Generalized Social Marginal Welfare Weights for Optimal Tax Theory *

Emmanuel Saez, UC Berkeley and NBER

Stefanie Stantcheva, MIT

January 2013

Abstract

This paper proposes a theory of optimal taxation using the tax reform approach and generalized social marginal welfare weights. A tax system is optimal if no budget neutral small reform can increase a weighted sum of (money metric) gains and losses across individuals. However, the social marginal welfare weights used for aggregating gains and losses are not derived from a standard social welfare function based on individual utilities but instead directly specified to reflect society's views for justice. Hence, our optimal tax formulas are no longer obtained by maximizing an objective function. Nevertheless, optimum tax formulas take the same form as standard utilitarian tax formulas by simply substituting standard marginal social welfare weights by those *generalized marginal social welfare weights*. Hence our theory nests standard theory and is equally tractable. We show how the use of suitable generalized social welfare weights can help resolve most of the puzzles of the traditional welfarist approach and account for existing tax policy debates and structures while retaining Pareto constrained efficiency. In particular, generalized welfare weights can provide a rich theory of optimal taxation even absent any behavioral responses. Generalized welfare weights can be derived from social justice principles, leading to a normative theory of taxation. We show how the most prominent alternatives to utilitarianism can be re-cast within our theory. Generalized welfare weights can also be derived from estimating actual social preferences of the public, leading to a positive theory of taxation. We use a simple online survey to illustrate this latter approach. Our theory brings back social preferences as a critical element for optimal tax theory analysis.

*Stefanie Stantcheva, MIT, stefanie@mit.edu; Emmanuel Saez, University of California, Department of Economics, 530 Evans Hall #3880, Berkeley, CA 94720, saez@econ.berkeley.edu. We thank Louis Kaplow, Ben Lockwood, Thomas Piketty, Matthew Weinzierl, and numerous seminar participants for useful discussions and comments. We acknowledge financial support from the Center for Equitable Growth at UC Berkeley.

1 Introduction

The dominant approach in optimal tax theory is to use the standard *welfarist* framework in which the government sets taxes and transfers to maximize a social welfare function which is an explicit function of individual utilities and solely of individual utilities. Social welfare is maximized subject to a government budget constraint and taking into account how individuals respond to taxes and transfers. The theory derives optimal tax formulas expressed as a function of the size of the behavioral responses to taxation and the value of redistribution. Behavioral responses to taxation are typically measured using elasticities such as the elasticity of pre-tax income with respect to tax rates and there is a wide empirical literature estimating them. The benefit of redistribution is measured by the variation of marginal social welfare weights across individuals. The marginal social welfare weight on a given individual measures the value that society puts on providing an additional dollar of consumption to this individual. With standard concave individual utilities, those marginal welfare weights decrease in after-tax income under the *utilitarian* criterion[1] so that redistributing from high to low incomes is socially valued.

Such optimal tax formulas do indeed capture the key equity-efficiency trade-off that is at the core of the public debate on the fair distribution of the tax burden and the proper level of means-tested transfers. Right-wing oriented citizens sometimes oppose redistributive taxation either because they believe that social marginal welfare weights on high earners should not necessarily be lower than social marginal welfare weights on low earners, or because they believe that redistribution discourages economic activity. More left-wing oriented citizens typically support progressive taxation both because they value redistribution and do not consider the efficiency costs to be too high. However, the standard utilitarian optimal tax approach also generates predictions for optimal taxes and transfers that are clearly at odds with actual tax systems.

First, if individuals do not respond to taxes, i.e., if pre-tax incomes are fixed, and individual utilities are concave, then utilitarianism recommends a 100% tax and full redistribution, a point originally made by Edgeworth (1897). In reality, even absent behavioral responses, many and perhaps even most people would still object to such levels of taxes on the grounds that it is unfair to fully confiscate individual incomes. Therefore, even absent behavioral responses, optimal taxation is probably not as trivial as standard utilitarian theory would suggest.

[1]The utilitarian criterion is the sum of individual utilities. It is the most commonly used one in the welfarist approach.

Second, views on taxes and redistribution seem largely shaped by views on whether the income generating process is fair and whether individual incomes are deserved or not. The public tends to dislike the redistribution of fairly earned income but is in favor of redistributing income earned unfairly or due to pure luck. Such distinctions are irrelevant for utilitarianism.[2]

Third and related, society assesses the value of transfers or the costs of taxes not only based on the actual economic situation of a given individual but also based on what this individual would have done absent that transfer or tax. For example, most people would value transfers to those unable to work (the "deserving poor") while fewer people value transfers to those who would stop working because of transfers themselves (the "undeserving free loaders") (Will 1993, Larsen 2008, Jeene, van Oorschot, and Uunk, 2011).

Fourth, under utilitarianism, optimal taxes should depend not only on income but also on all other observable characteristics which are correlated with intrinsic earning ability. Many such characteristics exist and are readily observable. Examples are height, gender, or race. Yet society seems highly reluctant to make taxes depend on such "tags".[3] Intuitively, the public would find it unfair to tax differently people with the same ability to pay. This is known as the horizontal equity concern and it cannot be easily reconciled with the welfarist approach.

Moreover, society seems to inherently value taxpayers relative to those receiving transfers, even beyond what would be captured by a person's net income. In addition, when assessing reforms, losers tend to carry more weight than winners, and it seems harder in practice to raise taxes than to lower them. Finally, the policy debate often revolves around specific but highly visible statistics such as the poverty rate so that there is a particular focus on poverty alleviation programs.

In this paper, we propose an alternative approach that nests the standard approach. Our approach does not maximize a specific social objective, yet allows to retain standard individual utilities, and is flexible enough to resolve most discrepancies between optimal tax theory and practice. It is motivated by the fact that governments certainly do not explicitly posit and maximize a social welfare function based on individual utilities. Actual tax policy debates

[2]Alternative theories of social justice have put emphasis on notions of compensation (individuals should be compensated for outcomes they are not responsible for) and responsibility (individuals should not be compensated for outcomes due to their choices). See in particular Roemer (1998), Fleurbaey and Maniquet (2006, 2007) and the comprehensive treatment of responsibility in Fleurbaey (2008) and Fleurbaey and Maniquet (2011).

[3]Akerlof (1978) pointed out that the use of such tags can help the government improve the tax system under a standard utilitarian social welfare criterion. More recently, Mankiw and Weinzierl (2010), using the case of height, point out that this reveals a deep weakness of standard welfarist optimal tax theory.

tend to focus instead on specific tax reforms, starting from a given situation, considering who the winners and losers are, and the broader consequences of the reform on economic activity and tax revenue. As is well known, the standard optimal tax approach can be recast in such tax reform terms, since the optimal tax system is precisely the one around which no further reform is desirable. Indeed, the first order conditions from maximizing a standard social welfare function always imply that a small budget neutral tax reform around the optimum has no overall impact on social welfare. In this tax reform approach, the gains and losses across individuals are aggregated using the social marginal welfare weights. The standard approach imposes a specific structure on those weights based on the individual utility functions and the social welfare function. Our approach replaces those standard weights by alternative *generalized social marginal welfare weights* that instead directly reflect society's views for justice. They are not necessarily derived from the underlying individual utility functions or the social welfare function. Our optimum tax formulas hence take the same form as standard welfarist tax formulas, but with the new *generalized social welfare weights* replacing standard social welfare weights. Therefore, our theory nests the standard theory and remains equally tractable. The optimal taxes we obtain can be seen as an equilibrium around which no marginal reform is desirable given the generalized social marginal welfare weights. Because the latter are always non-negative, our optimal taxes retain the key Pareto efficiency property.[4] At the optimum, there is no budget neutral small reform that the government could undertake and that could increase everybody's welfare, given its informational constraints.

Through a series of examples, we show how the use of suitable generalized social welfare weights can resolve most of the puzzles of the traditional utilitarian approach and account for existing tax policy debates and structures. First, we show that making generalized social marginal weights depend negatively on net taxes paid–in addition to net disposable income–eliminates the 100% tax result of Edgeworth and generates a non-trivial optimal tax theory even absent behavioral responses. In particular, this can be used to determine optimal family taxation (i.e., the treatment of couples and children) based on family equivalence scales. Second, we show that our theory allows to introduce an asymmetry in how losses vs. gains from tax reforms are valued by society. If gains from a tax reform are socially less valued than losses,

[4]Note however that due to the local nature of those weights, we can only guarantee local Pareto efficiency, that is, Pareto efficiency relative to small reforms around the status quo. The evaluation of large reforms requires non-marginal welfare weights and is left for future research.

perhaps because losers tend to complain more loudly than winners rejoice, there can be a wide range of tax equilibria rather than a single tax equilibrium. Third, our theory allows to make social weights 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. Fourth, generalized social welfare weights can also capture the fact that society prefers to tax income due to luck rather than income earned through hard work, and hence can capture the principles of compensation and responsibility that have been developed by Roemer (1998), Fleurbaey (2008), Fleurbaey and Maniquet (2011) in alternative normative theories. Fifth, we show that generalized social welfare weights can capture horizontal equity concerns as well. 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 the examples we present, the generalized welfare weights might appear ad-hoc and specified exactly so as to explain the puzzle at hand. However, we see this flexibility of our approach as a virtue that opens two avenues of investigation.

First, 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 which can also be seen as a particular case of our theory where the generalized social marginal welfare weights are concentrated solely on the most disadvantaged members of society. We show that a "locally Rawlsian" theory that endogenously and systematically divides the population into those deserving of support vs. not based on their situation and the tax/transfer system can generate a rich and realistic set of normative criteria. In particular, we show that the recent and influential social justice theories developed by Fleurbaey and Maniquet (2006, 2007, 2011) and Roemer (1998) and Roemer et al. (2003), as well as poverty alleviation objectives, can be recast in terms of such binary generalized social marginal welfare weights.

Second, generalized welfare weights could also be derived empirically, by estimating actual social preferences of the public, leading to a positive theory of taxation. There is indeed a small body of work trying to uncover perceptions of the public about various tax policies. Those 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.[5] Using a simple online survey with over 1000 participants, we illustrate how the public preferences can be mapped into generalized social marginal welfare weights. Our results confirm that the public views on redistribution are inconsistent with standard utilitarianism.

Naturally, many previous studies have proposed alternatives to the standard welfarist approach (see Kaplow (2008), Chapter 15, Fleurbaey (2008), Fleurbaey and Maniquet (2011), and Piketty and Saez (2013), Section 6 for surveys). The next section reviews those previous alternatives approaches in detail, highlighting the differences and complementarity with ours. Here, we just briefly emphasize that the most common alternative approach is to define some modified social objective but to keep the maximization of an objective principle intact. There have indeed been several attempts along these lines in the literature (Auerbach and Hassett (2002) for horizontal equity, Besley and Coate (1992) for poverty rate focus, Fleurbaey and Maniquet (2006, 2007, 2011) for responsibility, and Weinzierl (2012) for libertarianism). While such studies do succeed to some extent in extending social preferences over and above pure utilitarianism, this approach faces limitations. In fact, it greatly restricts the number of phenomena that can be captured, because it is necessary to retain individualistic social welfare functions, i.e., a welfarist objective, (as defined by Kaplow and Shavell, 2001) if one does not want to conflict with the Pareto principle.[6] Indeed, Kaplow and Shavell (2001) show that any non individualistic social function, that includes non-welfarist elements, necessarily leads to a violation of the Pareto principle in some circumstances.

Our paper is organized as follows. Section 2 outlines our approach and contrasts it with alternative approaches in the literature. Section 3 considers the simpler case with no behavioral responses to taxation. Section 4 considers the general case with behavioral responses and builds examples dealing with several specific puzzles observed in real tax policy. Section 5 shows how our approach can be reconciled with alternative concepts of social justice. Section 6 proposes empirical tests using survey data. Section 7 concludes.

[5]See Frohlich and Oppenheimer (1992), Cowell and Shoekkart (2001), Fong (2001), Devooght and Shoekkart (2003), Engelmann and Strobel (2004), Ackert, Martinez-Vazquez, and Rider (2007), Kuziemko, Norton, Saez, and Stantcheva, (2013), Weinzierl (2012b). Note also that the 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.

[6]An individualistic social welfare function is a function of individual utilities only. However, many notions of fairness cannot be accounted for through an individualistic welfare function. For example the aforementioned 'free loaders' concept, which fundamentally depends on what people *would have done* in a different tax system, or a reasonable notion of horizontal equity.

2 Outline of our Approach and Related Literature

2.1 Outline of our Approach

Before comparing our approach to the standard welfarist approach proposed in the previous literature, it is useful to outline the basic steps of our approach in the concrete case of optimal labor income taxation. Consider a discrete population of size I. Individual i has a standard utility function $u^i(c,z)$,increasing in disposable income c (individuals enjoy consumption) and decreasing in earnings z (labor effort is costly). The government sets an income tax $T(z)$ as a function of earnings so that $c = z - T(z)$. Individual i chooses z_i to maximize $u^i(z - T(z), z)$. It is easy to generalize our discussion to many commodities (and hence many potential tax bases), or to a continuous populations.

Standard welfare welfarist approach. In the standard welfarist approach, the government maximizes a social welfare function $G(u^1, .., u^I)$ that is a sole function of individual utilities. The function $G(.)$ is given exogenously and is not allowed to depend directly on the tax system. Kaplow and Shavell (2001) make the important point that including any other endogenous elements in $G(.)$ can always lead to Pareto dominated outcomes in some circumstances. The planner chooses the tax schedule $T(z)$ to maximize social welfare $G(u^1, .., u^I)$ subject to: (1) the aggregate budget constraint $\sum_i T(z^i) \geq E$ where E is exogenous (non-transfer related) government spending, (2) the fact that individual earnings z_i respond to taxes.

Using the standard envelope argument from the individual's optimization, a small tax reform $dT(z)$ changes utility $u^i(z_i - T(z_i), z_i)$ by $du^i = -u^i_c \cdot dT(z_i)$, i.e., the behavioral response dz_i can be ignored. Hence, $dT(z_i)$ measures the money-metric welfare impact of the tax reform on individual i and the net effect on social welfare is $-\sum_i G_{u^i} u^i_c \cdot dT(z_i) = -\sum_i g_i dT(z_i)$ with $g_i = G_{u^i} u^i_c$ the *social marginal welfare weight* for individual i.

Because individuals adjust their earnings z_i by dz_i following the reform, the change in taxes paid by person i is $dT(z_i) + T'(z_i)dz_i$, where $T'(z_i)dz_i$ is the fiscal change due to the behavioral response dz_i. $dT(z)$ is budget neutral if and only if $\sum_i [dT(z_i) + T'(z_i)dz_i] = 0$.

At the optimal schedule $T(z)$, no small, budget neutral tax reform $dT(z)$ exists that can increase social welfare. This implies that $\sum_i g_i dT(z_i) = 0$. To see this, if $\sum_i g_i dT(z_i) < 0$ the reform could increases social welfare. If $\sum_i g_i dT(z_i) > 0$ then $-dT(z)$ increases social welfare and is also budget neutral. Hence, the tax reform approach to optimal taxation can be stated

as follows.

Optimal tax criterion (standard welfarist approach). *If a tax system $T(z)$ is optimal, then for any budget neutral, small tax reform $dT(z)$, we have $\sum_i g_i dT(z_i) = 0$ with $g_i = G_{u^i} u_c^i$ the social marginal welfare weight on individual i.*

Knowing behavioral responses is necessary to assess whether a tax reform $dT(z)$ is budget neutral. Once this is known, assessing the welfare effects of $dT(z)$ just requires evaluating the mechanical effects of the reform on each individual (i.e., ignoring behavioral responses dz_i) and weighting them by the social marginal welfare weights $g_i = G_{u^i} u_c^i$.[7] Combining the government budget constraint condition $\sum_i [dT(z_i) + T'(z_i)dz_i] = 0$ with the social welfare condition $\sum_i g_i dT(z_i) = 0$ allows to derive optimal tax formulas expressed in terms of the social marginal welfare weights g_i as well as the behavioral responses to taxation dz_i, typically expressed in terms of elasticities (see Piketty and Saez, 2013 for a recent survey).

This small reform approach is actually much closer to what governments actually do than the abstract social welfare maximization primal approach. In effect, governments contemplating tax reforms get such reform proposals *scored* by government agencies to assess their effects on revenues and deficits, and assess which groups win or loose from the reform.[8] We naturally expect different political parties to hold different views on how to weigh gains and losses across different groups.

The standard approach however imposes strong conditions on how social welfare weights vary with the tax system and the economic circumstances of each individual. Consider the most widely used utilitarian case $G(u^1, .., u^I) = \sum_i u^i$ with separable utility $u^i(c, z) = u(c) - h^i(z)$ and uniform utility of consumption $u(c)$ across individuals. In that case, $g_i = u'(c_i)$ depends solely on the consumption level of individual i, regardless of how c_i is attained through a combination of work effort, ability, and taxation. The use of the social welfare function $G(u^1, .., u^I)$ allows some extra-flexibility but still imposes significant constraints as we shall see.

Generalized social welfare weights approach. Our approach generalizes the tax reform approach. For any tax system $T(z)$ along with all the other parameters of the economy, we can define *generalized* social marginal welfare weights $g_i \geq 0$ which measure how the government

[7] The tax reform approach only provides necessary first order conditions. This does not guarantee that the global maximum has been reached, i.e., we can be at a local extremum (see e.g., Guesnerie, 1995).

[8] For example, in the United States, the US Treasury but also the Congressional Budget Office, and the Joint Committee on Taxation routinely do such scoring. Outside non-profit organizations also provide scoring analysis.

values the marginal consumption of individual i.[9] In the following sections, using a series of examples, we show how such generalized weights can be made dependent on individual taxes and economic circumstances to capture various concepts of social justice. Using these generalized social welfare weights, we can define a tax system as optimal, in a parallel fashion to the standard approach.

Optimal tax criterion (generalized approach). *A tax system $T(z)$ is defined as optimal if and only if, for any budget neutral, small tax reform $dT(z)$, we have $\sum_i g_i dT(z_i) = 0$, with g_i the generalized social marginal welfare weight on individual i evaluated at $T(z)$.*

This approach has three key advantages relative to the standard welfarist approach and the alternatives proposed in the literature which we discuss in Section 2.2 below. First, it nests the standard welfarist approach, which is a particular case with weights $g_i = G_{u^i} u_c^i$. In particular, this implies that the standard optimum tax formulas remain valid–as optimal tax formulas are always expressed in terms of the g_i weights. Second, it ensures that any tax optimum is constrained Pareto efficient as long as the generalized weights g_i are all non-negative. To see this, the optimum is equivalent to maximizing the linear social welfare function $SWF = \sum_i \omega_i u^i$ with *Pareto weights* $\omega_i = g_i/u_c^i \geq 0$ where g_i and u_c^i are estimated at the optimum $T(z)$ (i.e., are taken as fixed in the maximization of SWF).[10] Hence, our approach can be reverse-engineered to obtain a set of Pareto weights ω_i and a corresponding standard social welfare function $\sum_i \omega_i u^i$. However, 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. Third, as we shall see, our approach allows great flexibility in the choice of the welfare weights g_i allowing us to incorporate elements that matter in actual tax policy debates and yet cannot be captured with the standard utilitarian approach. Finally, it is important to emphasize that our approach does not use a pre-specified social objective to be maximized. Hence, our optimum should be thought of as an equilibrium around which no small budget neutral reform is desirable when weighting gains and losses using the weights g_i.

[9] Strictly speaking, the weights measure only the *relative* value of consumption of individual i.

[10] There is one important caveat to note. This assumes that the first order condition characterizes the optimum. As mentioned above, that might not always be the case so that our generalized optimum may only be a local constrained Pareto optimum. We come back to this point below.

2.2 Related Literature

There have been many papers trying to move beyond utilitarianism and provide a more satisfactory treatment of social preferences for redistribution. Mirrlees himself (Mirrlees, 1974) highlighted a problem with the utilitarian approach. In the first-best, high-skilled agents "envy" low-skilled agents because they are forced to work more for the same disposable income, a result extended to the case with heterogeneous preferences by Choné and Laroque (2005). Indeed, apart from inequality aversion, captured by the concavity of the utility of consumption, there are no other fairness requirements or concepts of justice embodied in the utilitarian welfare function. The papers presented here have taken one of four possible avenues, other than ours, to remedy the shortcomings of the utilitarian framework.

First, one could focus on Pareto efficient taxation in order to find properties of optimal tax systems that are true for any social welfare function. This approach, adopted by Stiglitz (1987) and Werning (2007) in the case of the discrete and continuous optimal labor income tax models, remains very agnostic and unfortunately does not yield many practical policy recommendations as very few optimal tax results hold for any social welfare function. Our generalized weights are always non-negative, hence guaranteeing Pareto efficient taxation, but we impose more structure on them depending on social preferences we try to capture.

Second, one can try to directly augment the social welfare function to include other concerns society might have. This is the approach adopted in a recent paper by Weinzierl (2012) who argues that people, and hence society, exhibit 'normative diversity', that is, simultaneously use several normative concepts to judge policies. In addition to the utilitarian concept, he focuses on the libertarian "Equal Sacrifice" principle and considers as a social objective the minimization of a weighted sum of the utilitarian and libertarian functions. This allows him to explain why tagging is limited in the real world and why tags more strongly correlated with earning potential are used more. Our approach nests his specific normative reasoning with the weights set to equal to the derivatives of his social welfare function. Auerbach and Hassett (2002) consider a welfare function 'à la Atkinson' (1970) in which the arguments are the net consumptions of agents of different types, and there is aversion to after-tax income inequality, both across agents of the same 'type' (horizontal equity) and across agents of different income levels (vertical equity). In general, however, augmenting the social criterion with other considerations of fairness (or with any consideration other than individual welfare) can lead to Pareto inefficient outcomes

in some circumstances, which greatly restricts the generalizability of this approach. Indeed, as already mentioned in the introduction, and as shown by Kaplow and Shavell (2001) and Kaplow (2001), any non-welfarist social welfare function necessarily leads to a violation of the Pareto principle in some cases.[11] In that sense, we are nesting this approach, which consists in directly augmenting the social welfare function by extra-individualistic elements, only for the cases in which it generates Pareto efficient outcomes.[12]

Third, one can abandon the maximization of a social objective directly based on welfare and instead specify another objective, derived from specific normative concepts and core principles.[13] Fleurbaey and Maniquet use this axiomatic approach in a series of studies (Fleurbaey and Maniquet 2006, 2007, 2011). In Fleurbaey and Maniquet (2006), they consider the tradeoff between fairness and responsibility when agents differ both in their intrinsic earnings potential and their preferences for work. The two fundamental principles posited are the Pigou–Dalton principle (that transfers reducing income inequalities are acceptable, if performed between agents with the same preferences and labor supply), and the principle that "Laissez-faire" should be the social optimum if all agents had the same skill. Hence, unlike the utilitarian criterion, inequality in outcomes is acceptable if driven by different preferences for work. This leads to a measure of individual well-being in terms of an 'equivalent wage', the hypothetical wage rate that would allow an individual to reach her indifference curve at that current allocation if she could choose labor supply freely. The appropriate social criterion is then to minimize the maximal average tax rate paid by those with the lowest equivalent wage. The optimum involves the 'hard-working poor' receiving the greatest subsidy. In Section 5, we show how our generalized welfare weights can capture such a type of maximin social preferences. The three main advantages of their approach are that it does not require cardinal, inter-personal utility comparisons, that it preserves Pareto efficiency, and that the axioms postulated for social preferences are the foundation upon which the objective to be maximized is built. The difficulty is that many axioms might not always give rise to a tractable objective and that the objective

[11] This is the case for example when one introduces uncertainty (Kaplow, 2001). As explained in Kaplow (2001), in Auerbach and Hassett (2002), the welfare function depends on income differences between individuals derived using income levels in a reference distribution, and not just on individual welfare.

[12] Conversely and as we explained in Section 2.1, while it is possible to obtain our outcomes within the traditional approach *ex-post* by setting the traditional Pareto weights equal to our generalized welfare weights, it is not possible to choose the Pareto weights *ex-ante,* without knowing the tax system that arises in equilibrium.

[13] The two examples chosen here illustrate that Pareto efficiency may or may not be one of those driving principles.

needs to be re-derived from scratch for any new set of axioms considered. We show in Section 5 that Fleurbaey and Maniquet's axiomatic approach leads to a specific set of generalized social welfare weights. This allows to use standard methods to derive optimal tax formulas and hence connect the Mirrlees (1971) approach to the approach of Fleurbaey and Maniquet.

Besley and Coate (1992), as well as Kanbur, Keen, and Tuomala (1994), start from the fairness principle that everyone should be entitled to a minimal level of consumption and hence adopt as a criterion the minimization of the poverty rate, that is minimizing the number of people living below the poverty line. An issue is that their objective does not guarantee Pareto efficiency in all cases, as people could be forced to work despite prohibitively high disutility of labor, to push them above the poverty line. Our generalized social welfare weights can also be specified to capture poverty alleviation objectives (see Section 5).

The fourth method, which always guarantees that the Pareto principle holds, is to include the alternative, non-standard considerations directly into the *individual* utility functions and keep a utilitarian social welfare function. For example, Alesina an Angeletos (2005) introduce a disutility term at the individual level stemming from the amount of 'unfair' income in the economy. This however has two drawbacks. Firstly, it is not clear that social preferences are equivalent to (nor even respectful of) individual preferences. Secondly, it could lead to non-standard individual behaviors, if individual choice and social preference are non-separable in the utility function. In our framework, we remain completely agnostic about individual utilities, which can either be standard (this is the case we focus on) or incorporate any behavioral considerations one wishes (this is left for future research).

Although this paper is the first to formalize and systematically explore the concept of generalized social welfare weights, a number of studies in optimal taxation have implicitly used generalized social welfare weights. The Diamond and Mirrlees (1971) theory of optimal commodity taxation obtains formulas expressed directly in terms of social marginal welfare weights. Saez (2001, 2002) expresses optimal income tax formulas directly in terms of social marginal welfare weights and discusses (informally) how such weights can represent social preferences largely independently of individualistic utility functions.[14] Piketty and Saez (2012) also implicitly use generalized social welfare weights in the case of inheritance taxation to treat differently

[14]Recently and related, Lockwood and Weinzierl (2012) explore the effects of taste heterogeneity for optimal income taxation and show that it can substantially affect optimal tax rates through its effects on social marginal welfare weights.

in the social objective luck income (due to inheritances) from deserved income (due to labor).

3 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 the redistributive issues, and is useful as an introduction to our approach, especially as contrasted with the standard welfarist approach. We start by briefly reviewing the standard utilitarian setting.

3.1 Standard Utilitarian Approach

Consider an economy with a population normalized to one, an exogenous pre-tax earnings distribution $H(z)$, and a homogenous utility function $u(c)$ increasing and concave in disposable income c. Disposable income is equal to pre-tax earnings minus taxes so that $c = z - T(z)$. Assume that the government chooses the tax function $T(z)$ to maximize the utilitarian social welfare function:

$$SWF = \int_0^\infty u(z - T(z))dH(z) \quad \text{subject to} \quad \int T(z)dH(z) \geq 0 \quad (p),$$

where p is the Lagrange multiplier of the government budget constraint. As incomes z are fixed, a point-wise maximization with respect to $T(z)$ yields:

$$u'(z - T(z)) = p \quad \Rightarrow c = z - T(z) = \text{constant} \quad \text{across} \quad z.$$

Hence, utilitarianism with inelastic earnings and concave individual utility functions leads to complete redistribution of incomes. The government confiscates 100% of earnings and redistributes income equally across individuals (Edgeworth, 1897).

Let us denote by $g_i = u'(c_i)/p$ the social marginal welfare weight on individual i with consumption c_i. g_i measures the monetary value that society puts on an extra dollar of consumption for individual i. The optimum is such that all marginal welfare weights are set equal to one.

This simple case highlights three of the 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 100% 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, and as is well known, the utilitarian approach cannot handle well heterogeneity in individual utility functions. With heterogeneous utilities $u_i(c)$, the optimum is such that $u_i'(c_i)/p = 1$ for all i. Hence, consumption is no longer necessarily equal across individuals and is higher for individuals more able to enjoy consumption. This issue is known as the problem of inter-personal utility comparisons. In reality, society would be reluctant to redistribute based on preferences.[15]

3.2 Generalized Social Marginal Welfare Weights

A simple way to generalize the utilitarian approach is as follows. Instead of assuming that $g_i = u'(c_i)/p$, we can write directly $g_i = g(c_i, T_i)$ as a function of disposable income c_i as well as net taxes $T_i = T(z_i)$ paid by individual i.

It is natural to assume that $g(c, T)$ decreases in c to reflect the fact that society values additional consumption less (and hence accepts additional taxes more readily) for those with more disposable income, as under utilitarianism with a concave utility of consumption. This captures the old notion of "ability to pay". However, we can also assume that $g(c, T)$ increases with T as taxpayers contribute more to society's well being and are hence more deserving of additional consumption. Another interpretation is that individuals are in principle entitled to their income and hence become more deserving as the government taxes away their income. Conversely, those receiving a net subsidy from the government are perceived to be less deserving as they are debtors to society. The utilitarian case is a polar case in which g depends solely on c. The alternative polar case in which g depends solely on T would reflect the libertarian view according to which the level of one's disposable income is irrelevant and only the tax contribution matters for how socially deserving an individual is.[16]

Note that $g(c, T)$ is only defined up to a multiplicative constant. To simplify the presentation in some cases, we will assume that the latter is chosen so that the average $g(c, T)$ across the population is equal to one. Note also that we could have equivalently specified g as a function of c and z (instead of c and T as $c = z - T$).

The optimal tax system is such that no reform can increase social welfare at the margin, when the value of transfers is measured using the g weights. With no behavioral responses,

[15] Redistribution based on marginal utility is socially acceptable if there are objective reasons a person might need more disposable income, such as having a medical condition requiring high expenses, or a large family with many dependents.

[16] We assume away government funded public goods in our set-up for simplicity.

the optimal rule is very simple: the social welfare weights $g(z - T(z), T(z))$ should be constant across all income levels z. To see this, suppose by contradiction that $g(z_1 - T(z_1), T(z_1)) > g(z_2 - T(z_2), T(z_2))$. Then 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. Hence, setting the derivative of $g(z - T(z), T(z))$ with respect to z equal to zero, the optimal tax schedule is characterized by:

$$g_c \cdot (1 - T'(z)) + g_T \cdot T'(z) = 0 \quad \text{so that} \quad T'(z) = \frac{1}{1 - g_T/g_c} \tag{1}$$

where g_c and g_T denote the partial derivatives of g with respect to its first and second argument respectively. Note that $0 \leq T'(z) \leq 1$ as $g_c \leq 0$ and $g_T \geq 0$.

The standard utilitarian case, with $g(c, T) = g(c)$ implies that $T'(z) \equiv 1$ and the libertarian case with $g(c, T) = g(T)$ implies that $T'(z) \equiv 0$.

The specification $g(c, T) = g(c - \alpha T) = g(z - (1 + \alpha)T(z))$ where α is a constant parameter delivers an optimal linear tax rate with $T'(z) = 1/(1 + \alpha)$. In that case, paying one extra \$1 in taxes and consuming α more dollars leaves the person equally deserving. This means that, when earning an extra dollar, each person should be entitled to keep $\alpha/(1+\alpha)$ for consumption and pay an extra $1/(1 + \alpha)$ in taxes. This case intuitively captures the preferences of a society which finds everybody equally deserving at the margin when they are contributing a fraction $1/(1 + \alpha)$ of their incomes to taxes used to fund a uniform demogrant.

The optimal tax has an increasing marginal tax rate if $-g_T/g_c$ decreases with income z, i.e., if society feels that a higher income person should be entitled to keep a smaller fraction of her income than a lower income person.

We present in Section 6 results from a simple survey asking subjects to rank taxpayers with various incomes and tax burdens in terms of deservedness (of a tax break). We show that subjects put weight on both disposable income and the tax burden, allowing us to estimate α.

Link with the standard Pareto frontier approach. Instead of maximizing a given utilitarian social welfare function, another common approach is to derive the full Pareto frontier, by considering as an objective a weighted sum of utilities, where the weights are exogenous, nonnegative, and allowed to vary in an arbitrary fashion across individuals. Denoting by $\omega_z \geq 0$ the weight on an individual with income z, the government chooses $T(z)$ to maximize:

$$SWF\left((\omega_z)_z\right) = \int_0^\infty \omega_z u(z - T(z)) dH(z) \quad \text{subject to} \quad \int_0^\infty T(z) dH(z) \geq 0 \quad (p),$$

This program leads to the first order condition $\omega_z u'(z-T(z))/p \equiv 1$. By varying the exogenous weights (ω_z), one can recover all possible Pareto optima. However, it is not possible to specify the (ω_z) that deliver the same outcome as the generalized weights $g(c,T)$ without first solving the optimal tax using the $g(c,T)$ generalized weights. Furthermore, as we illustrate in examples below, the Pareto weights (ω_z) cannot depend directly or indirectly on tax policy, but instead only on parameters or characteristics exogenous to tax policy. Hence, specifying Pareto weights cannot be a substitute to our approach with generalized welfare weights.

Finally, it is not possible to obtain our optimum by considering a generalized social welfare function of the form

$$SWF = \int_0^\infty G(u(z-T(z)), T(z))dH(z),$$

that directly incorporates $T(z)$ in the social objective, over and beyond its effect on individual utility $z-T(z)$. Indeed, such an approach can lead to outcomes that can be Pareto dominated, as a simple thought-experiment can highlight. Suppose for example that $G(u,T)$ increases with taxes T (see the next subsection for an example of when this can realistically happen). In that case, increasing T has value per-se, independent of its effect on the individual disposable income. Hence, it could be desirable to raise taxes to burn money, i.e., have a government budget constraint that is slack and hence a situation that is Pareto dominated. Our approach using generalized non-negative marginal social welfare weights ensures that the optimum is always Pareto efficient.

3.3 Extensions

Luck income vs. deserved income: Endogenous desert criterion. An important belief society seems to hold is that it is fairer to tax income due to 'luck' than income earned through hard effort, and conversely, that it is fairer to insure against losses in income beyond an individual's control (see e.g., Fong, 2001, as well as Devooght and Shoekkart 2003 for how the notion of control over one's income is crucial to identify what is deserved income and Cowell and Shoekkart, 2001 for how perceptions of risk and luck inform redistributive preferences). Our framework can allow in a tractable way for such social preferences, which differentiate income streams according to their source. These preferences can also provide a micro-foundation for generalized social welfare weights $g(c,T)$ which are increasing in taxes T, as presented above.

Let us consider two sources of income $z = w + y$ where w is deserved income (due to one's

own effort) and y is luck income (due to one's luck). Suppose first that the government can observe separately y and w and impose separate taxes $T_y(y)$ and $T_w(w)$. Then naturally, $g(c, T_w)$ is independent of T_y because it is perceived as fair to tax lucky income. The optimum is to naturally confiscate 100% of luck income y and to tax w according to the rule derived above.

Suppose next that the government cannot observe y and w separately and can only observe total income z so that taxes have to be based on total income $T(z)$. Consider a society with sharp preferences for redistribution which considers that, ideally, all luck income should be fully redistributed, but that, by contrast, individuals are fully entitled to their deserved income. Let us denote by Ey average luck income in the economy. These social preferences can be captured by the following binary set of weights. A person is seen as deserving and has a weight of one if $c = z - T(z) \leq w + Ey$, i.e., disposable income is less than deserved income plus average luck income. Conversely, a person is seen as non-deserving and has a social weight of zero if $z - T(z) > w + Ey$. The average weight, i.e., the fraction of deserving individuals, at income level z for a given tax system (such that $z - T(z) = c$) is then given by:

$$\pi(c, z) = \int_{w+y=z} 1(c \leq w + Ey) f(w, y | w + y = z) dw dy.$$

This function is naturally decreasing in disposable income c (keeping total income z constant). Equivalently as $c = z - T$, π is increasing in T (keeping z constant). Increasing the tax on those with income z makes them more deserving on average. Hence, this model can provide a micro-foundation for the generalized weights $g(c, T) = \pi(c, c + T)$. Note importantly that, despite the absence of behavioral effects here, the social weights depend on the tax system, even controlling for c.

As above, the optimal tax system $T(z)$ is such that the average weight should be the same, across all income levels z. Hence, the presence of both (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 vs. 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.

This simple model already captures both the concept of compensation (i.e., individuals

should be compensated for outcomes such as luck income for which they are not responsible for) and the principle of responsibility (i.e., individuals should not be compensated for outcomes due to their merit or effort such as deserved income).

Taxpayers vs. transfer receivers. In practice, taxpayers–those for whom $T(z_i) > 0$–are perceived as more deserving than benefits recipients for whom $T(z_i) < 0$.[17] This can be captured by having $g(c, T)$ change discontinuously in T at $T = 0$. The optimum will then have a range of income levels for which $T(z) = 0$. Hence, there will be a transfer program at the very bottom up to z_1, and an income tax above an exemption z_2 with $z_2 > z_1$ and $T(z) = 0$ in the range (z_1, z_2). This fits with current practice where income tax exemptions for bottom earners are common and more readily accepted than direct transfers.

Family taxation. Our model can be used to discuss optimal family taxation, i.e., the treatment of couples and children, in a simple yet realistic way. Suppose $g(c, T)$ is the profile of weights for singles with no children. At the optimum, as we have seen, $T(z)$ is such that $g(z-T(z), T(z)) \equiv g$ is constant with z across all single individuals.

Consider first couples (and assuming away children). If the couple has earnings z_1, z_2, fully shares consumption, with no economies of scale in consumption for couples relative to singles, and pays total tax T then $c_1 = c_2 = (z_1 + z_2 - T)/2$. If social welfare weights only depend on consumption and the tax system (that is, society does not put intrinsic value on living as a couple as opposed to living single), the social marginal welfare weight for each member of the couple would naturally be $g((z_1 + z_2 - T)/2, T/2)$. At the optimum, those welfare weights should be equal to the (common) welfare weight g on singles and in particular on singles earning exactly the average income of the couple, $(z_1 + z_2)/2$, implying that $g((z_1 + z_2 - T)/2, T/2) = g((z_1 + z_2)/2 - T((z_1 + z_2)/2), T((z_1 + z_2)/2))$ This immediately implies that $T\left(z_1 + z_2\right)/2 = T((z_1 + z_2)/2)$. i.e., there should be perfect splitting of earnings, and then split earnings should be taxed according to the standard single schedule.[18]

Suppose now that couples do not split their incomes evenly. If couples do not share their incomes at all (and there are no economies of scale), then each member of the couple should be treated as a single individual.[19] Many countries do use such individualized tax systems. Hence,

[17] Besley and Coate (1992b) study the stigma of being a net welfare recipient and taxpayer 'resentment' towards those who do not pay taxes. Reutter et. al. (2009) document this empirically.

[18] This is the basic model of the French income tax system.

[19] If there is imperfect sharing, e.g., couples only share a fraction λ of their income, then the optimal system

even if we abstract from considerations related to the intrinsic value of living as a couple versus living as a single, our analysis is consistent with the general observation that those who believe that families fully share economic resources tend to support family-based taxation while those who believe that family members act more independently tend to support individual taxation.

Suppose now that there are economies of scale in consumption so that if the couple has a total disposable income of c, the per-person consumption equivalent is $c_1 = c_2 = (1+\delta)c/2$ with $0 \leq \delta \leq 1$, measuring the strength of economies of scale. In that case, the optimal tax T on the couple, which satisfies neutrality with respect to family choice, should be such that $g((1+\delta)(z_1+z_2-T)/2, T/2) = g$. This implies that there is a single income equivalent z such that $T(z) = T/2$ and $z - T(z) = (1+\delta)(z_1+z_2-T)/2$, i.e., the single equivalent income z is such that $(z_1+z_2)/2 = (z+\delta T(z))/(1+\delta)$. Once z is found, each member of the couple should pay $T(z)$ in taxes. In the case of a linear income tax at rate τ with demogrant R, the optimal tax on couples should also be linear at rate $\tau \cdot \frac{1+\delta}{1+\delta\tau} > \tau$ with demogrant (per person) $R/(1+\delta\tau) < R$. Effectively, couples are taxed more and receive a smaller demogrant because they benefit from economies of scales. Naturally, economies of scale are likely more important at the low-end than at the high-end of the income distribution. If δ is small for high incomes, then the tax rate on high-earning couples should asymptotically be the same as for high-earning singles.

The analysis of economies of scale can easily be extended to account for children. For simplicity, let us now ignore the distinction between couples and singles and instead focus on whether individuals have children or not. Suppose also that society does not put intrinsic value on children, so that the social welfare weights do not depend directly on whether someone has children or not.[20] Let us denote by $T(z,n)$ the tax on an individual with income z and n children. If having n children absorbs a fraction $\delta(n)$ of one's resources ($\delta(n)$ can be made dependent on the level of resources z as well), then the optimal neutral tax/transfer on a parent with n children should be such that $g((z-T(z,n))\cdot(1-\delta(n)), T(z,n)) = g$. For any pair (z,n) there is a "no child" equivalent income z^n such that $z^n - T(z^n,0) = (z-T(z,n))\cdot(1-\delta(n))$ and $T(z,n) = T(z^n,0)$. Hence, given the optimal tax schedule for childless individuals, we can

is to assume that member i earns $z_i' = \lambda(z_1+z_2)/2 + (1-\lambda)z_i$ and tax each member of the couple as if he/she were earning z_i'.

[20] Naturally, since having children affects both the resources and consumption available to a family, it enters the social marginal welfare weights indirectly.

derive the optimal tax schedule for those with children.

This discussion shows that our conceptual framework can easily be mapped into the equivalence scales used in applied welfare economics to obtain a simple and realistic theory of the taxation of families, even abstracting from behavioral responses. The standard utilitarian model cannot handle the issue of family taxation in a satisfactory way for two reasons (see also Kaplow, 2008, chapter 12 for a detailed review of the utilitarian approach to family taxation). First, without behavioral responses, the optimum always implies full redistribution so that the problem becomes degenerate, unlike in our setup. Second, with behavioral responses, the problem quickly becomes intractable and calls for redistribution across family types due to arguments related to tagging, which violate horizontal equity concerns (see below).[21]

Tax increases vs. tax decreases. Often, during actual tax reform discussions, potential losers tend to complain more loudly than potential winners rejoice. To capture such effects, we need to specify two distinct sets of social marginal weights, $g^+(c,T) > g^-(c,T) \geq 0$, where g^+ is the social marginal weight for tax increases and g^- is the social marginal weight for tax decreases.[22] This example shows that generalized weights can be made dependent not only on the current $T(z)$ but also on the nature of the contemplated small reform $dT(z)$ (in the present case the sign of $dT(z)$ at the individual level).[23] There will now be an interval set of equilibria instead of a single equilibrium. Any tax system such that $g^+(z-T(z),T(z)) \geq g^-(s-T(s),T(s))$ for all z,s is an equilibrium.

Note that such asymmetric effects can also be used to capture lack of trust in the government whereby the public does not trust that the proposed reform can actually be implemented under as favorable terms as advertised. For example, if the public believes that part of taxes collected and then redistributed are captured or wasted by the government (through inefficient bureaucracies, rewards to special interests or lobby groups, etc.), this would lead the public to discount projected gains for winners and inflate projected losses for losers.

This example also highlights one potential danger, which is the non-existence of an equi-

[21]For example, if singles are poorer on average than married individuals, and the marriage decision is inelastic to taxes, a utilitarian criterion calls for a large transfer from married to single until their average economic resources are equated. This is clearly not the way actual governments design family taxation.

[22]Note that in this situation, the two sets of weights cannot possibly both average to 1. If we normalize the set g^- such that $\int g_i^-(c,T)\,d\nu(i) = 1$, then the weights g^+ will average to more than 1. This makes sense if we consider that the weights g^- measure the value of \$1 uniformly distributed, while g^+ measure the value of \$1 uniformly taken. In a society averse to tax increases *per se*, the latter is costlier.

[23]Dependence on the small reform itself is crucial to capture horizontal equity concerns, as we show below.

librium for some sets of differential welfare weights, that is welfare weights depending on the nature of the reform itself. Here for example there would be no equilibrium if winners were weighted more heavily than losers (i.e., if $g^+(c,T) < g^-(c,T)$ for all c and T), since it would always be desirable to have additional transfers.

Conversely, it is also possible for *any* tax system to be an equilibrium if society is sufficiently averse to tax increases. An example is the case in which the weights only depend on c (that is: $g^+(c,T) = g^+(c)$ and $g^-(c,T) = g^-(c)$) and for every c_1 and c_2 such that $c_1 > c_2$, we have: $g^+(c_2) > g^+(c_1) > g^-(c_2) > g^-(c_1)$. Put differently, no matter how little taxes someone is paying, raising taxes is undesirable, even if it is done in order to give it to someone poorer.[24]

Making weights depend on non-income characteristics: Exogenous desert criterion with an application to equality of opportunity. One of the advantages of our approach is that we can allow the social marginal welfare weights to depend on any characteristics which are deemed relevant by society–other than disposable income–and to nevertheless preserve Pareto efficiency. This model can be usefully applied whenever individuals differ along several dimensions and society perceives some of these differences as unfair but others as fair. Across characteristics which are deemed to be unfairly distributed (for example, family background), society's preferences are redistributive (and *locally Rawlsian* in the case of a binary characteristic such as disadvantaged versus advantaged background). On the other hand, across characteristics where disparity is deemed to be fair (for example, income conditional on family background), society's preferences are libertarian.

Suppose that there is a number M of distinct groups of people which society can rank in terms of desert or merit, with group 1 being the most valued group and group M the least valued group. These social preferences are taken as given and may or may not be related to earnings potential, although most reasonable social preferences will likely focus on characteristics related to earnings. As an example, groups could be defined by family background, with children from advantaged backgrounds being privileged all throughout life (see Roemer et. al. 2003 and the example in the next section). Let P_m be the proportion of people from group m in the population. The density of individuals in group m who earn income z is given by $f_m(z)$, which is independent of the tax system due to the absence of behavioral effects. Within each group m,

[24]This type of situation can lead to a dead-lock in policy making, where any status quo, even if very suboptimal, can be sustained.

we can define the marginal social welfare weights as above, as a function of disposable income and taxes $g_m(z-T,T)$. Society's preferences imply that, at any given disposable income and tax level, a transfer to a group deemed more deserving is more valuable than a transfer to a less deserving group, i.e., $g_m(z-T,T) \geq g_n(z-T,T)$ for all z,T and all $m<n$. Let the fraction of members from group m at income level z, be denoted by $f(m|z) = f_m(z) P_m / \sum_{j=1}^{M} f_j(z) P_j$. The average welfare weight at income level z is then:

$$g(z-T(z),T(z)) = \sum_{m=1}^{M} f(m|z)\, g_m(z-T(z),T(z)) \tag{2}$$

Consider two cases. First, suppose that all inequality within a group is viewed as fairly based on merit and no redistribution is desired within groups. Inequality across groups by contrast is viewed as unfair. This implies that society places equal weights on all people (at all income levels) within a group but places higher weights on people with the same income level but in more deserving groups. Put differently, society's preferences are redistributive across groups,[25] and fully libertarian within a given group. Given that all individuals within a group are equally valued, we have $g_m(z-T(z),T(z)) = g_m(T(z))$, decreasing in T. If there can be group specific tax schedules T^m, (i.e., if there was the possibility of 'perfect tagging'), then the optimal tax system involves zero marginal taxes (i.e., constant taxes or transfers $T^m(z) = T^m$ for all z) within groups. Furthermore, the tax and transfers will be set such that there is perfect equalization of the (decreasing) $g_m(T^m)$ weights across groups. At any income level, this will involve taxing less individuals which are considered more deserving. On the other hand, suppose taxes can only depend on income and not on group belonging. At the optimum, given the absence of behavioral responses, the weights $g(z-T(z),T(z)) = \sum_{m=1}^{M} f(m|z)\, g_m(T(z))$ need to be equalized across income levels which involves higher taxes for income levels which are likely to have been earned by people in less deserving groups. This is a "reversed tagging" situation, in which, instead of using the group as a tag for income, we use income itself as a tag for group belonging (being able to do so is of course due to no behavioral responses).[26]

[25] With only two groups or with several groups where only one group is valued, we have a *locally Rawlsian* case, in which we only care about the least advantaged group. With more than two groups, there are still positive weights on groups less valued than the least advantaged one, which we call redistributive preferences.

[26] As an application of this 'reverse tagging', consider groups based on how fairly people earn their income. In the extreme, group 1 could be pure rent-seekers who exert no productive effort but instead extract resources from others. Group 0 could be hardworking people. If rent-seekers exhibit higher incomes, society might use high incomes as a reverse tag for rent-seeking and want high taxes on top incomes. In a less stark way, this sort of argument is sometimes heard, especially in dysfunctional economies, when the elite is considered to be

Secondly, and more realistically, suppose that society also cares about within-group inequality so that welfare weights within the groups, $g_m\left(z-T\left(z\right),T\left(z\right)\right)$, are strictly decreasing in disposable income. If taxes can be made dependent on the group, the optimal tax schedule for group m, $T^{*m}\left(z\right)$ would be set according to (1), that is: $T^{*m\prime}(z)=-g_{mc}/\left(g_{mT}-g_{mc}\right)$ so that weights are fully equalized within each group.[27] The level of the group-specific demogrants, or equivalently, of inter-group transfers are set to equalize all group-specific weights $g_m\left(T^{*m}\right)$. On the other hand, if taxes can only depend on total income, they will be set according to (1), using as weights g as defined in (2).

As an application, consider family background, ranked from poorest ($m=1$) to richest ($m=M$).[28] While low earners are valued more, those coming from poorer family backgrounds are viewed as more deserving at any income level. Since a better background provides a boost to earnings all throughout life, $f\left(m|z\right)/f\left(n|z\right)$ for $m<n$ will be decreasing in z. Hence, higher income levels act as an imperfect tag for a better family background and will be taxed more, even if within a given family background, society has perfectly libertarian preferences.[29]

4 Optimal Tax Theory with Behavioral Responses

In this section, we introduce behavioral responses. To keep the presentation as simple as possible, we consider the case with no income effects, so that the utility function of individual i takes the form $u(c-h_i(z))$ where $h_i(z)$ is the disutility of earning z. We assume that $h_i'(z)>0$, $h_i''\left(z\right)>0$ and and $u'(.)>0$, $u''\left(\right)<0$. The absence of income effects simplifies the presentation. We focus on linear taxes, at a rate τ and assume that all tax revenues are rebated as a uniform demogrant R. Hence, tax policy can be summarized by the one dimensional variable τ.

Faced with a linear tax rate τ, individual i chooses his income z_i so as to maximize his utility $u\left(z_i(1-\tau)+R-h_i(z_i)\right)$. Thus, $h_i'(z_i)=1-\tau$, and we can rewrite taxable income as a function of the net retention rate $(1-\tau)$ only, i.e. $z_i=z_i(1-\tau)$.[30] Aggregating over the

extractive. Piketty, Saez and Stantcheva (2011) also show that within the standard framework, the presence of rent-seeking among top earners tends to push up the top tax rate.

[27]This determines the tax schedule within group m only up to a constant, which is the group-specific demogrant, $T^m\left(0\right)$. Given the tax system T^{*m}, this leads to a common, equalized group-specific welfare weight $g_m\left(T^{*m}\right)$.

[28]Income of individual i in group m might for example take the form $z_i=m+l_i$ where l_i is an inelastic and heterogeneous labor supply (due to individual preferences, assumed orthogonal to family background).

[29]The next section will consider the exogenous desert model in a more realistic fashion, including behavioral responses.

[30]This is due to the absence of income effects.

population, total earnings, denoted by $Z(1-\tau)$, are also a function of the net retention rate $(1-\tau)$ with elasticity $e = [(1-\tau)/Z]dZ/d(1-\tau)$. Using the government's budget constraint, the lumpsum demogrant is then $R = \tau Z(1-\tau)$.

4.1 Standard Utilitarian Approach

In the standard utilitarian approach, the government chooses τ and R to maximize

$$SWF = \int_i u((1-\tau)z_i(1-\tau) + R - h_i(z_i(1-\tau)))d\nu(i) \quad \text{s.t.} \quad R \leq \tau Z(1-\tau) \quad (p)$$

with p the multiplier on the budget constraint, and $\nu(i)$ the CDF of individual types i. Using the envelope theorem applied to each individual i's utility maximization problem, the first order conditions with respect to R and τ for the government are simply:

$$\int u_i' d\nu(i) = p,$$

$$\int_i u_i' z_i d\nu(i) = p\left[Z - \frac{dZ}{d(1-\tau)}\right].$$

Denoting by $g_i = u_i'/p$ the normalized social marginal welfare weight on person i, the first equation states that the g_i's average to one. This is a consequence of the absence of income effects: the government is indifferent between \$1 of public funds and \$1 uniformly distributed to all. The two equations can be combined to obtain the standard optimal tax formula

$$\tau = \frac{1-\bar{g}}{1-\bar{g}+e} \quad \text{with} \quad \bar{g} = \frac{\int g_i z_i d\nu(i)}{Z} \tag{3}$$

where $\bar{g}$ is the average social marginal welfare weight weighted by pre-tax income z. $\bar{g}$ can also be seen as the average normalized income z_i/Z weighted by the social welfare weights g_i.

The optimal tax rate τ balances the equity-efficiency tradeoff. It decreases with the elasticity of income e, which measures the efficiency costs of taxation and increases with $(1-\bar{g})$ which measures the value of redistribution. When utility is linear, $g_i \equiv 1$, and hence $\bar{g} = 1$. Since there is no value for redistribution, $\tau = 0$. When utility is *any* concave function, g_i decreases with z_i and hence $\bar{g} < 1$ and $\tau > 0$. Note also that $\bar{g}$ increases with τ as increasing taxation combined with an increased lump-sum demogrant reduces the difference in utility across individuals with different earnings abilities. Therefore, in general, equation (3) defines an unique solution.

4.2 Generalized Social Welfare Weights

Instead of specifying the social welfare weights as derived from the social welfare function, i.e., $g_i = u_i'((1-\tau)z_i + R - h_i(z_i))/p$, in our new approach, we directly write them as a function of disposable income and taxes, $g_i = g(c_i, T(z_i))$ with $T(z_i) = \tau z_i - R$ as in the prior section. Formula (3) continues to apply with such weights. To see this, let us derive it again, using the tax reform (or "perturbation") approach to optimal taxation. Suppose the government considers changing τ by some small amount $d\tau$. The impact of the reform on the utility of individual i, measured in monetary terms is just the direct effect on consumption $-z_i d\tau + dR$, since, by the envelope theorem, the change in behavior dz_i has no first order impact on utility. Budget balance $R = \tau Z(1-\tau)$ hence requires that $dR = [Z - \tau dZ/d(1-\tau)]d\tau = Zd\tau[1 - e\tau/(1-\tau)]$. Therefore, using weight g_i to measure the net social benefit of the reform for individual i, at the optimum we must have that the reform has a null value:

$$\int_i g_i \cdot \left[-z_i + Z \cdot \left(1 - e\frac{\tau}{1-\tau}\right)\right] d\nu(i) = 0$$

This can be immediately rewritten as equation (3).

As we shall see in the next subsections, our approach can capture a number of additional effects beyond the equity-efficiency trade-off captured by the standard welfarist approach. Before moving on to those, let us first show how our approach can nest the most widely used alternatives to utilitarianism in a more natural way than social welfare maximization.

Libertarian case. Under libertarianism, any individual is fully entitled to his pre-tax income and society should not be responsible for those with lower earnings.[31] This can be captured again in our framework by assuming that $g_i = g(T(z_i))$ is increasing in $T(z_i)$. This will immediately deliver $T(z_i) \equiv 0$ (hence, $\tau = 0$) as the optimal policy. In the standard framework, the way to obtain a zero tax at the optimum is to either assume that utility is linear or to specify a *convex* transformation of $u(.)$ in the social welfare function which undoes the concavity of $u(.)$. This seems much more artificial than directly stating that society considers redistribution as unjust confiscation.

[31] This view could for example be justified in a world where individuals differ solely in their preferences for work but not in their ability to earn. In that case, there is no good normative reason to redistribute from goods 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).

Rawlsian case. The Rawlsian case is the polar opposite of the Libertarian case. Society cares most about the least fortunate and hence sets the tax rate to maximize her welfare. In terms of a social welfare function, this can be captured by a maxi-min criterion. In our framework, it can be done instead by assuming that all the social welfare weight is concentrated on the least advantaged. If the latter have zero earnings, independently of taxes, then a reform is desirable if and only if it increases the demogrant R, leading to the revenue maximizing tax rate $\tau = 1/(1+e)$ at the optimum.[32]

Political economy. Our framework can also naturally incorporate political economy considerations. The most widely used model for policy decisions among economists is the median-voter model. In our model, each individual has single peaked preferences about the tax rate τ. This is because indirect utility $\psi_i(\tau) = u\left((1-\tau)z_i(1-\tau) + \tau Z(1-\tau) - h_i(z_i(1-\tau))\right)$ is single peaked at τ_i^* which is the solution to $-z_i + Z - \tau dZ/d(1-\tau)$, i.e., $\tau_i^* = (1 - z_i/Z)/(1 - z_i/Z + e)$. Hence, the median voter is the voter with median income z_m and the political equilibrium has:

$$\tau_m = \frac{1 - z_m/Z}{1 - z_m/Z + e}.$$

Note that $\tau_m > 0$ when $z_m < Z$ 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. As with the Rawlsian case, the weights g is not strictly speaking a function but a distribution with mass one at the median.[33]

Role of behavioral responses. Suppose there are no behavioral responses so that $e = 0$. In that case, the optimum τ is such that $\bar{g} = 1$, i.e., social welfare weights g_i should not be correlated with pre-tax income z_i. Under a standard welfarist model with g_i decreasing with c_i, $\bar{g} = 1$ only when $\tau = 1$. In contrast, with generalized social welfare weights $\bar{g} = 1$ is possible even with $\tau < 1$. For example, with social weights $g_i = g(c_i, T_i)$ as in Section 2, $\bar{g} = 1$ when τ is set high enough so that the correlation between deservedness and pre-tax income disappears.[34]

[32] This can also be seen as follows. Under the Ralwsian criterion and assuming the most disadvantaged have no earnings, $\bar{g} = 0$ because $g_i = 0$ whenever $z_i > 0$, hence formula (3) also leads immediately to $\tau = 1/(1+e)$.

[33] One may object that this result can also be recovered using a standard Pareto weighted sum of utilities with all the weight placed on the median earner. However, the difficulty of this approach is that the median could be endogenous to tax policy τ and hence impossible to specify ex-ante, before solving the problem using our generalized social marginal welfare weights (since, recall, that the standard exogenous Pareto weights should not depend directly on the tax policy).

[34] Another way to see this is as follows. With $\tau = 1$, deservedness and pre-tax incomes would be positively correlated (when g increases with T) so that $\bar{g} > 1$, calling for a lower tax rate.

With behavioral responses $e > 0$, then $\bar{g} = 1 - e\frac{\tau}{1-\tau} < 1$, i.e., the optimum τ is set so that the correlation between deservedness and pre-tax income remains negative. Hence, as in the standard model, the presence of behavioral responses lowers the optimum tax rate τ everything else being the same.

Equality of opportunity. Consider again the model of exogenous desert from the previous section, as applied to family background, now adding behavioral responses. Individuals can either come from an advantaged (in the notation of the previous section, $m = 1$) or disadvantaged background ($m = 0$). A good background gives an unfair advantage to earnings, such as through better access to schools from early on, but conditional on background, the level of earnings is based purely on merit (e.g., taste for work). The natural way to set the marginal welfare weights in this situation is to assume uniform and positive weights (say, equal to one) among those from a disadvantaged background ($g_0(z - T, T) = 1$ for all z) and weights equal to zero among those from an advantaged background ($g_1(z - T, T) = 0$ for all z). Those weights capture the two normative principles already mentioned. First, that there is value in redistributing from those coming from an advantaged background to those coming from a disadvantaged background conditional on having the same rank in the distribution of earnings in each background group. Second, that there is no value in redistributing across individuals with different earnings conditional on background because those different earnings are purely due to merit. If tax policy cannot be conditioned directly on background, then the social marginal welfare weight at a given earnings level is given by the formula in (2) which simplifies to $g(z - T(z), T(z)) = f(0|z) = f_0(z) P_0 / \sum_{j=0}^{1} f_j(z) P_j$, the fraction of individuals with a disadvantaged family background at the given earnings level. Since people with a disadvantage background typically have lower earnings, their frequency is decreasing at higher pre-tax earnings and so are the weights. Those are the implicit social welfare weights used in the analysis of Roemer et al. (2003) (see Piketty and Saez, 2013 for an application to the case of the optimum top income tax rate in such a model).

Skill or preferences for work. The exogenous desert model can also capture different notions of fairness analyzed in the literature when people differ according to both ability and taste for work (see for example Fleurbaey and Maniquet 2006, 2007, 2011). Suppose that the desert groups are split according to intrinsic ability or skill, which leads to a higher wage rate. In

addition, within each skill group, people have heterogeneous disutilities from work, with 'lazier' people disliking work more. For simplicity, consider two skill levels (high and low skill) and two preference types (lazy and hard-working). If society is libertarian along the preference dimension (that is, conditional on a skill level it views differences in earnings due to different preferences for work as fair), but Rawlsian along the skill dimension (that is, at any given level of income, it values most a transfer towards the low skill agent), then we can again capture this in a natural way with binary weights which are constant and equal to 1 for all low-skilled individuals and equal to 0 for all high-skilled individuals. The marginal social welfare weight at income level z will then be equal to the fraction of low skill individuals at income level z, so that the marginal social welfare weight will again be decreasing in income, even though conditional on a skill level (that is, given equality of opportunity), society is not bothered by inequality in outcomes. As shown by Fleurbaey (1994), the principle of compensation and the principle of responsibility are generally mutually incompatible. This example shows that generalized social marginal welfare weights can be used very simply to model the trade-off between those two principles.

4.3 Transfers and Free Loaders

In practice, behavioral responses are inherently tied to social welfare weights since one of the biggest complaint against redistribution is that it benefits "free loaders", that is those who stop working precisely because of the generosity of the redistributive system. Yet, standard welfarism cannot capture such effects, since the social welfare weight on a given individual depends solely on her current situation, and not on whether her current situation arises from responses to taxes and transfers.

The simplest way to illustrate this is to consider a simple model in which individuals can either work and earn a uniform wage w, or not work and earn zero. Utility is $u(c_l - \theta l)$ where $l \in \{0, 1\}$ takes the value 1 if an individual works and 0 otherwise and consumption c_l is equal to c_0 if an individual is out of work and to $c_1 = w \cdot (1 - \tau) + c_0$ if she works, where τ is the linear earnings tax rate. Taxes fund the demogrant transfer c_0. The cost of work θ is distributed according to a cdf $P(\theta)$. Individual θ works if and only if $\theta \leq c_1 - c_0 = (1 - \tau) \cdot w$. Hence, the fraction of people working is $P(w(1 - \tau))$. As before, let e be the elasticity of aggregate earnings $Z(1 - \tau) = wP(w(1 - \tau))$ with respect to the retention rate $(1 - \tau)$.

Under the utilitarian objective, the government maximizes:

$$SWF = \int_{\theta>(1-\tau)w} u(c_0)dP(\theta) + \int_{\theta\leq(1-\tau)w} u(c_0+(1-\tau)w-\theta)dP(\theta) \quad \text{s.t.} \quad c_0 = \tau w P((1-\tau)w) \quad (p).$$

Routine computations show that the optimal tax formula takes the form:

$$\frac{\tau}{1-\tau} = \frac{(1-P)(\bar{g}_0 - \bar{g}_1)}{e},$$

where $\bar{g}_0 = u'(c_0)/p$ is the average social welfare weight on non-workers and $\bar{g}_1 = \int_{\theta\leq(1-\tau)w} u'(c_0+(1-\tau)w-\theta)dP(\theta)/(p\cdot P)$ the average social welfare weight on workers. As before, because of no income effects, the average social welfare weight across the population is one, so that $(1-P)\bar{g}_0 + P\bar{g}_1 = 1$.

In the utilitarian framework, the social welfare weight placed on the unemployed depends only on c_0 and is completely independent of whether they would have worked absent taxes and transfers. By contrast, the public policy debate focuses on whether the unemployed are deserving of support or not. Transfer beneficiaries are only 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), Ellwood and Bane (1996)). 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 support from the government. The origins of the US welfare system since 1935, starting with the Aid to Families with Dependent Children (AFDC) and continuing with the Temporary Assistance to Needy Families (TANF) federal assistance programs highlights exactly that logic. The goal was to help children of single parents or whose families had low or no income, rather than a more general population, among which there could be many free-loaders.

Naturally, the fraction of "free loaders" among the unemployed increases with the generosity of transfers and with the behavioral elasticity e. Under standard utilitarianism, free loaders and deserving poor are treated equally in the social welfare function. With generalized welfare weights, it is possible to treat those two groups differentially.

Formally, let us define the deserving poor as those with $\theta > w$, (those who would not work, even absent any transfer), and the free loaders as those with $w \geq \theta > w \cdot (1-\tau)$ (those who do not work because of the welfare program only). Denoting by $P_0 = P(w)$ the fraction working when $\tau = 0$, there are $P(w(1-\tau))$ workers, $1 - P_0$ deserving poor, and $P_0 - P(w(1-\tau))$ free loaders.

Let us assume that society sets social marginal welfare weights for the deserving poor, $\bar{g}_0$ and those who work, $\bar{g}_1$ as under utilitarianism, but sets weights to zero for free loaders. Weights still average to one so that $(1-P_0)\bar{g}_0 + P\bar{g}_1 = 1$. The optimum tax rate becomes

$$\frac{\tau}{1-\tau} = \frac{(1-P)\left[\frac{1-P_0}{1-P} \cdot \bar{g}_0 - \bar{g}_1\right]}{e}.$$

Two points are worth noting about this formula. First, since $(1-P_0)/(1-P) < 1$ taxes will naturally be lower relative to the utilitarian case. Effectively, in the new formula $\bar{g}_0$ is replaced by $\bar{g}_0 \frac{1-P_0}{1-P} = \frac{(1-P_0)\cdot g_0 + (P_0-P)\cdot 0}{1-P}$ which is the average social marginal weight including the deserving poor (with weight $\bar{g}_0$) and the free loaders (with weight 0). In the extreme case in which all unemployed are free-loaders, the optimal transfer (and hence the taxes financing it) is zero. This corresponds to the (admittedly extreme) view that all unemployment is created by an over-generous welfare system. As long as there are some deserving poor though, taxes and transfers will be positive. Second, when e is larger, $(1-P_0)/(1-P)$ is smaller and hence a higher elasticity reduces the optimal tax rate not only through the standard efficiency effect but also through the social welfare weight channel as it negatively affects society's view on how deserving the poor are.

Application: Transfers over the Business Cycle. Individuals are less likely to be responsible for their unemployment status in a recession than in an expansion. In an expansion when jobs are easy to find, long unemployment spells are more likely to be due to low search efforts than in a recession when jobs are difficult to find even with large search efforts. If society wants to redistribute toward the hard-searching unemployed–i.e., those who would not have found jobs even absent unemployment benefits–then it seems desirable to have time limited benefits during good times combined with expanded benefit durations in bad times. Our online survey presented in Section 6 shows indeed that support for the unemployed depends critically on whether they can or cannot find jobs.

4.4 Luck versus Deserved Income

Let us consider again the case with luck vs. deserved income. Suppose that total income of individual i, z_i, has two components: standard income w_i, earned at a disutility cost $h_i(w_i)$ and a stochastic 'luck' income, y_i received at no cost. Faced with a linear tax τ, agent i again generates earnings $w_i(1-\tau)$ and receives a random luck shock, y_i, independent of taxes, leading to total observed earnings $z_i(1-\tau) = w_i(1-\tau) + y_i$, which are decreasing in τ.

Let us assume as in Section 3.3 that society believes that individuals are entitled to deserved income but not to luck income with binary social welfare weights based on $c = z - T(z) \lessgtr w + Ey$ so that the average social marginal welfare weight $g(z)$ at pre-tax income level z is given by the fraction of individuals with income z such that $z - T(z) < w + Ey$. Given such welfare weights, the derivation of the optimal tax formula is the same as above, leading to:

$$\tau = \frac{1-\bar{g}}{1-\bar{g}+e} \quad \text{with} \quad \bar{g} = \frac{\int g(z_i) z_i}{Z \int g(z_i)} \tag{4}$$

Under a standard utilitarian criterion, $\bar{g}$ increases with τ so that the solution of (4) is unique.[35]

In contrast, in this luck vs. deserved income model, $\bar{g}$ is not necessarily increasing with τ so that multiple equilibria (i.e., both low tax rate and high tax rate equilibria) are possible as in the model of Alesina and Angeletos (2005).

To see this, consider the case where deserved income is highly elastic to τ. Assume that, when taxes are low (e.g., $\tau \simeq 0$), luck income is small relative to deserved income on average, i.e., $Ey \ll Ew$. In that case, with low τ, e is large (as elastic deserved income is the main component of total income). Furthermore, a low tax rate is desirable in that case, as income is mostly deserved.[36] Hence, a close to zero tax rate is a stable optimum.

Conversely, suppose the tax rate τ is very high (and close to one). Because w is highly elastic, deserved income is now small. Let us assume it is quantitatively small relative to luck income. In that case, the tax base z has low elasticity and hence $e \simeq 0$. This implies that the optimal tax rate should indeed be high. The welfare effect further reinforces this as most income is undeserved so that a close to one tax rate is also a stable optimum.

Thus, economies with social preferences favoring hard-earned income over luck income can end up in two possible situations. In the low tax equilibrium, people work hard, luck income

[35] This also requires that the elasticity e does not depend on τ (or at least does not vary too quickly with τ).

[36] More precisely, imposing a positive tax rate would make higher incomes (above average) more deserving than lower incomes (below average), leading to a $\bar{g}$ above 1.

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 equilibrium, 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 take away that unfair luck income (itself favored by the high taxes in the first place). This shows that our framework can encompass the important multiple equilibria outcomes of Alesina and Angeletos (2005) without departing as drastically from optimal income tax techniques as in the model of Alesina and Angeletos (2005).

This example also illustrates that our theory delivers only locally Pareto efficient equilibria (i.e., equilibria where no small reform can improve everybody's welfare). In the situation described, the low tax equilibrium typically Pareto dominates the high tax equilibrium. Hence, starting from the high tax equilibrium, a large tax reform moving the economy to the low tax equilibrium can be Pareto improving.

5 Link with 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.

5.1 Horizontal Equity Concerns

The standard utilitarian framework leads to the conclusion that if agents can be separated into different groups, based on attributes, so-called 'tags', which are correlated with income and exogenous to taxes, then an optimal tax system should have differentiated taxes for those groups. 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 only mildly elastic to taxes (such as, arguably, the number of children or marital status). Mankiw and Weinzierl (2010) explore a tax schedule differentiated by height and use this stark example as a critique of the standard utilitarian framework. 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 framework for optimal tax policy which wishes to respect society's

preferences needs to be able to include them.[37]

To see how our approach allows us to think about horizontal equity, consider two groups which differ according to some observable and perfectly inelastic attribute $m \in \{1, 2\}$ and according to their taxable income elasticities (respectively denoted e_1 and e_2).[38] Let λ_1 and λ_2 be the fraction each group represents in the total population.

As a benchmark, the standard utilitarian approach would lead to two different tax rates, τ_1 and τ_2 for the two groups, such that:

$$\tau_m = \frac{1 - \bar{g}_m}{1 - \bar{g}_m + e_m} \quad \text{with} \quad \bar{g}_m = \frac{\lambda_m \int_{i \in m} g_i z_i d\nu_m(i)}{Z_m}$$

where $\bar{g}_m$ is the income weighted average social marginal welfare weight for group m and ν_m is the CDF of types, conditional on being in group m. A group would tend to be taxed more if it is less elastic, in the spirit of the traditional Ramsey inverse elasticity rule.

In our framework, we can however incorporate society's belief that different taxes on people with the same earnings are not fair. To do so, the social marginal welfare weight on each group can be specified as a function of both tax rates τ_1 and τ_2. Suppose we start from a situation with equal tax rates $\tau_1 = \tau_2 = \tau$, and consider a reform introducing differential taxes (and hence, horizontal inequity). Using standard weights $g_1(c_1, \tau_1) = g_2(c_2, \tau_2)$ at $c_1 = c_2$ and $\tau_1 = \tau_2$, if $e_1 < e_2$, it would be desirable in the standard utilitarian framework to perform a small reform

[37] A few comments, essentially based on Kaplow (2001), seem important. Kaplow (2001) forcefully argues that Horizontal Equity (HE) per se does not have an independent normative appeal, but that it often only proxies for losses in welfare due to unequal treatment. We are refraining from any judgment here on how important HE truly is as a normative criterion, or on where the social concern for it stems from. We simply take as given, driven by casual empirical observations, that society values it, whether it is reasonable or not, and we show how our framework can capture it. Secondly, Kaplow (2001), and Kaplow and Shavell (2001) highlight that Horizontal Equity considerations, in particular as modeled in Auerbach and Hassett (2000) will conflict with the Pareto principle in some cases. Our non-negative social welfare weights guarantee that this can never occur in our setup, so that the pursuit of horizontal equity will not come at the expense of welfare. Of course, the concept of horizontal equity per se remains subject to the valid criticisms raised in Kaplow (2001). To name a few, the benchmark against which society judges horizontal inequity (most often, pre-tax income) may itself be endogenous to tax policy or unfairly achieved (through other luck shocks), as well as arbitrary (which order one assigns to luck shocks matters for what is considered to be the fair 'status quo' distribution. Ideally, one would like to have some fundamental, morally justifiable definition of what constitutes "equals", rather than just based on pre-tax income. Our approach is somewhat more flexible than previous ones in that our groups could be based on any characteristic. A somewhat appealing interpretation for horizontal equity in our view is the fear of *unfair* discrimination by the government of otherwise "equals" such as based on religion, gender, interest groups lobbying, etc..

[38] Note that there are two possible ways to think about tags. The first and more standard one, in line with the aforementioned papers, considers two groups which differ in terms of their average earning abilities, in a world in which the government is unable to observe individual abilities. With nonlinear taxation, self-selection constraints are then relaxed thanks to tagging. The second approach, pursued here and more relevant for linear taxation, is to consider groups which differ in terms of their *elasticities* to taxes.

$(d\tau_1, d\tau_2)$, increasing taxes on group 1 and reducing them on group 2 (i.e., $d\tau_1 > 0 > d\tau_2$). But with generalized welfare weights and a preference for horizontal equity, $\tau_1 = \tau_2$ can be an equilibrium, despite different elasticities. Consider for example "differential weights" for reforms introducing horizontal inequity, with which consumption at the margin for a group gaining from the horizontal inequity would be less valued than for a group losing from it. A simple way is to assume that $g(c_i, \tau, d\tau_2 > d\tau_1) = 0$ for all $i \in 1$ and $g(c_i, \tau, d\tau_2 > d\tau_1) = g(c_i, \tau) / \int_{j \in 2} g(c_j, \tau)$ for all $i \in 2$ (and vice versa for a reform such that $d\tau_2 < d\tau_1$).

To see what taxes τ can be equilibria, consider that the value of any reform $(d\tau_1, d\tau_2)$ around an equilibrium with equal treatment $\tau_1 = \tau_2 = \tau$ must be non-positive, that is:

$$Z_1 \left(-\bar{g}_1 + [1 - e_1 \tau/(1-\tau)]\right) d\tau_1 + Z_2 \left(-\bar{g}_2 + [1 - e_2 \tau/(1-\tau)]\right) d\tau_2 \leq 0$$

where:

$$\bar{g}_m = \frac{\lambda_m \int_{i \in m} g(c_i, \tau_i, d\tau_2, d\tau_1) z_i d\nu_m(i)}{Z_m}$$

Since this must hold for any reform, it is sufficient to check that it holds for four 'basic' reforms, namely for $d\tau_1 > 0 = d\tau_2$, for $d\tau_2 > 0 = d\tau_1$, for $d\tau_2 < 0 = d\tau_1$ and for $d\tau_1 < 0 = d\tau_2$. Checking those four cases, leads to the following range of possible equilibrium taxes:

$$\min\left\{\frac{1}{1+e_1}, \frac{1}{1+e_2}\right\} \geq \tau \geq \max\left\{\frac{1 - \bar{g}_1(c, \tau, d\tau_1 > d\tau_2)}{1 - \bar{g}_1(c, \tau, d\tau_1 > d\tau_2) + e_1}, \frac{1 - \bar{g}_2(c, \tau, d\tau_2 > d\tau_1)}{1 - \bar{g}_2(c, \tau, d\tau_2 > d\tau_1) + e_2}\right\}$$

Hence, if the weight on the group threatened to suffer from a potential horizontal inequity is sufficiently large, there is a large possible interval of equilibrium taxes, despite the differential elasticities of both groups. This is because any deviation from equal taxes is penalized sufficiently heavily. In the limit, suppose that $\bar{g}_1(c, \tau, d\tau_1 > d\tau_2)$ tends to 1. Then any non-negative tax below the smallest revenue maximizing level across the two groups can be sustained as an equilibrium. Note however, that due to the fact that our social welfare weights are always non-negative, we cannot have a Pareto-dominated situation (as illustrated by the upper bounds equal to the revenue maximizing rates for each group).

We can also address the more general question of what *different* tax rates on the two groups can be sustained in equilibrium. To do so, we need to define the weights more generally as functions of both the tax levels and the tax changes. Let the weight for person i be $g_i = g(c_i, \tau_m, \tau_n, d\tau_m, d\tau_n)$. Consider the following example. Start from a set of standard welfare

weights $\{g(c_i, \tau_m)\}_i$ for each agent which sum to 1 across the population: $\int_{i \in m} g(c_i, \tau) d\nu(i) + \int_{i \in n} g(c_i, \tau) d\nu(i) = 1$. Now define the weights capturing horizontal equity concerns as follows:

i) If in the status quo $\tau_m > \tau_n$, a reform is introduced with $d\tau_m < d\tau_n$ (so that the group which is already taxed more is helped by a reduction in its taxes), then let $g(c_i, \tau_m > \tau_n, d\tau_m < d\tau_n) = g(c_i, \tau_m) / \int_{i \in m} g(c_i, \tau_m)$ for $i \in m$ and $g(c_i, \tau_m > \tau_n, d\tau_m > d\tau_n) = 0$ for $i \in n$.

ii) If in the status quo $\tau_m > \tau_n$, a reform is introduced with $d\tau_m > d\tau_n$ (so that the group which is already taxed more is hurt further by a tax increase), then let: $g_i = g(c, \tau_m > \tau_n, d\tau_m > d\tau_n) = g_i^+ = g^+(c, \tau_m > \tau_n, d\tau_m < d\tau_n)$ for all $i \in m$, where g^+ is a mean-preserving spread of g (informally put, the set of weights g^+ place even less value on high consumption people, and even more weight on low consumption people). This means that the average weight on group m, weighted by income will be higher for the proposed reform ii) than for reform i). Again, let $g_i = 0$ for all $i \in n$.

With this set of weights, a situation with $\tau_1 > \tau_2$ can be an equilibrium, if and only if:

$$Z_1 \left(-\bar{g}_1 + [1 - e_1 \tau_1/(1 - \tau_1)]\right) d\tau_1 + Z_2 \left(-\bar{g}_2 + [1 - e_2 \tau_2/(1 - \tau_2)]\right) d\tau_2 \leq 0$$

It is again sufficient to check that this equality holds for the following 'basic' tax reforms (since all possible other reforms can be expressed as combinations of these basic reforms) as explained in the Appendix. The equilibrium range of taxes obtained for a case with $\tau_1 < \tau_2$ is then:

$$\frac{1 - \bar{g}_1^+}{1 - \bar{g}_1^+ + e_1} \leq \tau_1 \leq \frac{1 - \bar{g}_1}{1 - \bar{g}_1 + e_1} \quad \text{and} \quad \tau_2 = \frac{1}{1 + e_2}$$

The last equality implies that if the government wants to set τ_2 at a lower level than τ_1, then it must necessarily be set at the revenue maximizing rate.

For a tax τ_1 to exist in this range, we need to have a valid range, so that we require:

$$\frac{1}{1 + e_2} < \frac{1 - \bar{g}_1}{1 - g_1 + e_1}$$

which reduces to $e_2 (1 - \bar{g}_1) > e_1$. Despite this condition being itself endogenous to τ_1 (since $\bar{g}_1$ is), it means broadly speaking that e_2 will have to be sufficiently larger than e_1 to justify τ_2 being set lower (and the more so, the more people are averse to horizontal inequity as captured in a larger $\bar{g}_1$).

Intuitively, our generalized social marginal welfare weights punish the group which benefits from tagging (i.e., which is taxed less as a result of tagging), so that the latter will only gain

if the its elasticity is really sufficiently large relative to the other group's. 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 that case, the revenue-maximizing rate is imposed on the group 'favored' by the tax system, in the same way that the revenue-maximizing rate is imposed on everyone except the poorest with Rawlsian social preferences. In this framework, a tradeoff appears between efficiency (setting taxes based on the differential elasticities) and social preferences for horizontal equity (pushing for taxes to be equalized despite different elasticities). In other words, we can rephrase the Rawlsian famous criterion as follows:

"It is permissible to discriminate against a group using taxes and transfers only in the case where such discrimination allows to improve the welfare of the group discriminated against."

One possible application of this analysis would be reforms focusing on gender-differentiated taxation. Indeed, there is ample empirical evidence that single mothers or secondary earners are more elastic in their labor supply than prime age men (see e.g., Blundell and MaCurdy, 1999). Yet, almost no country has adopted gender-differentiated taxes, despite the standard Ramsey consideration. Our welfare weights can make sense of the absence of such taxes.

5.2 Poverty Alleviation

The poverty rate, defined as the fraction of households below a given disposable income threshold (the poverty threshold) gets 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 used government objectives incorporating such government objectives. Besley and Coate (1992) and Kanbur, Keen, and Tuomala (1994) show how adopting poverty minimization indexes affects optimal tax analysis. Importantly, they show that the outcomes can be Pareto dominated. In this section, we show how the use of generalized welfare weights allows to incorporate in a simple way in the traditional optimal tax analysis poverty alleviation considerations while maintaining the Pareto principle.

We consider the optimal nonlinear income tax problem as we are particularly interested in the profile of taxes and transfers.

We assume that individuals differ solely through their ability parameter n, distributed with

cdf $F(n)$ and density $f(n)$. Individual n has utility

$$u^n(c,z) = c - nh(z/n)$$

with $h(.)$ denoting the increasing and convex cost of work function. We normalize h so that $h'(1) = 1$ and $h(0) = 0$. Each individual chooses z to maximize $z - T(z) - nh(z/n)$ leading to first order condition $1 - T'(z) = h'(z/n)$. Hence, we have $z = n\phi(1 - T')$ where $\phi(.)$ is the inverse of $h'(.)$. When $T' = 0$, $z = n$ so that we can interpret n as "potential income." Positive marginal tax rates depress real income relative to potential income. We denote by e the elasticity of earnings z with respect to the net-of-tax rate $1 - T'$.

The government maximizes

$$SWF = \int \omega_n G(z - T(z) - nh(z/n))dF(n)$$

Denoting by $g_n = \omega_n G'(u_n)/p$ the social marginal welfare weight on individual n, no income effects implies that $\int g_n dF(n) = 1$. The standard optimal tax formula takes the following form (see the derivation in appendix):

$$\frac{T'(z_n)}{1 - T'(z_n)} = \frac{1}{e} \cdot \frac{\int_n^\infty (1 - g_m)dF(m)}{nf(n)}. \tag{5}$$

The demogrant is then defined by the government budget constraint.

Let us now consider criteria of poverty alleviation. Let us denote the poverty threshold by $\bar{c}$. Anybody with disposable income $c < \bar{c}$ is poor. 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. Hence, the marginal tax rates would take the form

$$\frac{T'(z_n)}{1 - T'(z_n)} = \frac{1 - g}{e} \cdot \frac{1 - F(n)}{nf(n)} \quad \text{or} \quad T'(z_n) = \frac{1 - g}{1 - g + a_n \cdot e},$$

where $a_n = nf(n)/[1 - F(n)]$ is the local Pareto parameter. 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 $g = 0$ (which corresponds to the Rawlsian case), tax revenue cannot fund a demogrant as large as $\bar{c}$. Let us assume that n^* is the ability level at the poverty threshold so that $F(n^*)$ individuals are poor at the optimum. There are two ways to consider poverty alleviation, one is to minimize the poverty gap, the other to minimize the poverty rate.

Poverty gap alleviation. Suppose the government cares about the consumption of people living in poverty. A natural way to capture this is to assume that social welfare weights are concentrated among those living in poverty, i.e., with disposable income c below the poverty threshold $\bar{c}$. We can therefore specify the welfare weights as follows: $g(c) = \bar{g} > 0$ if $c < \bar{c}$ and $g(c) = 0$ if $c \geq \bar{c}$.[39] The normalization of social welfare weights implies that $\bar{g}F(n^*) = 1$. The optimum marginal tax rates take the following form:

$$\frac{T'(z_n)}{1-T'(z_n)} = \frac{1}{e} \cdot \frac{1-F(n)}{nf(n)} \quad \text{if} \quad n > n^*$$

$$\frac{T'(z_n)}{1-T'(z_n)} = \frac{\bar{g}-1}{e} \cdot \frac{F(n)}{nf(n)} \quad \text{if} \quad n < n^*$$

Because $\bar{g}F(n^*) = 1$, the marginal tax rate is continuous at the poverty level. The marginal tax rate is Rawlsian above n^* and positive (and typically large) below n^*. The shape of optimal tax rates is quite similar to the standard utilitarian case.

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. Hence, the social marginal welfare weights are concentrated solely at the poverty threshold $\bar{c}$. Hence $g(c) = 0$ below $\bar{c}$ and above $\bar{c}$, and $g(c) = \bar{g}$ at $\bar{c}$ ($\bar{g}$ is finite if a positive fraction bunch at the poverty threshold as we shall see, otherwise $g(c)$ would be a Dirac distribution). The optimum marginal tax rates take the following form:

$$\frac{T'(z_n)}{1-T'(z_n)} = \frac{1}{e} \cdot \frac{1-F(n)}{nf(n)} \quad \text{if} \quad n > n^{**}$$

$$\frac{T'(z_n)}{1-T'(z_n)} = \frac{1}{e} \cdot \frac{-F(n)}{nf(n)} \quad \text{if} \quad n < n^*$$

Hence, there is a kink in the optimal tax schedule with bunching at the poverty threshold $\bar{c}$. All individuals with n such that $n^* \leq n \leq n^{**}$ bunch at the poverty level so that z_n is constant and $c_n = \bar{c}$ in that range. The marginal tax rate is Rawlsian above the poverty threshold and is negative below the poverty threshold so as to push as many people 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 rate.

[39] A less extreme version of this assumption would set $g(c) = \underline{g}$ above $\bar{c}$ with $\underline{g} < \bar{g}$. It is easy to adapt our results to that case.

5.3 Link with Fleurbaey-Maniquet: Work Preferences vs. Skills

Fleurbaey and Maniquet (2006, 2007, 2011) have considered optimal income tax models where individuals differ in skills and in preferences for work (Fleurbaey and Maniquet, 2011, chapters 10 and 11 present their framework in detail). Based on the "Compensation objective" and the "Responsibility objective", they develop 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). In this section, we outline how two of the criteria developed by Fleurbaey and Maniquet translate into profiles of social marginal welfare weights. This allows us to connect their theory to our approach with generalized social welfare weights.

For simplicity, we consider the case of the nonlinear income tax on earnings as in the continuous Mirrlees model.[40] We assume away income effects using quasi-linear utilities of the form: $u^i = c - h^i(z/w_i)$ where w_i is the skill of individual i so that $l = z/w_i$ is labor supply required to earn income level z. Skills are distributed in $[w_{\min}, w_{\max}]$ with $w_{\min} > 0$ and labor supply $l \in [0,1]$ so that $l = 1$ represents full-time work. Heterogeneity in work preferences are embodied in the individual specific disutility of work function $h^i(.)$. The key contribution of Saez (2001) is to derive an optimal income tax formula that generalizes the formulas of Mirrlees (1971) to situations with heterogeneous populations (i.e., situations where individuals differ not only in skills but also possibly in preferences). The optimal marginal tax rate for earnings level z can be expressed as follows:

$$\frac{T'(z)}{1-T'(z)} = \frac{1}{e} \cdot \frac{\int_z^\infty (1-g(z'))dH(z')}{zh(z)}, \tag{6}$$

where e is the average elasticity of earnings with respect to the net-of-tax rate $1-T'$ at earnings level z, $H(z)$ is the cumulative earnings distribution function, $h(z)$ the earnings density,[41] and $g(z')$ the average social marginal welfare weight at earnings level z'. Because of no income effects, social marginal welfare weights average to one in the population. Formula (6) is a generalization of formula (5) in the case with heterogeneous populations.

Fleurbaey and Maniquet consider various social criteria. We focus on two of them that produce explicit optimal tax formulas, the w-equivalent leximin criterion and the $w_{\min}$-equivalent

[40] Fleurbaey and Maniquet consider also the case with discrete populations as well as the case where the government can also observe hours of work (but individuals can choose to work for a wage lower than their skill).

[41] More precisely, it is the virtual density that would hold at z if the income tax system were linearized at z.

leximin criterion.[42]

The w-equivalent leximin criterion. The w-equivalent criterion satisfies the Equal Preferences Transfer Principle and a weakened version of the Equal Skills Transfer Principle.[43] To rank different allocations, it starts by defining an equivalent skill level for every allocation and agent, which is the skill level for which an agent would be indifferent between his current allocation and the best allocation he could achieve if he freely chose labor supply at that skill level. Agents whose equivalent skill is lower are naturally disadvantaged and are considered to be the worst-off ones. The w-equivalent criterion ranks allocations according to which one provides a higher equivalent skill level to the worst-off agent. Intuitively, redefining an equivalent skill level appropriately neutralizes differences in preferences by allowing agents to freely choose on a budget set and favors those with an unfairly low intrinsic skill level. Differences in preferences are not compensated for at all under this criterion: the hard-working ones among the low-skilled will be most rewarded.

Under their w-equivalent leximin social criteria, the optimal tax system maximizes the net transfers to those with the minimum skill $w_{\min}$ working full-time, i.e., $l = 1$ and hence earning $z = w_{\min}$. The optimal marginal tax rate is negative in the earnings range $[0, w_{\min}]$ (Theorem 11.5 in Fleurbaey and Maniquet, 2011), and positive for incomes above $w_{\min}$. Mapping this criterion into our social marginal welfare weights requires that the weights are fully concentrated on those with skill $w_{\min}$ who work full time $l = 1$ and hence earn $w_{\min}$. Social welfare weights are zero on any other earnings level, so that $g(z)$ is a Dirac distribution concentrated at $w_{\min}$. Applying formula (6), this implies indeed that $T'(z)/[1-T'(z)] = (1/e)[-H(z)/(zh(z))] < 0$ for $z < w_{\min}$ and $T'(z)/[1 - T'(z)] = (1/e)[(1 - H(z))/(zh(z))] > 0$ for $z > w_{\min}$.[44] This criterion concentrates social welfare weights on the hard working, low-skilled workers, which justifies the

[42]The authors prove that it is impossible to simultaneously satisfy the Equal Preferences Transfer Principle and the Equal Skills Transfer Principle in their pure form. The Equal Preferences Transfer Principle is weakened to the "Equal Welfare Selection Principle" which states that if all agents had the same preferences, the only efficient allocations should be the ones which are symmetric for everyone. The Equal Skills transfer principle is weakened to the "Laissez-Faire Principle" which states that if agents face the same budget set, they should be left to optimize without intervention. However, there is an asymmetry. While the combination of the Equal Preferences transfer and the Laissez Faire principles leads to the w-equivalent criterion, the combination of the weakened Equal Preferences and the Equal Skills transfer Principle does not satisfy Pareto efficiency and separability at the same time. Hence, the authors instead combine a yet modified Equal Skills Transfer principle (described below) with the Equal Preferences Transfer to obtain the $w_{\min}$-equivalent criterion.

[43]More precisely, it satisfies the Laissez-Faire Principle described above.

[44]Note that this optimal tax system is isomorphic to the tax schedule minimizing the poverty rate discussed above although the ethical justification does not arise from the same justice principles.

use of an in-work benefit such as the Earned Income Tax Credit, i.e., larger transfers for low income workers than for those not working.[45]

The $w_{\min}$-equivalent leximin criterion. The $w_{\min}$-equivalent criterion also satisfies the Equal Preferences Transfer Principle and a weakened version of the Equal Skills Transfer Principle.[46] For each allocation and each agent, it defines an equivalent lump-sum transfer which is the transfer that would make an agent indifferent between his current allocation and the allocation he would receive were he allowed to choose labor supply freely at the minimum wage level, $w_{\min}$ and received that lump-sum transfer. This equivalent transfer tries to partially capture the difference in utility attributable to preferences, since it is computed when agents are left to all work at the same $w_{\min}$ level. Agents with low equivalent lump-sum transfers are the hard-working ones and favored by the social criterion. However, it allows for some compensation as well, in the sense that the hard-working ones will be less favored than if the pure Equal Skills Transfer Principle were applied. Intuitively, it will favor low skill people, even if they do not work very hard, and will hence partially redistribute across preferences as well. According to this criterion, an allocation is preferred to another if the smallest equivalent lump-sum transfer across all agents is higher.

Fleurbaey and Maniquet show that 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, consistent with the intuition that this criterion focuses on low-skilled agents but does not reward hard-working ones among them as much as the previous criterion. The optimal tax system maximizes this transfer and has positive marginal tax rate above $w_{\min}$ (Theorem 11.4 in Fleurbaey and Maniquet, 2011). Using (6), this optimal tax system implies that $\int_z^\infty [1 - g(z')]dH(z') = 0$ for $0 \leq z \leq w_{\min}$. Differentiating with respect to z, we get $g(z) = 1$ 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.[47] Social marginal welfare weights are then zero above $w_{\min}$ so that (6) implies that $T'(z)/[1 - T'(z)] = (1/e)[(1 - H(z))/(zh(z))] > 0$

[45] Saez (2002) made a related point in the discrete version of the Mirrlees model. Namely, if the social marginal welfare weight on non-workers is below one, then an in-work benefit is optimal even in a model with only intensive labor supply responses.

[46] This modified version applies the Equal Skills Transfer principle only to pairs of agents such that the richer agent is also the more hard-working one (in the sense that at any given budget set, he would choose to work more). This in essence allows for some 'compensation' for being 'lazy' and, symmetrically, reduces the reward for being hard-working.

[47] Effectively, the social objective gives them average weight because all those earners work strictly less than $l = 1$, implying that part of the reason for their low earnings is low taste for work.

for $z > w_{\min}$.[48]

This criterion, and the average weights $g(z)$ implied by it, are founded on the following underlying generalized social marginal welfare weights. Weights for individual i are a function of the skill level and the tax paid (equivalently, of the transfer received), i.e., $g_i = g(w_i, T(z_i))$ such that: i) $g_i = 0$ for $w_i > w_{\min}$, for any $T(z_i)$ (there is no social welfare weight placed on those with skill above $w_{\min}$ no matter how much they pay in taxes) and ii) $g(w_{\min}, \cdot)$ is an (endogenous) Dirac distribution concentrated on $T_{\max} = \max_{\{i:w_i=w_{\min}\}} T(z_i)$ (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. Since there are agents with skill level $w_{\min}$ found at every income level below $w_{\min}$ (an assumption made by Fleurbaey and Maniquet), the sole equilibrium is to have equal transfers, i.e., $T'(z) = 0$ in the $[0, w_{\min}]$ earnings range. Weights are zero above earnings $w_{\min}$ since those with the lowest skill $w_{\min}$ can at most earn $w_{\min}$, even when working full time.

The Fleurbaey and Maniquet approach can be seen as complementary to ours: they derive social preferences from reasonable axioms, which aim to capture society's views above and beyond utilitarianism. As highlighted in the two examples above, there are suitable welfare weights that can then be specified to capture those same social preferences. Their approach could be used in conjuction to ours. Indeed, we could apply that same axiomatic method *directly* to the welfare weights (and we hope that future work will do so), so as to draw conclusions about the properties of reasonable weights. Our approach has three main advantages i) we do not need to worry about the Pareto principle (which they need to explicitly account for, case by case, since they work with a social objective function) ii) we do not need to specify global axioms, only local ones, iii) we can easily adapt the optimal tax formulas derived in the large welfarist literature, using our generalized weights, instead of having to re-derive optimal tax formulas from scratch.

[48] As social marginal welfare weights $g(z)$ average to one in the full population, this means that there is an atom at $w_{\min}$.

6 Empirical Testing using Survey Data

The next step in this research agenda is to provide empirical foundation for our theory, in addition to the existing papers cited in Section 2.2. The basic tool we use is a series of online survey questions destined to elicit people's preferences for redistribution and their concepts of fairness. The questions 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, namely, such as whether marginal utility of income matters (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 3, 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 slight more than 1100 respondents.[49] The complete details of the survey are presented in appendix A.3. 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.

Table 2 reports preferences for giving a tax break and or a benefit increase across individuals in various scenarios. Panel A considers two individuals with the same earnings, same taxes, and same disposable income but who differ in the 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 individual taste based marginal utilities should not be relevant for tax policy as long as disposable income is the same. 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

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

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.

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 majority), 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 some some support to the Fleurbaey and Maniquet social criteria. Long hours of work do seem to make a person more deserving than short hours of work, conditional on having the same total earnings.

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 abled bodied unemployed or welfare recipient not looking for work. This provides very strong support to the "free loaders" theory that ability and willingness to work are the key determinants of deservedness of transfer recipients. Those results are consistent with a broad body of work discussed in Section 2.2.

Next, in the spirit of our analysis of Section 3 with fixed incomes, we analyze whether revealed social marginal welfare weights depend on disposable income and/or taxes paid. Table 3 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 3 considers two families A and B with similar disposable income but dissimilar pre-tax income (and hence taxed paid). Family B has lower taxes and pre-tax incomes than family A. We keep family B constant and vary family A taxes and disposable income. Overall, subjects overwhelmingly find family A more deserving than family B implying that disposable

income is not a sufficient statistics (as in the utilitarian case) to determine deservedness, and that taxes paid enter deservedness positively. Panel B in Table 3 considers two families A and B with similar taxes paid but dissimilar pre-tax income (and hence 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. Overall, subjects overwhelmingly find family B more deserving than family A implying that taxes paid is not a sufficient statistics (as in the libertarian case) to determine deservedness and that disposable income enters negatively deservedness. Therefore, Table 3 provides compelling non-parametric evidence that both taxes and disposable income matter for social marginal welfare weights as we posited in Section 3.

Finally, Table 4 estimates the weights placed by social preferences on both disposable income and taxes paid. Recall the simple linear form discussed above, $g(c,T) = g(c - \alpha T)$, 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.[50] 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. 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 $\frac{dT}{dc}|_{S=\text{constant}} = -\frac{\beta_T}{\beta_c}$. Table 4 shows the implied α and the optimal marginal tax rates in four subsamples.[51] The implied α is between 0.37 and 0.65, so that the implicit optimal marginal tax rates are relatively high, ranging from 61% to 74%. In

[50] Annual incomes could take one of 7 values \$10K, \$25K, \$50K, \$100K, \$200K, \$500K, \$1 million, and taxes could take one of 5 values, 5%, 10%, 20%, 30%, and 50%.

[51] 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).

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. This simple exercise confirms the results from Table 3 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 can help calibrate the weights more precisely.

7 Conclusion

This paper has proposed a generalized theory of optimal taxation using the tax reform approach and generalized social marginal welfare weights. A tax system is optimal if no budget neutral marginal reform can increase the sum of (money metric) gains and losses across individuals weighted using the generalized social marginal welfare weights. Optimum tax formulas take the same form as standard utilitarian tax formulas by simply substituting standard marginal social welfare weights by those *generalized marginal social welfare weights*. Hence our theory nests standard theory and is equally tractable. As we have shown through a series of examples, the use of suitable generalized social welfare weights can help resolve most of the puzzles of the traditional welfarist approach and account for existing tax policy debates and structures while retaining (local) Pareto constrained efficiency. In particular, generalized welfare weights can provide a rich theory of optimal taxation even absent any behavioral responses. 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 desert, 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 violations of horizontal equity) may have the highest priority. In that case, those deserving of support are those discriminated against whenever horizontal inequities arise. This implies that horizontal inequities can only arise if they help the group discriminated against, dramatically lowering the scope for such policies (such as tagging) that arise with the standard welfarist approach and that are not observed as frequently in the real world.

Second, deserving individuals will be those who face difficult economic situations through no fault of their own. This captures the principle of compensation. Health shocks come to mind, explaining why virtually all advanced countries adopt generous public health insurance that effectively compensate individuals for the bad luck of having poor health and 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 (Piketty and Saez, 2012)[52] can therefore be justified on such grounds. Naturally, public education or inheritance taxation cannot fully erase inequalities due to background. This leaves a role for taxes and transfers based on income that aim at correcting for remaining inequities in opportunity as in the theory of Roemer et al. (1993), which can also be nested in our choice of welfare weights.

Third, even conditional on background, there remains substantial inequality in incomes. Part of this inequality is due to choices (preferences for leisure vs. preferences for goods) 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. 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).

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

[52]Stantcheva (2012) considers optimal education and human capital policies, distinguishing between policies which improve equality of opportunity versus those who exacerbate already existing skill differences.

than for higher income individuals. This principle might be particularly strong at the low income end to justify the use of anti-poverty programs.

Social preferences of the public are indeed shaped by beliefs about what drives disparities in individual economic outcomes (effort, luck, background, etc.) as in the model of Piketty (1995). Generalized welfare weights could also be derived empirically, by estimating actual social preferences of the public, leading to a positive theory of taxation. There is indeed a small body of work trying to uncover perceptions of the public about various tax policies using surveys (see e.g., Fong, 2001 and Frohlich and Oppenheimer, 1992). More ambitiously, economists can also cast light on those mechanisms and hence enlighten public perceptions so as to move the debate up to higher level normative principles.

A Appendix

A.1 Horizontal Equity Proofs

The value of any differential reform on the two groups, $d\tau_1$ and $d\tau_2$ can be derived as usual: There is a direct consumption effect on group i equal to: $-z_i d\tau + dR$. By budget balance, we have that: $R = \tau_1 Z_1(1-\tau_1) + \tau_2 Z_2(1-\tau_2)$ (Z_i denotes the total income earned by group i), so that $dR = [Z_1 - \tau dZ_1/d(1-\tau_1)]d\tau_1 + [Z_2 - \tau dZ_2/d(1-\tau_2)]d\tau_2 = Z_1 d\tau_1[1 - e_1\tau_1/(1-\tau_1)] + Z_2 d\tau_2[1 - e_2\tau_2/(1-\tau_2)]$

Weights sum to 1 over the whole population, $\int_i g_i d(v_i) = 1$. and hence $\int_{i\in 1} g_i d\nu(i) = 1 - \int_{i\in 2} g_i d\nu(i)$. Also, letting $v_m(i)$ denoted the CDF of types conditional on being in group m, and by λ_m the fraction of types m in the population, we have: $\lambda_1 \int_{i\in 1} g_i d\nu(i) + \lambda_2 \int_{i\in 2} g_i d\nu(i) = 1$. A situation with $\tau_1 = \tau_2 = \tau$ is an equilibrium iff:

$$-\lambda_1 \int_{i\in 1} g_i z_i d\nu_1 d\tau_1 + \lambda_1 \int_{i\in 1} g_i Z_1 \left[1 - \frac{e_1\tau}{1-\tau}\right] d\nu_1 d\tau_1 + \lambda_1 \int_{i\in 1} g_i Z_2 \left[1 - \frac{e_2\tau}{1-\tau}\right] d\nu_1 d\tau_2 +$$
$$-\lambda_2 \int_{i\in 2} g_i z_i d\nu_2 d\tau_2 + \lambda_2 \int_{i\in 2} g_i Z_1 \left[1 - \frac{e_1\tau}{1-\tau}\right] d\nu_2 d\tau_1 + \lambda_2 \int_{i\in 2} g_i Z_2 \left[1 - \frac{e_2\tau}{1-\tau}\right] d\nu_2 d\tau_2 \leq 0$$

or, simplifying:

$$Z_1\left(-\bar{g}_1 + [1 - e_1\tau/(1-\tau)]\right) d\tau_1 + Z_2\left(-\bar{g}_2 + [1 - e_2\tau/(1-\tau)]\right) d\tau_2 \leq 0$$

for any $d\tau_1$ and $d\tau_2$, where $\bar{g}_m$ is as defined in the main text. It is sufficient to check that this holds for four 'basic' reforms, namely i) $d\tau_1 > 0 = d\tau_2$, ii) $d\tau_2 > 0 = d\tau_1$, iii) $d\tau_1 < 0 = d\tau_2$, iv) $d\tau_2 < 0 = d\tau_1$. All other reforms can be written as combinations of these four basic reforms.[53]Hence, for $d\tau_1 > 0 = d\tau_2$, $\bar{g}_1 = \bar{g}_1(c, \tau, d\tau_1 > d\tau_2)$ and $g_2 = 0$, the following condition must hold:

$$\tau \geq \frac{1 - \bar{g}_1}{(1 - \bar{g}_1 + e_1)}$$

And symmetrically, one should not be able to profitably increase τ_2, that is for $d\tau_2 > 0 = d\tau_1$ the following condition must hold:

$$\tau \geq \frac{1 - \bar{g}_2}{(1 - \bar{g}_2 + e_2)}$$

In addition, it should not be possible to reduce either τ_1 or τ_2 leading to:

$$\min\left\{\frac{1}{1+e_2}, \frac{1}{1+e_1}\right\} \geq \tau$$

Existence here is not a problem, since this is a well-defined interval for τ as long as both average welfare weights are less than 1, weakly.

[53]For example consider a more general reform $d\tau_2 > d\tau_1 > 0$. If the effect of only $d\tau_2$ is positive, then we are ruling that out by the basic reform $d\tau_2 > 0$. If the effect of $d\tau_2$ is itself negative, we can not 'compensate' it by increasing $d\tau_1$ so much that the net effect is positive. Because that would require the effect of $d\tau_1$ alone to be positive, which we rule out as well in one of the four basic reforms.

General Case. With the set of weights as defined in the main text, a situation with $\tau_1 > \tau_2$ can be an equilibrium, if and only if:

$$Z_1 \left(-\bar{g}_1 + [1 - e_1\tau_1/(1-\tau_1)]\right) d\tau_1 + Z_2 \left(-\bar{g}_2 + [1 - e_2\tau_2/(1-\tau_2)]\right) d\tau_2 \leq 0$$

It is again sufficient to check that this equality holds for the following 'basic' tax reforms (since all possible other reforms can be expressed as combinations of these basic reforms).

1. For $d\tau_1 > d\tau_2 = 0$:
$$\frac{1 - \bar{g}_1^+}{1 - \bar{g}_1^+ + e_1} \leq \tau_1$$
2. For $d\tau_1 < d\tau_2 = 0$:
$$\tau_1 \leq \frac{1 - \bar{g}_1}{1 - \bar{g}_1 + e_1}$$
3. And finally, both $d\tau_2 > d\tau_1 = 0$ and $d\tau_2 < d\tau_1 = 0$ taken together lead to:
$$\tau_2 = \frac{1}{1 + e_2}$$

A.2 Poverty Alleviation

The individual first order condition $h'(z/n) = 1 - T'$ implies that the elasticity of earnings with respect to $1 - T'$ is $e = \frac{1-T'}{z}\frac{dz}{d(1-T')} = h'(z/n)/[(z/n)h''(z/n)]$. We denote by c_n, z_n, and u_n the consumption, earnings, and utility of individual n. Using the envelope theorem, we have $du_n/dn = -h(z/n) + (z/n)h'(z/n)$.

The government maximizes a social welfare function,

$$W = \int \omega_n G(u_n) f(n) dn \quad \text{s.t.} \quad \int c_n f(n) dn \leq \int z_n f(n) dn \quad (p),$$

where ω_n is the Pareto weight on individual n. Following Mirrlees (1971), in the maximization program of the government, u_n is regarded as the state variable, z_n as the control variable, while $c_n = u_n + nh(z_n/n)$. Therefore, the Hamiltonian is

$$H = [\omega_n G(u_n) + p \cdot (z_n - u_n - nh(z_n/n))] f(n) + \psi(n) \cdot [-h(z_n/n) + (z_n/n)h'(z_n/n)],$$

where $\psi(n)$ is the multiplier of the state variable. The first order condition with respect to z is

$$p\left(1 - h'(z_n/n)\right) f(n) + \frac{\psi(n)}{n} \cdot ((z_n/n)h''(z_n/n)) = 0.$$

The first order condition with respect to u is

$$-\frac{d\psi(n)}{dn} = [\omega_n G'(u_n) - p] f(n).$$

Denoting by $g_n = \omega_n G'(u_n)/p$ the social marginal welfare weight on individual n and using the transversality condition $\psi(\infty) = 0$, we can integrate this equation to

$$\psi(n) = p \int_n^\infty [g_m - 1] f(m) dm.$$

Plugging this expression into the first order condition with respect to z, noting that $h'(z_n/n) = 1 - T'(z_n)$, and $(z_n/n)h''(z_n/n) = h'(z_n/n)/e$, we have:

$$pT'(z_n)f(n) = \frac{1}{n \cdot e}(1 - T'(z_n)) \int_n^\infty [1 - g_m] f(m) dm,$$

which can be rearranged as equation (5) in the text. Note that the transversality condition $\psi(0) = 0$ implies that $\int g_n f(n) dn = 1$, i.e., social welfare weights average to one.

In the poverty alleviation case, $g_n = 0$ when $c_n > \bar{c}$, which leads to the formula in the text.

In the poverty minimization case, all the weights are concentrated at $\bar{c}$. If individuals with $n \in (n^*, n^{**})$ bunch at $\bar{c}$ then $g_n = 0$ below n^* and above n^{**} and $g_n = 1/[F(n^{**}) - F(n^*)]$ for $n \in (n^*, n^{**})$ which leads to the formulas in the text.

A.3 Online Survey

Our survey was conducted in December 2012 on Amazon's Mechanical Turk service, using a sample of 1100 respondents,[54] 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 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)

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

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 yer, 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 yer, 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 yer, 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."

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.

Akerlof, George. 1978. "The Economics of Tagging as Applied to the Optimal Income Tax, Welfare Programs, and Manpower Planning," *American Economic Review* 68(1), 8-19.

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

Atkinson, Anthony B. 1970. "On the Measurement of Inequality," *Journal of Economic Theory* 2(3), 244-63

Almas, Ingvild, Alexander Cappelen, Erik Sorensen, and Bertil Tungodden. 2010. "Fairness and the Development of Inequality Acceptance," *Science* 328(5982), 1176-1178

Auerbach, Alan and Kevin Hassett. 2002. "A New Measure of Horizontal Equity," *American Economic Review*, 92(4), 1116-1125.

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.

Besley, Timothy and Stephen Coate. 1992b. "Understanding Welfare Stigma: Taxpayer Resentment and Statistical Discrimination," *Journal of Public Economics* 48(2), 165-183.

Blundell, Richard and Thomas MaCurdy. 1999. "Labour Supply: A Review of Alternative Approaches," In: Ashenfelter, O., Card, D. (Eds.), *Handbook of labour Economics*, Volume 3A, North-Holland, Amsterdam.

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

Bosmans, Kristof and Erik Schokkaert. 2004. "Social welfare, the veil of ignorance and purely individual risk: an empirical examination," *Research on Economic Inequality* 11, 85-114

Choné, Philippe and Guy Laroque. 2005. "Optimal incentives for labor force participation," *Journal of Public Economics* 89.(2), 395-425

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

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

Chone, Philippe and Guy Laroque. 2005. "Optimal Incentives for Labor Force Participation," *Journal of Public Economics* 89, 395–425.

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

Diamond, Peter and James Mirrlees. 1971. "Optimal Taxation and Public Production I: Production Efficiency and II: Tax Rules." *American Economic Review*, 61: 8-27 and 261-278.

Edgeworth, Francis Y. 1897. "The Pure Theory of Taxation," *Economic Journal* 7, 46-70, 226-238, and 550-571.

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

Ellwood, David and Mary J. Bane. 1996. *Welfare Realities: from Rhetoric to Reform*, Cambridge: Harvard University Press.

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. 2006. "Fair Income Tax," *Review of Economic Studies* 73, 55-83.

Fleurbaey, Marc and François Maniquet. 2007. "Help the Low Skilled or let the Hard-working Thrive? A Study of Fairness in Optimal Income Taxation," *Journal of Public Economic Theory* 9(3), 467-500.

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

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.

Guesnerie, Roger. 1995. *A Contribution to the Pure Theory of Taxation*, Cambridge University Press: Cambridge.

Jeene, Marjolein, Wim van Oorschot, and Wilfred Uunk. 2011. "Popular Criteria for the Welfare Deservingness of Disability Pensioners: The Influence of Structural and Cultural Factors," *Journal of Social Indicators Research*, 1-15.

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)," *Inequality and Tax Policy,* Hassett and Hubbard, eds., American Enterprise Institute, 75–97.

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

Kaplow, Louis, and Steven Shavell. 2001. "Any Non-welfarist Method of Policy Assessment Violates the Pareto Principle," *Journal of Political Economy* 109(2), 281–86.

Kuziemko, Ilyana, Michael Norton, Emmanuel Saez, and Stefanie Stantcheva. 2013. "The Effects of Information about Inequality and Taxes on Preferences for Redistribution: Evidence from a Randomized Survey Experiment", Working paper.

Larsen, Christian Albrekt. 2008. "The Institutional Logic of Welfare Attitudes: How Welfare Regimes Influence Public Support," *Comparative Political Studies* 41(2), 145-168.

Lockwood, Benjamin and Matthew Weinzierl. 2012. "De Gustibus non est Taxandum: Theory and Evidence on Preference Heterogeneity and Redistribution," NBER Working Paper No. 17784.

Mankiw, Gregory 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 A. 1971. "An Exploration in the Theory of Optimal Income Taxation." *Review of Economic Studies*, 38: 175-208.
Mirrlees, James A. 1974. "Notes on Welfare Economics, Information and Uncertainty" in M. Balch, D. McFadden and S. Wu (eds.) *Essays in Equilibrium Behavior under Uncertainty* (Amsterdam: North-Holland), 243–258
Piketty, Thomas. 1995 "Social Mobility and Redistributive Politics," *Quarterly Journal of Economics*, 110(3), 551-584.
Piketty, Thomas and Emmanuel Saez. 2012. "A Theory of Optimal Inheritance Taxation," CEPR Discussion Paper No. 9241.
Piketty, Thomas and Emmanuel Saez. 2013. "Optimal Labor Income Taxation," *Handbook of Public Economics,* Volume 5, (Amsterdam: North Holland).
Piketty, Thomas, Emmanuel Saez, and Stefanie Stantcheva. 2011. "Optimal Taxation of Top Labor Incomes: A Tale of Three Elasticities," NBER Working Paper No. 17616
Reutter, Linda, Miriam Stewart, Gerry Veenstra, Rhonda Love, Dennis Raphael, and Edward Makwarimba. 2009. "Who Do They Think We Are, Anyway? Perceptions of and Responses to Poverty Stigma," *Qualitative Health Research*, 19(3), 297-311.
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. 2002. "Optimal Income Transfer Programs: Intensive Versus Extensive Labour Supply Responses." *Quarterly Journal of Economics*, 117(2): 1039-73.
Stantcheva, Stefanie. 2012 "Optimal Human Capital Policies and Redistribution over the Lifecycle," MIT Working Paper
Stiglitz, Joseph. 1987. "Pareto Efficient and Optimal Taxation and the New New Welfare Economics," *Handbook of Public Economics,* Volume 2, (Amsterdam: North Holland).
Weinzierl, Matthew C. 2012. "Why Do We Redistribute So Much But Tag So Little? The Principle of Equal Sacrifice and Optimal Taxation," NBER Working Paper No. 18045.
Weinzierl, Matthew C. 2012b. "The Promise of Positive Optimal Taxation: A Generalized Theory Calibrated to Survey Evidence on Normative Preferences Explains Puzzling Features of Policy," NBER Working Paper No. 18599.
Werning, Ivan. 2007. "Pareto Efficient Taxation," MIT working paper
Will, Jeffrey. 1993. "The Dimensions of Poverty: Public Perceptions of the Deserving Poor," *Social Science Research* 22, 312–332.
Zoutman, Floris, Bas Jacobs, and Egbert Jongen. 2012. "Revealed Social Preferences of Dutch Political Parties", Tinbergen Institution Discussion Paper.

Table 1: Generalized Social Marginal Welfare Weights

	Actual practice	Standard Welfarist Criterion	Generalized Social Marginal Welfare Weights
	(1)	(2)	(3)
Pareto efficiency	Desirable	Yes	Yes (local) if g are not negative
Optimal taxes with fixed incomes	Non-degenerate	Degenerate (full redistribution desirable)	Non-degenerate if g depend directly on taxes paid (in addition to consumption)
Luck income vs. deserved income	Important	Cannot be distinguished	Can be distinguished if g depends on luck vs. deserved income
Free loaders	Important	Cannot be captured	Can be captured if g depends on hypothetical behavior (work or not absent transfers)
Tax increase/decrease asymmetry	Important	Cannot be captured	Can be captured if g depends on direction of small tax reform
Tagging	Used minimally	Highly desirable	Can be made undesirable if g depends on horizontal inequities (g also needs to depend on small tax reform)

Note: 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: Revealed Social Preferences

	(1)	(2)	(3)	(4)
A. Consumption lover vs. Frugal				
	Consumption lover > Frugal	Consumption lover = Frugal	Consumption lover < Frugal	
# obs. = 1,125	4.1%	74.4%	21.5%	
B. Hardworking vs. leisure lover				
	Hardworking > Leisure lover	Hardworking = Leisure lover	Hardworking < Leisure lover	
# obs. = 1,121	42.7%	54.4%	2.9%	
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
% assigned first rank	57.5%	37.3%	2.7%	2.5%
% assigned last rank	2.3%	2.9%	25.0%	70.8%

Notes: This table reports preferences for giving a tax break and or a benefit increase across 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% 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 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 or welfare recipient not looking for work.

Table 3: 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	48.5%	54.7%	65.4%
A=B	38.9%	37.3%	27.9%
A<B	12.6%	8.0%	6.7%
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	7.8%	3.5%	3.1%
A=B	29.4%	40.3%	23.8%
A<B	62.7%	56.2%	73.1%

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).

Table 4: Calibrating Social Welfare Weights

	(1)	(2)	(3)	(4)
	Probability of being deemed more deserving			
d(Tax)	0.0017***	0.0052***	0.0156***	0.0154***
	(0.0003)	(0.00194)	(0.00194)	(0.00216)
d(Net Income)	-0.00456***	-0.00912***	-0.024***	-0.024***
	(0.000124)	(0.000285)	(0.000784)	(0.000937)
Number of observations	11,450	8,368	5,816	3,702
Implied α	0.37	0.58	0.65	0.64
Implied Optimal MTR	73%	63%	61%	61%

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 mil} and taxes 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 mil. Column (3) excludes families with income of 500K or more. Column (4) further excludes in addition families with income below 10K. The implied α is obtained as (the negative of) the ratio of the coefficient on d(Tax) over the one on d(Net income). The optimal implied constant 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.