapters 10 and 11 present their fair income tax framework in detail. A number of studies in standard optimal income tax theory has also considered models with heterogeneity in both preferences and skills (see Boadway et al. 2002, Cuff, 2000, Lockwood and Weinzierl, 2015, and the surveys by Kaplow, 2008 and Boadway 2012).

[8] Our approach using formula (2) requires estimating weights by income level. It is of course not always straightforward to derive aggregated weights by income level (see Fleurbaey and Maniquet, 2015 for a discussion

We specialize our general framework to the utility function: $u_i = c_i - v(z_i/w_i, \theta_i)$ where w_i is again the skill of individual i and θ_i captures heterogeneous preferences for work. Hence labor supply is $l_i = z_i/w_i$ and it is assumed that $l \in [0,1]$ so that $l = 1$ represents full-time work. Again, formula (2) provides the optimal marginal tax rate in this model.

The $w_{\min}$-equivalent leximin criterion proposed by Fleurbaey and Maniquet puts full weight on those with $w = w_{\min}$ who receive the smallest net transfer from the government.

This criterion leads to an optimal tax system with zero marginal tax rates in the earnings range $[0, w_{\min}]$. Therefore, all individuals with earnings $z \in [0, w_{\min}]$ receive the same transfer. The optimal tax system maximizes this transfer and has positive marginal tax rate above $w_{\min}$, with $T'(z) = 1/(1 + \alpha(z) \cdot e(z)) > 0$ for $z > w_{\min}$ (Theorem 11.4 in Fleurbaey and Maniquet, 2011). Using (2), this optimal tax system implies that $\bar{G}(z) = 1$ for $0 \leq z \leq w_{\min}$, i.e., $\int_z^{\infty}[1 - g(z')]dH(z') = 0$. Differentiating with respect to z, we get $\bar{g}(z) = 1$ for $0 \leq z \leq w_{\min}$. This implies that the average social marginal welfare weight on those earning less than $w_{\min}$ is equal to one. Because the government tries to maximize transfers to those earning less than $w_{\min}$, social marginal welfare weights are zero above $w_{\min}$.[9]

This criterion, and the average weights $g(z)$ implied by it, can be founded on the following underlying generalized social marginal welfare weights. Let $T_{\max} \equiv \max_{(i:w_i=w_{\min})}(z_i - c_i)$. Formally, the weights are functions: $g_i = g(c_i, z_i; w_i, w_{\min}, T_{\max})$ where $x_i^b = w_i$, $x_i^u = \theta_i$, and $x_i^s = (w_{\min}, T_{\max})$, where $w_{\min}$ is an exogenous aggregate characteristic, while $T_{\max}$ is an endogenous aggregate characteristic. Note that, as discussed in the outline of our approach, the characteristics that appear in the utility function but not in the social welfare weights are characteristics that society does not want to redistribute accross. This is the case here for preferences for work θ_i, which are not considered fair to compensate for. This is in contrast to the "Free Loaders" case in section II.B, where the cost of work was viewed as caused by health differentials or disability, which are considered as fair to compensate for.

More precisely, the weights that rationalize the Fleurbaey-Maniquet tax system are such that: $g(c_i, z_i; w_i, w_{\min}, T_{\max}) = \tilde{g}(z_i - c_i; w_i, w_{\min}, T_{\max})$ with i) $\tilde{g}(z_i - c_i; w_i, w_{\min}, T_{\max}) = 0$ for $w_i > w_{\min}$, for any $(z_i - c_i)$ (there is no social welfare weight placed on those with skill above $w_{\min}$ no matter how much they pay in taxes) and ii) $\tilde{g}(.; w_{\min}, w_{\min}, T_{\max})$ is an (endogenous) Dirac distribution concentrated on $z - c = T_{\max}$ (that is, weights are concentrated solely on those with skill $w_{\min}$ who receive the smallest net transfer from the government). This specification forces the government to provide the *same* transfer to all those with skill $w_{\min}$. Otherwise, if an individual with skill $w_{\min}$ received less than others, all the social welfare weight would concentrate on her and the government would want to increase transfers to her. When there are agents with skill level $w_{\min}$ found at every income level below $w_{\min}$, the sole optimum is to have equal transfers, i.e., $T'(z) = 0$ in the $[0, w_{\min}]$ earnings range. Weights are zero above earnings $w_{\min}$ as $w_{\min}$-skilled individuals can at most earn $w_{\min}$, even when working full time.

of this important point).

[9] As social marginal welfare weights $\bar{g}(z)$ average to one, this implies there is a welfare weight mass at $w_{\min}$.

C Empirical Testing using Survey Data

The next step in this research agenda is to provide empirical foundations for our theory. There is already a small body of work trying to uncover perceptions of the public about various tax policies. These approaches either start from the existing tax and transfers system and reverse-engineer it to obtain the underlying social preferences (Christiansen and Jansen 1978, Bourguignon and Spadaro 2012, Zoutman, Jacobs, and Jongen 2012) or directly elicit preferences on various social issues in surveys.[10]

In this section, using a simple online survey with over 1000 participants, we elicit people's preferences for redistribution and their concepts of fairness. Our results confirm that public views on redistribution are inconsistent with standard utilitarianism. We then show how actual elicited social preferences can be mapped into generalized social marginal welfare weights.

The questions of our survey are clustered in two main groups. The first set serves to find out what notions of fairness people use to judge tax and transfer systems. We focus on the themes addressed in this paper, such as taxes paid matter (keeping disposable income constant), whether the wage rate and hours of work matter (keeping earned income constant), or whether transfer recipients are perceived to be more or less deserving based on whether they can work or not. The second set has a more quantitative ambition. As described in Section II.A, it aims at estimating whether and how social marginal welfare weights depend both on disposable income c and taxes paid T.

Our survey was conducted in December 2012 on Amazon's Mechanical Turk service, using a sample of slightly more than 1100 respondents.[11] The complete details of the survey are presented next in Section C.1. The survey asks subjects to tell which of two families (or individuals) are most deserving of a tax break (or a benefit increase). The families (or individuals) differ in earnings, taxes paid, or other attributes. The results are presented in Section C.2

C.1 Online Survey Description

Our survey was conducted in December 2012 on Amazon's Mechanical Turk service, using a sample of 1100 respondents,[12] all at least 18 years old and US citizens. The full survey is available online at `https://hbs.qualtrics.com/SE/?SID=SV_9mHljmuwqStHDOl`. The first part of the survey asked some background questions, including: gender, age, income, employment status, marital status, children, ethnicity, place of birth, candidate supported in the 2012 election, political views (on a 5-point spectrum ranging from "very conservative" to "very liberal"), and

[10] See Yaari and Bar-Hillel (1984), Frohlich and Oppenheimer (1992), Cowell and Shokkaert (2001), Fong (2001), Devooght and Schokkaert (2003), Engelmann and Strobel (2004), Ackert, Martinez-Vazquez, and Rider (2007), Gaertner and Schokkaert (2012), Weinzierl (2014), Kuziemko, Norton, Saez, and Stantcheva (2015). Our focus on tax reform and on (local) marginal welfare weights might make it much easier to elicit social preferences than if trying to calibrate a global objective function.

[11] The full survey is available online at `https://hbs.qualtrics.com/SE/?SID=SV_9mHljmuwqStHDOl`

[12] A total of 1300 respondents started the survey, out of which 200 dropped out before finishing.

State of residence. The second part of the survey presented people with sliders on which they could choose the (average) tax rates that they think four different groups should pay (the top 1%, the next 9%, the next 40% and the bottom 50%). The other questions focused on eliciting views on the marginal social welfare weights and are now described in more detail. Parts in italic are verbatim from the survey, as seen by respondents.

Utilitarianism vs. Libertarianism. The question stated: "*Suppose that the government is able to provide some families with a $1,000 tax break. We will now ask you to compare two families at a time and to select the family which you think is most deserving of the $1,000 tax break.*" Then, the pair of families were listed (see right below). The answer options given were: "*Family A is most deserving of the tax break", "Family B is most deserving of the tax break" or "Both families are equally deserving of the tax break".*

The series shown were:

Series I: (tests utilitarianism)

1) *Family A earns $50,000 per year, pays $14,000 in taxes and hence nets out $36,000.*
   Family B earns $40,000 per year, pays $5,000 in taxes and hence nets out $35,000.
2) *Family A earns $50,000 per year, pays $15,000 in taxes and hence nets out $35,000.*
   Family B earns $40,000 per year, pays $5,000 in taxes and hence nets out $35,000.
3) *Family A earns $50,000 per year, pays $16,000 in taxes and hence nets out $34,000.*
   Family B earns $40,000 per year, pays $5,000 in taxes and hence nets out $35,000.

For purely utilitarian preferences, only net income should matter, so that the utilitarian-oriented answers should be 1) B is most deserving, 2) Both are equally deserving, 3) A is most deserving. Hence utilitarian preferences should produce a large discontinuity in preferences between A and B when we move from scenario 1) to scenario 2) to scenario 3).

Series II: (tests libertarianism)

1) *Family A earns $50,000 per year, pays $11,000 in taxes and hence nets out $39,000.*
   Family B earns $40,000 per year, pays $10,000 in taxes and hence nets out $30,000.
2) *Family A earns $50,000 per year, pays $10,000 in taxes and hence nets out $40,000.*
   Family B earns $40,000 per year, pays $10,000 in taxes and hence nets out $30,000.
3) *Family A earns $50,000 per year, pays $9,000 in taxes and hence nets out $41,000.*
   Family B earns $40,000 per year, pays $10,000 in taxes and hence nets out $30,000.

For purely libertarian preferences, only the net tax burden should matter, so that the libertarian-oriented answers should be 1) A is most deserving, 2) Both are equally deserving 3) B is most deserving. Hence libertarian preferences should produce a large discontinuity in preferences between A and B when we move from scenario 1) to scenario 2) to scenario 3).

To ensure that respondents did not notice a pattern in those questions, as they might if they were put one next to each other or immediately below each other, we scattered these pairwise comparisons at different points in the survey, in between other questions.

Testing for the weight put on net income vs. taxes paid. In this part of the survey, we created fictitious households, by combining different levels of earnings and taxes paid. Each fictitious household is characterized by a pair (y, τ) where y denotes gross annual income, which could take values in $Y = \{\$10,000; \$25,000; \$50,000; \$100,000; \$200,000; \$500,000; \$1,000,000\}$ and where τ is the tax rate, which could take values in $T = \{5\%, 10\%, 20\%, 30\%, 50\%\}$. All possible combinations of (y, τ) were created for a total of 35 fictitious households. Each respondent was then shown 5 consecutive pairs of fictitious households, randomly drawn from the 35 possible ones (uniformly distributed) and ask to pick the household in each pair which was most deserving of a \$1000 tax break. As an example, a possible draw would be:

"*Which of these two families is most deserving of the \$1,000 tax break?*
Family earns \$100,000 per year, pays \$20,000 in taxes, and hence nets out \$80,000
Family earns \$10,000 per year, pays \$1,000 in taxes, and hence nets out \$9,000"

Test of utilitarianism based on consumption preferences. Utilitarian social preferences lead to the stark conclusion that people who enjoy consumption more should also receive more resources. To test this, we asked respondents:

"Which of the following two individuals do you think is most deserving of a \$1,000 tax break?
- Individual A earns \$50,000 per year, pays \$10,000 in taxes and hence nets out \$40,000. She greatly enjoys spending money, going out to expensive restaurants, or traveling to fancy destinations. She always feels that she has too little money to spend.
- Individual B earns the same amount, \$50,000 per year, also pays \$10,000 in taxes and hence also nets out \$40,000. However, she is a very frugal person who feels that her current income is sufficient to satisfy her needs."

The answer options were again that A is most deserving, B is most deserving, or that both A and B are equally deserving.

Test of Fleurbaey and Maniquet social preferences. To test whether social preferences deem hard-working people more deserving, all else equal, we asked respondents:

"Which of the following two individuals is most deserving of a \$1,000 tax break?
- Individual A earns \$30,000 per year, by working in two different jobs, 60 hours per week at \$10/hour. She pays \$6,000 in taxes and nets out \$24,000. She is very hard-working but she does not have high-paying jobs so that her wage is low.
- Individual B also earns the same amount, \$30,000 per year, by working part-time for 20 hours per week at \$30/hour. She also pays \$6,000 in taxes and hence nets out \$24,000. She has a good wage rate per hour, but she prefers working less and earning less to enjoy other, non-work activities."

The answer options were again that A is most deserving, B is most deserving or that both A and B are equally deserving.

Test of the free loaders model. To test whether the concept of free loaders presented in

the main text is relevant for social preferences, we created 4 fictitious individuals and asked people to rank them according to who they deem most deserving. Ties were allowed. The exact question was:

"*We assume now that the government can increase benefits by $1,000 for some recipients of government benefits. Which of the following four individuals is most deserving of the $1,000 increase in benefits? (...)*
- Individual A gets $15,000 per year in Disability Benefits because she cannot work due to a disability and has no other resources.
- Individual B gets $15,000 per year in Unemployment Benefits and has no other resources. She lost her job and has not been able to find a new job even though she has been actively looking for one.
- Individual C gets $15,000 pear year in Unemployment Benefits and has no other resources. She lost her job but has not been looking actively for a new job, because she prefers getting less but not having to work.
- Individual D gets $15,000 per year in Welfare Benefits and Food Stamps and has no other resources. She is not looking for a job actively because she can get by living off those government provided benefits."

C.2 Results

C.2.1 Qualitative Social Preferences

Table A1 reports preferences for giving a tax break and or a benefit increase across individuals in various scenarios.

Marginal utility of income. Panel A considers two individuals with the same earnings, same taxes, and same disposable income but who differ in their marginal utility of income. One person is described as "She greatly enjoys spending money, going out to expensive restaurants, or traveling to fancy destinations. She always feels that she has too little money to spend." while the other person is described as "She is a very frugal person who feels that her current income is sufficient to satisfy her needs." Under standard utilitarianism, the consumption loving person should be seen as more deserving of a tax break than the frugal person. In contrast, 74.4% of people report that consumption loving is irrelevant suggesting that marginal utilities driven by individual taste should not be relevant for tax policy as long as disposable income is held constant. This fits with the view described in this paper that, in contrast to welfarism, actual social welfare weights have little to do with tastes for enjoying consumption. Furthermore, in sharp contrast to utilitarianism, 21.5% think the frugal person is most deserving and only 4.4% of people report that the consumption loving person is the most deserving of a tax break. This result is probably due to the fact that, in moral terms, "frugality" is perceived as a virtue while "spending" is perceived as an indulgence.

Hard worker vs. leisure lover. Panel B considers two individuals with the same earnings, same taxes, and same disposable income but different wage rates and hence different work hours: one person works 60 hours a week at \$10 per hour while the other works only 20 hours a week at \$30 per hour. 54.4% of respondents think hours of work is irrelevant. This suggests again that for a majority (albeit a small one), hours of work and wage rates are irrelevant for tax policy as long as earnings are the same. A fairly large group of 42.7% of subjects think the hardworking low wage person is more deserving of a tax break while only 2.9% think the part-time worker is most deserving. This provides support to the fair income tax social criteria of Fleurbaey and Maniquet discussed in Section B.5. Long hours of work do seem to make a person more deserving than short hours of work, conditional on having the same total earnings.

Transfer recipients and free loaders. Panel C considers transfer recipients receiving the same benefit levels. Subjects are asked to rank 4 individuals in terms of deservedness of extra benefits: (1) a disabled person unable to work, (2) an unemployed person actively looking for work, (3) an unemployed person not looking for work, (4) a welfare recipient not looking for work. Subjects rank deservedness according to the order just listed. In particular, subjects find the disabled person unable to work and the unemployed person looking for work much more deserving than the able-bodied unemployed or welfare recipient not looking for work. This provides very strong support to the "free loaders" theory laid out in Section II.B that ability and willingness to work are the key determinants of deservedness of transfer recipients. These results are consistent with a broad body of work discussed above.

Disposable income vs. taxes paid. In the spirit of our analysis of Section II.A with fixed incomes, we analyze whether revealed social marginal welfare weights depend on disposable income and/or taxes paid. Table A2 presents non-parametric evidence showing that both disposable income and taxes paid matter and hence that subjects are neither pure utilitarians (for whom only disposable income matters) nor pure libertarians (for whom only taxed paid matter).

Panel A in Table A2 considers two families A and B with similar disposable income but dissimilar pre-tax income (and hence, different taxes paid). Family B has lower taxes and pre-tax incomes than family A. We keep family B constant and vary family A's taxes and disposable income. Overall, subjects overwhelmingly find family A more deserving than family B. To put it simply, most people find that a family earning \$50,000 and paying \$15,000 in taxes is more deserving of a tax break than a family earnings \$40,000 and paying \$5,000 in taxes. This implies that disposable income is not a sufficient statistics to determine deservedness, and that taxes paid enter deservedness positively. This contradicts the basic utilitarian model of Section II.A.

One small caveat in this interpretation is that if respondents consider consumption and labor to be complementary in utility, they might be choosing to compensate people who earn more income through higher consumption. However, as shown by Chetty (2006), labor supply fluctuations are not very correlated with consumption changes, so that consumption and labor cannot be complementary enough to explain our results.

Panel B in Table A2 considers two families A and B with similar taxes paid but dissimilar pre-tax income (and hence dissimilar disposable income as well). Family B has lower pre-tax and disposable income than family A. We again keep family B constant and vary family A taxes and disposable income. Subjects overwhelmingly find family B more deserving than family A. To put it simply, most people find that a family earning $40,000 and paying $10,000 in taxes is more deserving of a tax break than a family earnings $50,000 and paying $10,000 in taxes. This implies that taxes paid is not a sufficient statistics to determine deservedness and that disposable income affects deservedness negatively. This contradicts the basic libertarian model.

Therefore, Table A2 provides compelling non-parametric evidence that both taxes and disposable income matter for social marginal welfare weights as we posited in Section II.A.

C.2.2 Quantifying Social Preferences

Table A3 provides a first attempt at estimating the weights placed by social preferences on both disposable income and taxes paid. Recall the simple linear form discussed above, $\tilde{g}(c,T) = \tilde{g}(c-\alpha T)$, for which the optimal marginal tax rate with no behavioral effects is constant at all income levels and equal to $T' = 1/(1+\alpha)$. To calibrate α, we created 35 fictitious families, each characterized by a level of taxes and a level of net income.[13] Respondents were sequentially shown five pairs, randomly drawn from the 35 fictitious families and asked which family is the most deserving of a $1,000 tax break. This menu of choices allows us in principle to recover the social preferences $\tilde{g}(c,T)$ of each subject respondent.

Define a binary variable S_{ijt} which is equal to 1 if fictitious family i was selected during random display t for respondent j, and 0 otherwise. The regression studied is:

$$S_{ijt} = \beta_0 + \beta_T dT_{ijt} + \beta_c dc_{ijt},$$

where dT_{ijt} is the difference in tax levels and dc_{ijt} is the difference in net income levels between the two fictitious families in the pair shown during display t to respondent j. Under our assumption on the weights, $dc/dT = \alpha$ represents the slope of the (linear) social indifference curves in the (T,c) space. Families (that is, combinations of c and T) on higher indifference curves have a higher probability of being selected by social preferences. Hence, there is a mapping from the level of social utility derived from a pair (T,c) and the probability of being selected as most deserving in our survey design. The constant slope of social preferences, α, can then be inferred from the ratio $\frac{dc}{dT}|_{S=\text{constant}} = -\frac{\beta_T}{\beta_c}$. Table A3 shows the implied α and the optimal marginal tax rates in four subsamples.[14] The implied α is between 0.37 and 0.65, so that the implicit optimal marginal tax rates are relatively high, ranging from 61%

[13] Annual incomes could take one of 7 values $10K, $25K, $50K, $100K, $200K, $500K, $1 million, and taxes paid (relative to income) could take one of 5 values, 5%, 10%, 20%, 30%, and 50%.

[14] First, using the full sample and then dropping higher income groups ($1 million and above and $500K and above respectively) or the lowest income group ($10K).

to 74%. In part, this reflects our implicit assumption of no behavioral effects, which would otherwise tend to reduce the optimal tax rates at any given level of redistributive preferences. Interestingly, the implied marginal tax rates decrease when higher income fictitious families are not considered. Columns 5 and 6 highlight an interesting heterogeneity between respondents who classify themselves as "liberal" or "very liberal" (in column 5), and those who classify themselves as "conservative" or "very conservative" (in column 6). Liberals' revealed preferred marginal tax rate is 85%, while that of conservatives is much lower at 57%. Liberals put a very small weight on taxes paid relative to disposable income (column 5) while conservative put almost an equal weight on taxes paid relative to disposable income (column 6). Hence, liberals are relatively close to the utilitarian polar case while conservative are about mid-way between the utilitarian and libertarian polar cases.

This simple exercise confirms the results from Table A2 that both net income and the tax burden matter significantly for social preferences and that it is possible to determine the relative weight placed on each. More complex and detailed survey work in this spirit could help calibrate the weights more precisely.

Appendix References

Ackert, Lucy, Jorge Martinez-Vazquez, and Mark Rider. 2007. "Social Preferences and Tax Policy Design: Some Experimental Evidence," *Economic Inquiry* 45(3), 487-501.

Alesina, Alberto and George-Marios Angeletos. 2005. "Fairness and Redistribution," *American Economic Review* 95(3), 960-980.

Atkinson, Anthony B. 1975. "La 'Maxi-Min' et l'imposition optimale des revenus," *Cahiers du Séminaire d'Econométrie* 16, 73–86.

Boadway, Robin. 2012. *From Optimal Tax Theory to Tax Policy,* Cambridge: MIT Press.

Boadway, Robin, Maurice Marchand, Pierre Pestieau, and Maria del Mar Racionero. 2002. "Optimal Redistribution with Heterogeneous Preferences for Leisure," *Journal of Public Economic Theory* 4(4), 475–798.

Bourguignon, François and Amedeo Spadaro. 2012. "Tax-benefit Revealed Social Preferences," *Journal of Economic Inequality* 10(1), 75-108.

Chetty, Raj. 2006. "A New Method of Estimating Risk Aversion," *American Economic Review* 96(5), 1821-1834.

Christiansen, Vidar and Eilev Jansen. 1978. "Implicit Social Preferences in the Norwegian System of Indirect Taxation," *Journal of Public Economics* 10(2), 217–245.

Cowell, Frank and Erik Schokkaert. 2001. "Risk perceptions and distributional judgments," *European Economic Review* 45 (4-6), 941-952.

Cuff, Katherine. 2000. "Optimality of Workfare with Heterogeneous Preferences," *Canadian Journal of Economics* 33, 149–174.

Devooght, Kurt and Erik Schokkaert. 2003. "Responsibility-sensitive Fair Compensation in Different Cultures," *Social Choice and Welfare* 21, 207-242

Engelmann, Dirk and Martin Strobel. 2004. "Inequality Aversion, Efficiency and Maximin Preferences in Simple Distribution Experiments," *American Economic Review* 94(4), 857–69.

Fleurbaey, Marc. 1994. "On Fair Compensation", *Theory and Decision* 36, 277–307.

Fleurbaey, Marc. 2008. *Fairness, Responsability and Welfare*, Oxford: Oxford University Press.

Fleurbaey, Marc and François Maniquet. 2011. *A Theory of Fairness and Social Welfare*, Cambridge: Cambridge University Press.

Fleurbaey, Marc and François Maniquet. 2015. "Optimal Taxation Theory and Principles of Fairness," unpublished.

Fong, Christina. 2001. "Social Preferences, Self-interest, and the Demand for Redistribution," *Journal of Public Economics* 82(2), 225–246.

Frohlich, Norman and Joe A. Oppenheimer. 1992. *Choosing Justice: An Experimental Approach to Ethical Theory*, Berkeley University of California Press.

Gaertner, Wulf and Erik Schokkaert. 2012. *Empirical Social Choice: Questionnaire-Experimental Studies on Distributive Justice,* Cambridge: Cambridge University Press.

Kaplow, Louis. 2008. *The Theory of Taxation and Public Economics*, Princeton University Press: Princeton.

Kuziemko, Ilyana, Michael Norton, Emmanuel Saez, and Stefanie Stantcheva. 2015.

"How Elastic are Preferences for Redistribution? Evidence from Randomized Survey Experiments"*American Economic Review* 105(4), 1478–1508.

Lockwood, Benjamin and Matthew Weinzierl. 2015. "De Gustibus non est Taxandum: Heterogeneity in Preferences and Optimal Redistribution." *Journal of Public Economics* 124: 74–80.

Weinzierl, Matthew C. 2014. "The Promise of Positive Optimal Taxation: Normative Diversity and a Role for Equal Sacrifice." *Journal of Public Economics* 118, 128–142.

Yaari, Menahem E. and Maya Bar-Hillel. 1984. "On Dividing Justly," *Social Choice and Welfare* 1(1), 1–24.

Zoutman, Floris, Bas Jacobs, and Egbert Jongen. 2012. "Revealed Social Preferences of Dutch Political Parties", Tinbergen Institution Discussion Paper.

Table A1: Revealed Social Preferences

	(1)	(2)	(3)	
A. Consumption lover vs. frugal				
obs. = 1,125	Consumption lover > Frugal	Consumption lover = Frugal	Consumption lover < Frugal	
	4.1 %	74.4 %	21.5 %	
	(0.6 %)	(1.3 %)	(1.2 %)	
B. Hardworking vs. leisure lover				
obs. = 1,121	Hardworking > Leisure lover	Harworking = Leisure Lover	Harworking < Leisure Lover	
	42.7 %	54.4 %	2.9 %	
	(1.5 %)	(1.5 %)	(0.5 %)	
C. Transfer Recipients and Free Loaders				
obs. = 1,098	Disabled person unable to work	Unemployed looking for work	Unemployed not looking for work	Welfare recipient not looking for work
Average rank (1-4) assigned	1.4	1.6	3.0	3.5
	(0.018)	(0.02)	(0.023)	(0.025)
% assigned first rank	57.5 %	37.3 %	2.7 %	2.5 %
	(1.3 %)	(1.3 %)	(0.4 %)	(0.4 %)
% assigned last rank	2.3 %	2.9 %	25.0 %	70.8 %
	(0.4 %)	(0.4 %)	(1.1 %)	(1.2 %)

Notes: This table reports preferences for giving a tax break and or a benefit increase to individuals in various scenarios. Panel A considers two individuals with the same earnings, same taxes, and same disposable income but high marginal utility of income (consumption lover) vs. low marginal utility of income (frugal). In contrast to utilitarianism, 74.4% of people report that consumption loving is irrelevant and 21.5% think the frugal person is most deserving. Panel B considers two individuals with the same earnings, same taxes, and same disposable income but different wage rates and hence different work hours. 54.4% think hours of work is irrelevant and 42.7% think the hardworking low wage person is more deserving. Panel C considers out-of-work transfer recipients receiving the same benefit levels. Subjects find the disabled person unable to work and the unemployed person looking for work much more deserving than the abled bodied unemployed person or the welfare recipient not looking for work. For all statistics, standard errors are reported in parentheses below each estimate.

Table A2: Utilitarian vs. Libertarian Preferences

	(1)	(2)	(3)
A. Utilitarian Test			
	Family B: z = \$40,000, T=\$5,000, c=\$35,000		
Most deserving family	Family A: z=\$50,000 T=\$14,000 c=\$36,000	Family A: z=\$50,000 T=\$15,000 c=\$35,000	Family A: z=\$50,000 T=\$16,000 c=\$34,000
A>B	48.8 %	54.8 %	65.2 %
	(1.5 %)	(1.5 %)	(1.4 %)
A=B	38.8 %	37.3 %	28.0 %
	(1.4 %)	(1.4 %)	(1.3 %)
A<B	12.4 %	7.9 %	6.8 %
	(1.0 %)	(0.8 %)	(0.7 %)
B. Libertarian Test			
	Family B: z = \$40,000, T=\$10,000, c=\$30,000		
Most deserving family	Family A: z=\$50,000 T=\$11,000 c=\$39,000	Family A: z=\$50,000 T=\$10,000 c=\$40,000	Family A: z=\$50,000 T=\$9,000 c=\$41,000
A>B	7.7 %	3.6 %	3.1 %
	(0.8 %)	(0.6 %)	(0.5 %)
A=B	29.1 %	40.0 %	23.7 %
	(1.3 %)	(1.5 %)	(1.3 %)
A<B	63.2 %	56.4 %	73.2 %
	(1.4 %)	(1.5 %)	(1.3 %)

Notes: Sample size 1,111 subjects who finished the survey. Subjects were asked which of Family A vs. Family B was most deserving of a \$1,000 tax break in 6 scenarios with different configurations for pre-tax income z, taxes paid T, and disposable income $c = z - T$. The table reports the fraction of subjects reporting that family A is more deserving ($A > B$), families A and B are equally deserving ($A = B$), family B is more deserving ($A < B$). Standard errors are in parentheses.

Table A3: Calibrating Social Welfare Weights

	Probability of being deemed more deserving in pairwise comparison					
Sample	Full	Excludes cases with income of $1m	Excludes cases with income of $500K +	Excludes cases with income of $500K + and $10K or less	Liberal subjects only	Conservative subjects only
	(1)	(2)	(3)	(4)	(5)	(6)
d(Tax)	0.0017***	0.0052***	0.016***	0.015***	0.00082***	0.0032***
	(0.0003)	(0.0019)	(0.0019)	(0.0022)	(0.00046)	(0.00068)
d(Net Income)	-0.0046***	-0.0091***	-0.024***	-0.024***	-0.0048***	-0.0042***
	(0.00012)	(0.00028)	(0.00078)	(0.00094)	(0.00018)	(0.00027)
Number of observations	11,450	8,368	5,816	3,702	5,250	2,540
Implied α	0.37	0.58	0.65	0.64	0.17	0.77
	(0.06)	(0.06)	(0.07)	(0.09)	(0.12)	(0.16)
Implied marginal tax rate	73 %	63 %	61 %	61 %	85 %	57 %

Notes: Survey respondents were shown 5 randomly selected pairs of fictitious families, each characterized by levels of net income and tax, for a total of 11,450 observations, and asked to select the family most deserving of a $1,000 tax break. Gross income was randomly drawn from {$10K, $25K, $50K, $100K, $200K, $500K, $1 million} and tax rates from $\{5\%, 10\%, 20\%, 30\%, 50\%\}$. The coefficients are from an OLS regression of a binary variable equal to 1 if the fictitious family was selected, on the difference in tax levels and net income levels between the two families of the pair. Column (1) uses the full sample. Column (2) excludes fictitious families with income of $1 million. Column (3) excludes families with income of $500K or more. Column (4) further excludes in addition families with income below $10K. Column (5) shows the results for all families but only for respondents who classify themselves as "liberal" or "very liberal," while Colum (6) shows the results for respondents who classify themselves as "conservative" or "very conservative." The implied α is obtained as (the negative of) the ratio of the coefficient on d(Tax) over the one on d(Net income). Bootstrap standard errors in parentheses. The optimal implied constant marginal tax rate (MTR) under the assumption of no behavioral effects is, as in the text, $MTR = 1/(1+\alpha)$. The implied MTRs are high, between 61% and 74%, possibly due to the assumption of no behavioral effects. In addition, the implied MTR declines when respondents are not asked to consider higher income fictitious families. Respondents who consider themselves Liberals prefer higher marginal tax rates than those who consider themselves Conservatives.