Generalized Social Marginal Welfare Weights Imply Inconsistent Comparisons of Tax Policies

Itai Sher
University of Massachusetts Amherst

First version: February 15, 2021
This version: June 7, 2023

Abstract
This paper concerns Saez and Stantcheva’s (2016) generalized social marginal welfare weights (GSMWW), which aggregate losses and gains due to tax policies, while incorporating non-utilitarian ethical considerations. The approach evaluates local tax changes without a global social objective. I show that local tax policy comparisons implicitly entail global comparisons. Moreover, whenever welfare weights do not have a utilitarian structure, these implied global comparisons are inconsistent. I argue that broader ethical values cannot in general be represented simply by modifying the weights placed on benefits to different people, and a more thoroughgoing modification of the utilitarian approach is required. (JEL D60, D63, D71, H21, H23, I3)

*email: isher@umass.edu. I am grateful for helpful comments and discussions with Matthew Adler, Eduardo Davila, Maya Eden, Wojciech Kopczuk, Maria Koumenta, Benjamin Lockwood, Juan Moreno-Cruz, Louis Perrault, Paolo Piacquadio, Peter Sher, and Matthew Weinzierl and to seminar audiences at UC Riverside, the Welfare Economics and Economic Policy virtual seminar, the University of Chicago Harris Public Policy’s Political Economy Workshop, and the PPE Society annual meeting.
1 Introduction

The traditional optimal tax literature, building on the classic work of Mirrlees (1971), has adopted a broadly utilitarian normative framework. As argued by several recent authors, including Weinzierl (2014, 2017) and Fleurbaey and Maniquet (2018), the omission of other ethical principles that people care about, such as libertarianism, equality of opportunity, and desert, is a serious problem for the classical approach. Saez and Stantcheva (2016) have proposed a general, relatively simple, way of addressing these concerns: They argue that one can modify the optimality conditions of the standard approach so that these can incorporate broader values while maintaining the structure of the standard optimal taxation theory. According to Saez and Stantcheva’s generalized social marginal welfare weights (GSMWW) approach, all one has to do is substitute for the standard utilitarian welfare weights – corresponding to the marginal utility of consumption – other welfare weights reflecting broader values. Such generalized welfare weights can effectively be used as a kind of “get out of jail free” card that allows one to ignore normative issues on the assumption that they can be incorporated simply by appropriate selection of welfare weights. In this paper, I show formally that this solution to the problem of incorporating broader values into optimal tax does not work because it leads to inconsistencies. It is not possible, in general, to capture broad ethical principles simply by means of welfare weights. Broadening the normative considerations that bear on taxation will require a more thoroughgoing revision of optimal tax theory.

I now discuss the specific contributions of this paper. The GSMWW approach only claims to make local comparisons among tax policies and accordingly to find local optima. Indeed, Saez and Stantcheva write, “In our approach ... there is no social welfare objective primitive that the government maximizes”. The first contribution of the paper is to show how to collect the local comparisons made by generalized welfare weights into implied global social comparisons. In particular, for any system of welfare weights $g$, I define strict and weak rankings $\prec_g$ and $\sim_g$ over tax policies, which capture (some of) the global social comparisons implied by welfare weights (see Section 3). Second, I define a critical property of welfare weights, structural utilitarianism (see Section 4), which is essential to the question of whether welfare weights are consistent. Third, I show that if welfare weights $g$ are structurally utilitarian, then welfare weights are consistent in the sense that there exists a social welfare function that generates those welfare weights (see Theorem 1 in Section 4). More specifically, I show that in the case in which welfare weights are structurally utilitarian, the case in which they can be generated by a generalized utilitarian social welfare function of the form $\int F_i(U_i) \, di$, where $U_i$ represents agent $i$’s utility and $F_i(U_i)$ is an agent-specific monotonic transformation of this utility. Fourth, I show that when welfare weights are not structurally utilitarian, then they are inconsistent in the sense of the following theorem:

---

1 I am grateful to an anonymous referee for suggesting the formulation of the problem in this paragraph as well as some of the wording.

2 For utilitarianism to be meaningful and for generalized utilitarianism to be meaningfully different than utilitarianism, we must assume that we are given utilities $U_i$ that are cardinal and interpersonally comparable.
Generalized Welfare Weights Inconsistency Theorem. If welfare weights weights $g$ are not structurally utilitarian, then they are inconsistent in the sense there exist tax policies $T_0, T_1, T_2, T_3$, each of which raises the same revenue, and such that welfare weights imply a social preference cycle of the form: $T_0 \sim^g T_1 \prec^g T_2 \sim^g T_3 \sim^g T_0$.

This is Theorem 3 in Section 6. Theorem 2 in Section 5 is a simpler version of the result with a more accessible proof. Putting together the third and fourth contributions, it follows that structural utilitarianism is necessary and sufficient for welfare weights to be consistent. So the generalized welfare weights approach does not meaningfully add anything beyond what is already available by means of a generalized utilitarian social welfare function, a framework which is long established; the additional possibilities offered by generalized welfare weights are inconsistent.

Some ethical values can be captured in a generalized utilitarian framework by making the transformations $F_i$ suitably dependent on agent characteristics. But Saez and Stanczewska suggest that libertarian values can be captured by making welfare weights a function of total taxes paid and that a poverty alleviation imperative can be captured by making weights a function of consumption, and I show that such weights lead to inconsistent judgements (see Sections 5.2 and 6.4). Section 7 continues the discussion of the significance of these results and their relation to the literature.

2 Model

This section presents the model of Saez and Stanczewska (2016). I assume all functions are smooth – meaning infinitely differentiable – unless their domain is discrete or explicitly stated otherwise.

2.1 Standard aspects of the model

There is a continuum of agents uniformly distributed on the interval $I = [0, 1]$. Each agent $i \in I$ has observable characteristics $x_i$ drawn from the set $X$ and unobservable characteristics $y_i$ drawn from the set $Y$. I assume that $X$ and $Y$ are either discrete or subsets of Euclidean spaces. Let $c_i$ be agent $i$’s consumption and $z_i$ be agent $i$’s income. Agent $i$ has the quasilinear utility function $U_i(c_i, z_i) = u(c_i - v_i(z_i))$, where $v_i(z_i) = v(z_i, x_i, y_i)$ is the cost of earning income $z_i$ given characteristics $(x_i, y_i)$. Assume $v$ is increasing and strictly convex in $z_i$, $v'_i(z_i) > 1$ for sufficiently large $z_i$, and $u$ is increasing and concave. I assume for simplicity for all $i$, $v'_i(0) < 1$, so that, in the absence of taxes, all agents would earn a positive income. A tax policy is a function $T: Z \times X \to \mathbb{R}$, where $T(z, x)$ is the tax paid by citizens with income $z$ given observable characteristics $x$, and $Z = \mathbb{R}_+$ is the set of possible income levels. I write $T_i(z_i) = T(z_i, x_i)$, so that $T_i$ gives $i$’s personalized tax on the basis of $i$’s observable characteristics. Let $\mathcal{T}$ be the set of all tax policies. I assume that $\mathcal{T}$ has the formal structure requisite to support the exposition that follows, and make more precise assumptions below\footnote{Indeed, Mirrlees (1971) posited a generalized utilitarian social welfare function, although with with a common transformation $F(U_i)$ of utility for all agents $i$.}. Given a tax policy $T$, we have $c_i = z_i - T_i(z_i)$.

\footnote{More specifically, in Section 5.1 I precisely define a subset $\hat{\mathcal{T}}$ of $\mathcal{T}$ with which my main results are concerned.}
Define \( z_i(T) \) to be \( i \)'s optimal income when facing tax system \( T \), and \( c_i(T) = z_i(T) - T_i(z_i(T)) \); formally, \( z_i(T) \in \arg \max_{z_i} U_i(z_i - T_i(z_i), z_i) \). The agent’s indirect utility from tax policy \( T \) is then \( U_i(T) = U_i(c_i(T), z_i(T)) \). Let \( R(T) = \int T_i(z_i(T)) \, \text{d}t \) be the revenue generated by \( T \).

### 2.2 Generalized welfare weights

The novelty in the GSMWW approach is the way that tax systems are evaluated. We assume a system \( g(c_i, z_i; x_i, y_i) \) of generalized social marginal welfare weights. Thus, we assign a certain weight to each agent depending on their consumption \( c_i \), income \( z_i \), and characteristics \( x_i, y_i \). Formally, a system of generalized social welfare weights is a function \( g : \mathbb{R} \times Z \times X \times Y \to \mathbb{R} \) such that \( g(c_i, z_i; x_i, y_i) > 0, \forall c_i, z_i, x_i, y_i \). Define \( g_i(c_i, z_i) = g(c_i, z_i; x_i, y_i) \). The intuitive interpretation of generalized social marginal welfare weights is that they measure the marginal social value of consumption for each person \( i \), and ratios of welfare weights \( g_i(c_i, z_i)/g_j(c_j, z_j) \) measure social marginal rates of substitution of consumption for agents \( i \) and \( j \). Given a tax system \( T \), the local marginal welfare weight \( g_i(T) = g_i(c_i(T), z_i(T)) \) is endogenously determined. The key innovation of the approach is to assess small tax reforms via local marginal welfare weights rather than by reference to a global objective.

I now present some illustrative examples from Saez and Stantcheva (2016). **Utilitarian weights**: \( g_i(c_i, z_i) = \frac{\partial}{\partial c_i} U_i(c_i, z_i) = u'(c_i - v_i(z_i)) \). These are the standard utilitarian weights that prioritize benefits according to the marginal utility for consumption. **Libertarian weights**: \( g_i(c_i, z_i) = \tilde{g}(z_i - c_i) = \tilde{g}(t_i) \), where \( t_i = z_i - c_i \) is the tax paid and we assume that \( \tilde{g}'(t_i) > 0 \). That is, the more tax a person has already paid, the greater the weight placed on that person. **Libertarian-utilitarian mix**: \( g_i(c_i, z_i) = \hat{g}(c_i - v_i(z_i), z_i - c_i) = \hat{g}(\hat{u}_i, t_i) \) where \( \hat{u}_i = c_i - v_i(z_i) \) with \( \frac{\partial \hat{u}_i}{\partial c_i} < 0 \) and \( \frac{\partial \hat{u}_i}{\partial t_i} > 0 \); the first inequality can be interpreted as saying that weights are increasing in marginal utility for consumption (since \( u'(c_i - v_i(z_i)) \) is decreasing in \( c_i - v_i(z_i) \)) and the second says that they are also increasing in taxes paid. **Poverty alleviation**: \( g(c_i, z_i) = 1 \) if \( c_i < \bar{c} \) where \( \bar{c} \) is the poverty threshold and \( g(c_i, z_i) = 0 \) otherwise; that is, we put positive and equal weight on those beneath the poverty line, and no weight on those below the poverty line.\(^5\) **Counterfactuals**: Welfare weights can be made to depend on how much someone would have worked in the absence of taxes (which depends on their type) in comparison to how much they work in the presence of taxes. **Equality of opportunity**: Weights can be made to depend on one’s rank in the income distribution conditional on one’s background conditions. Such weights go beyond the formal framework in that they depend on the entire income distribution and not just on \( c_i, z_i, x_i, \) and \( y_i \); Saez and Stantcheva present several examples that go beyond the basic formal framework they present.

### 2.3 Local optimality and local improvements

A tax reform is a function \( \Delta T \in \mathcal{F} \) whose interpretation is that it represents some change to the status quo tax policy. Define \( \Delta T_i(z_i) = \Delta T(z_i, x_i) \). Say tax reform \( \Delta T \) is locally budget neutral.
at \( T \) if \( \frac{d}{d\varepsilon}|_{\varepsilon=0} R(T + \varepsilon\Delta T) = 0 \). For any tax policy \( T \in \mathcal{T} \), tax reform \( \Delta T \in \mathcal{T} \), and system of social welfare weights \( g \) say that a locally budget neutral tax reform \( \Delta T \) is \textbf{locally desirable} if

\[
\int g_i(T) \Delta T(z_i(T)) \, di < 0. \tag{1}
\]

In other words, \( \Delta T \) is locally desirable at \( T \) if the cost of the tax change to different individuals, weighted by the local welfare weights, due to a small version of the reform \( \varepsilon\Delta T \), is negative. Say that tax system \( T \in \mathcal{T} \) satisfies the \textbf{local optimal tax criterion} if

\[
\forall \Delta T \in \mathcal{T}, \left[ \frac{d}{d\varepsilon}|_{\varepsilon=0} R(T + \varepsilon\Delta T) = 0 \Rightarrow \int g_i(T) \Delta T(z_i(T)) \, di = 0 \right]. \tag{2}
\]

Saez and Stantcheva (2016) say that (2) gives a necessary condition for local optimality of a tax system \( T \): Any locally budget neutral tax reform has no local aggregate effect on welfare when changes in tax liability are weighted by generalized social welfare weights evaluated at \( T \). This local optimality condition is a key tool for the GSMWW approach. It is used to derive optimal tax formulas that are analogous to the standard optimal tax formulas and thereby to analyze specific optimal tax problems incorporating broader non-utilitarian values.

In the traditional utilitarian framework, the goal is to choose a tax policy \( T \) to maximize the utilitarian objective \( \int U_i(T) \, di \) subject to a revenue constraint \( R(T) \geq E \), where \( E \) represents required government expenditures. Given this formulation, employing utilitarian weights \( g_i(c_i, z_i) = \frac{\partial}{\partial c_i} U_i(c_i, z_i) = u'(c_i - v_i(z_i)) \), and using the envelope theorem, the local optimal tax criterion (2) is a necessary condition for \( T \) to be an optimum, and (1) is a sufficient condition for a small version of the reform \( \Delta T \) to be a local improvement. However, in the GSMWW framework there is no global objective from which to derive these conditions; so, the conditions for a locally desirable reform and for a local optimum are posited by analogy to the utilitarian case.

### 3 The global social comparisons implied by welfare weights

Generalized social marginal welfare weights provide \textit{local} comparisons: conditions for a local improvement and for local optimality of tax policies. This section shows how to derive \textit{global} social preferences implicit in welfare weights.

#### 3.1 Modifying tax policies

To derive global comparisons, I need to smoothly vary tax policies in a parametric way. To do this, I append a (real-valued) parameter \( \theta \) to our tax policies \( T^{\theta} \). Varying \( \theta \) corresponds to changing tax policy in some way. For example, if \( T^{\theta} = T + \theta\Delta T \), where \( T \) is a tax policy and \( \Delta T \) is a tax reform, then \( \theta \) measures the size of the reform. Alternatively, consider a (non-individualized) linear tax \( T^{\theta}(z) = \theta z + \kappa(\theta) \), where, when we vary \( \theta \), we vary both the marginal tax rate and the lumpsum tax \( \kappa(\theta) \). In general, let \( \Theta = [\underline{\theta}, \overline{\theta}] \) be an interval in the real line, where \( \underline{\theta} < \overline{\theta} \).

Consider a parameterized collection \( (T^{\theta})_{\theta \in \Theta} \) of tax policies. Below I sometimes use the abbreviated
notation \((T^\theta)\) rather than \((T^\theta)_{\theta \in \Theta}\). Given \((T^\theta)\), define \(T_i(z, \theta) = T_i^\theta(z)\), so that \(T_i(z, \theta)\) can be regarded as a real-valued function with domain \(Z \times \Theta\). For any tax policy \(T\), agent \(i\), and income \(z_i\), let \(\hat{U}_i^T(z_i) = z_i - T_i(z_i) - v_i(z_i)\) be agent \(i\)’s utility if \(i\) chooses \(z_i\) in response to \(T\), using a representation of utility that omits the outer utility function \(u(\cdot)\). Let \(\hat{T}\) be a subset of well-behaved of tax policies: \(\hat{T}\) is the set of tax policies \(T\) such that (i) for each \(i \in I\), the optimal income \(z_i(T)\) exists and is unique, \(z_i(T) > 0\), and the second order condition holds with a strict inequality at the optimum \(\frac{d^2}{dz_i^2} \hat{U}_i^T(z_i(T)) < 0\); (ii) for each \(i \in I\), the map \(z \mapsto T_i(z)\) is smooth, and (iii) there are at most finitely many \(i' \in I\) at which, for some \(z\), the map \(i \mapsto T_i(z)\) is discontinuous.\(^6\)

Let \(T\) be the set of all parameterized collections of tax policies \((T^\theta)\) such that

1. \(T^\theta \in \hat{T}, \forall \theta \in \Theta\).

2. For all \(i \in I\), the map \((z, \theta) \mapsto T_i(z, \theta)\) is smooth.

Given a family \((T^\theta)\), write \(z_i(\theta) = z_i(T^\theta)\), \(c_i(\theta) = c_i(T^\theta)\), \(U_i(\theta) = U_i(T^\theta)\), \(g_i(\theta) = g_i(T^\theta)\), and \(\hat{U}_i^T(z_i(T^\theta))\) for, respectively, \(i\)’s optimal income, optimal consumption, indirect utility, and welfare weight at \(T^\theta\).

### 3.2 Global social comparisons implied by welfare weights

Consider a system of generalized social welfare weights \(\theta\). I will now define \(\prec^\theta\), the strict social preferences implied by \(\theta\), and \(\sim^\theta\), the social indifferences implied by \(\theta\); that is, for a pair of tax policies \(T_0\) and \(T_1\), \(T_0 \prec^\theta T_1\) means that welfare weights \(\theta\) imply that \(T_1\) is strictly socially preferred to \(T_0\), and \(T_0 \sim^\theta T_1\) means that welfare weights \(\theta\) imply that \(T_1\) is socially indifferent to \(T_0\).

As above, let \((T^\theta)_{\theta \in \Theta}\) be a parameterized collection of tax policies in \(T\), and let \(\theta_0, \theta_1 \in \Theta\) be such that \(\theta_0 < \theta_1\). Consider the following principles:

- **Global improvement principle.** Suppose that for all \(\hat{\theta} \in [\theta_0, \theta_1]\),

\[
\int g_i(\hat{\theta}) \left. \frac{\partial}{\partial \theta} \right|_{\theta = \hat{\theta}} T_i(z_i(\hat{\theta}), \theta) \, dz = 0
\]

\(\hat{\theta}\) that is, increasing \(\theta\) is locally socially desirable at \(\hat{\theta}\). Then \(T^\theta_0 \prec^\theta T^\theta_1\): \(T^\theta_1\) is socially preferred to \(T^\theta_0\).

- **Global indifference principle.** Suppose that for all \(\hat{\theta} \in [\theta_0, \theta_1]\),

\[
\int g_i(\hat{\theta}) \left. \frac{\partial}{\partial \theta} \right|_{\theta = \hat{\theta}} T_i(z_i(\hat{\theta}), \theta) \, dz = 0
\]

\(\hat{\theta}\) that is, welfare weights don’t detect any change in social welfare as \(\theta\) changes. Then \(T^\theta_0 \sim^\theta T^\theta_1\): \(T^\theta_0\) and \(T^\theta_1\) are socially indifferent.\(^7\)

\(^6\)As \(u(\cdot)\) is strictly increasing, \(u(c_i - v_i(z_i))\) and \(c_i - v_i(z_i)\) represent the same preferences over bundles \((c_i, z_i)\).

\(^7\)Observe that the set \(\hat{T}\) implicitly depends on the profile of cost of earning income functions \((v_i(\cdot))_{\iota \in I}\).
We can think of these two principles as axioms that put constraints on the relations \(<^g\) and \(\sim^g\), which are meant to capture, respectively, the strict social preferences and social indifferences implied by welfare weights \(g\). Henceforth, I shall assume that \(<^g\) and \(\sim^g\) satisfy these principles.

To understand these principles, consider first the standard utilitarian case. The utilitarian social welfare of tax policy \(T^g\) is \(W_{\text{util}}(\theta) = \int U_i(\theta) \, \text{d}i\). Because \(U_i(\theta) = U_i\left(\frac{z_i(\theta) - T_i(z_i(\theta), \theta), z_i(\theta)}{c_i(\theta)}\right)\), it follows from the envelope theorem that for any \(\hat{\theta} \in \Theta\),

\[
\frac{d}{d\theta} U_i\left(\hat{\theta}\right) = -\frac{\partial}{\partial c_i} U_i\left(c_i\left(\hat{\theta}\right), z_i\left(\hat{\theta}\right)\right) \frac{\partial}{\partial \theta} T_i\left(z_i\left(\hat{\theta}\right), \theta\right). 
\]

That is, the envelope theorem tells us that, the marginal effect of a change in tax policy on an agent’s utility is the product of the agent’s marginal utility of consumption and the marginal direct effect on the change in \(\theta\) on the agent’s tax bill, and we can ignore the indirect effects due to changes in behavior – the choices of consumption and income – as taxes change. So in the utilitarian case,

\[
\frac{d}{d\theta} W_{\text{util}}\left(\hat{\theta}\right) = -\int \frac{d}{d\theta} U_i\left(\hat{\theta}\right) \, \text{d}i = -\int \frac{\partial}{\partial c_i} U_i\left(c_i\left(\hat{\theta}\right), z_i\left(\hat{\theta}\right)\right) \frac{\partial}{\partial \theta} T_i\left(z_i\left(\hat{\theta}\right), \theta\right) \, \text{d}i \\
= -\int g_i\left(\hat{\theta}\right) \frac{\partial}{\partial \theta} T_i\left(z_i\left(\hat{\theta}\right), \theta\right) \, \text{d}i.
\]

Given this equation, \(\Box\) becomes \(\frac{d}{d\theta} W_{\text{util}}\left(\hat{\theta}\right) > 0\) and \(\Box\) becomes \(\frac{d}{d\theta} W_{\text{util}}\left(\hat{\theta}\right) = 0\), so that the global improvement principle says that if utilitarian welfare is increasing as we very \(\theta\) from \(\theta_0\) to \(\theta_1\), then utilitarian welfare is greater at \(\theta_1\) than at \(\theta_0\), and the global indifference principle says that if utilitarian welfare is unchanging as we vary \(\theta\), then utilitarian welfare is the same at \(\theta_1\) as at \(\theta_0\). In the utilitarian case, these principles are obviously valid.

In the case of generalized welfare weights, the global improvement and indifference principles are posited by analogy with the utilitarian case. This is the same as the the justification for Saez and Stantcheva’s definitions for a local desirability of a tax reform and local optimality of a tax policy, which substitute generalized welfare weights \(g_i\left(\hat{\theta}\right)\) for utilitarian welfare weights \(\frac{\partial}{\partial c_i} U_i\left(c_i\left(\hat{\theta}\right), z_i\left(\hat{\theta}\right)\right)\) in principles that are valid for utilitarianism. Indeed, when the parameterized family of tax reforms has the form \(T^\theta = T + \theta \Delta T\) and \(\hat{\theta} = 0\), \(\Box\) simplifies to \(\Box\), Saez and Stantcheva’s condition for a locally desirable tax reform.

The following useful result assumes the global improvement principle and follows from our smoothness assumptions – see the Appendix for the proof.

**Proposition 1. Local improvement principle.**

Let \(g\) be a system of welfare weights, let \((T^\theta)_{\theta \in [g]}\) in \(T\) be a parameterized family of tax policies,

\[\text{saez and stantcheva apply this condition to locally revenue neutral tax reforms; in my main theorem, i use the global improvement and indifference principles to construct a cycle when revenue remains constant.}\]
and let $\theta_0 \in [\underline{\theta}, \bar{\theta})$. If $\int g_i(\theta_0) \frac{\partial}{\partial \theta} \bigg|_{\theta=\theta_0} T_i (z_i(\theta_0), \theta) \, d\theta < 0$, then there exists $\theta_1 \in (\theta_0, \bar{\theta}]$ such that for all $\theta \in (\theta_0, \theta_1)$, $T^{\theta_0} \prec^g T^{\theta}$. Similarly, if $\int g_i(\theta_0) \frac{\partial}{\partial \theta} \bigg|_{\theta=\theta_0} T_i (z_i(\theta_0), \theta) \, d\theta > 0$, then there exists $\theta_1 \in (\theta_0, \bar{\theta}]$ such that for all $\theta \in (\theta_0, \theta_1)$, $T^{\theta_0} \succ^g T^{\theta}$.

3.3 Pareto

Certain Pareto conditions, which are useful below, are implicit in the welfare weights framework. In particular, it follows from (5), which was derived using the envelope theorem, and the fact that the marginal utility of consumption is positive that that the following relation holds that:

$$\forall i, \forall \theta, \quad \frac{d}{d\theta} U_i \bigg( \hat{\theta} \bigg) \geq 0 \Leftrightarrow \frac{\partial}{\partial \theta} \bigg|_{\theta=\hat{\theta}} T_i \bigg( z_i \big( \hat{\theta} \big), \theta \bigg) \geq 0. \quad (6)$$

That is, $\frac{d}{d\theta} U_i \bigg( \hat{\theta} \bigg)$ and $\frac{\partial}{\partial \theta} \bigg|_{\theta=\hat{\theta}} T_i \bigg( z_i \big( \hat{\theta} \big), \theta \bigg)$ always have the opposite sign when nonzero, and otherwise both are zero. This shows that the term $- \frac{\partial}{\partial \theta} \bigg|_{\theta=\hat{\theta}} T_i \bigg( z_i \big( \hat{\theta} \big), \theta \bigg)$ captures preferences in the sense that it points in the same direction as preferences do in response to a change in $\theta$; and it also shows why the global improvement and indifference principles respect preferences. If, at $\hat{\theta}$, an increase in $\theta$ makes all agents better off, the terms $\frac{\partial}{\partial \theta} \bigg|_{\theta=\hat{\theta}} T_i \bigg( z_i \big( \hat{\theta} \big), \theta \bigg)$ will be negative for all agents, and so $\int g_i \big( \hat{\theta} \big) \frac{\partial}{\partial \theta} \bigg|_{\theta=\hat{\theta}} T_i \bigg( z_i \big( \hat{\theta} \big), \theta \big) \, d\theta < 0$, no matter what (positive) welfare weights $g$ are used. This is formalized by the following proposition, which is proved in the Appendix and assumes, as above, that welfare weights are always positive, and also assumes the global improvement and indifference principles.

**Proposition 2.** Let $(T^\theta)_{\theta \in \Theta}$ be in $T$, and let $\theta_0, \theta_1 \in \Theta$ be such that $\theta_0 < \theta_1$.

1. **Pareto indifference along paths.** Suppose that all agents are indifferent among all tax policies $T^\theta$ for $\theta \in [\theta_0, \theta_1]$. Then for all systems of welfare weights $g$, $T^{\theta_0} \sim^g T^{\theta_1}$.

2. **Weak Pareto along paths.** Suppose that for all $\hat{\theta} \in [\theta_0, \theta_1]$ and all agents $i$, $\frac{d}{d\theta} U_i \bigg( \hat{\theta} \bigg) > 0$ so that, for all agents, tax policies become more desirable as $\theta$ increases within $[\theta_0, \theta_1]$. Then for all systems of welfare weights $g$, $T^{\theta_0} \prec^g T^{\theta_1}$.

The Pareto principles stated above are weaker than the standard principles because they only apply to paths of smoothly changing tax policies along which the direction of preferences is constant. Say that a social welfare function is **Paretian along paths** if it satisfies a weakened version of the Pareto principle, analogous to the properties that the above proposition shows to be satisfied by all systems of welfare weights. A formal statement of this property of social welfare functions, as well as of what it means for a system of welfare weights to implement a social welfare function and a proof of the following corollary is in the Appendix.

---

9One could also establish a **strong Pareto along paths** principle along similar lines that handles the case where some but not all agents prefer that $\theta$ increases in $[\theta_0, \theta_1]$, while the rest are indifferent.
Corollary 1. Any social welfare function that is not Paretian along paths cannot be implemented by any system of generalized social welfare weights.

The corollary shows that the expressive power of welfare weights is limited in the sense that non-Paretian (in a weak sense of Paretian) objectives cannot be implemented by welfare weights.

4 Structural utilitarianism

The key condition for generalized welfare weights to be consistent is structural utilitarianism.

Definition 1. A system of welfare weights \( g \) is structurally utilitarian if and only if \( \forall i \in I, \forall z_i, z'_i \in Z, \forall c_i, c'_i \in \mathbb{R}, \)

\[
c_i - v_i (z_i) = c'_i - v_i (z'_i) \Rightarrow g_i (c_i, z_i) = g_i (c'_i, z'_i). \tag{7}
\]

To interpret this definition, observe that given quasilinear utility \( U_i (c_i, z_i) = u (c_i - v_i (z_i)) \), we have \( \frac{\partial}{\partial c_i} U_i (c_i, z_i) = u' (c_i - v_i (z_i)) \). Thus the marginal utility of consumption \( \frac{\partial}{\partial c_i} U_i (c_i, z_i) \) is determined by the quantity \( c_i - v_i (z_i) \), and given our assumption that the outer utility function \( u (\cdot) \) is strictly convex, the condition \ref{7} for structural utilitarianism is equivalent to:

\[
\frac{\partial}{\partial c_i} U_i (c_i, z_i) = \frac{\partial}{\partial c_i} U_i (c'_i, z'_i) \Rightarrow g_i (c_i, z_i) = g_i (c'_i, z'_i). \tag{8}
\]

Thus, structural utilitarianism allows that welfare weights are not necessarily equal to the marginal utility of consumption \( \frac{\partial}{\partial c_i} U_i (c_i, z_i) \), the utilitarian welfare weight, but it requires that welfare weights are determined by the marginal utility of consumption in the sense that, if \( i \)'s marginal utility of consumption does not change, then \( i \)'s welfare weight does not change. Note that the condition is imposed separately on each agent \( i \); it is a condition on how that agent's welfare weight changes as their allocation \((c_i, z_i)\) changes, and no relation is posited between the welfare weights of different agents \( i \) and \( j \). So structural utilitarianism is consistent with welfare weights being dependent on agents' characteristics \((x_i, y_i)\). Recalling that utility is given by \( U_i (c_i, z_i) = u (c_i - v_i (z_i)) \), utility is also determined by the quantity \( c_i - v_i (z_i) \). When the outer utility function \( u (\cdot) \) is both strictly increasing and concave, \( \frac{\partial}{\partial c_i} U_i (c_i, z_i) = \frac{\partial}{\partial c_i} U_i (c'_i, z'_i) \) if and only if \( U_i (c_i, z_i) = U_i (c'_i, z'_i) \), and the condition \ref{7} for structural utilitarianism is also equivalent to:

\[
U_i (c_i, z_i) = U_i (c'_i, z'_i) \Rightarrow g_i (c_i, z_i) = g_i (c'_i, z'_i). \tag{9}
\]

Thus, structural utilitarianism can also be interpreted as saying that \( i \)'s welfare weight doesn't change when \( i \)'s utility doesn't change. The coincidence of \ref{8} and \ref{9} depends on the assumption of quasilinear utility, and, indeed, the condition required to guarantee consistency of welfare weights changes when utility is no longer assumed quasilinear.

Define \( \hat{U}_i (c_i, z_i) = c_i - v_i (z_i) \). \( \hat{U}_i (c_i, z_i) \) is a utility function over \((c_i, z_i)\) pairs that is ordinally equivalent to \( U_i (c_i, z_i) \). Define the variable \( \hat{u}_i \) by \( \hat{u}_i = \hat{U}_i (c_i, z_i) \). The variable gives a representation
of the agent’s utility as a function of consumption $c_i$ and income $z_i$. We can then re-express welfare weights as a function $\hat{g}_i(\hat{u}_i, z_i)$ of utility and income $(\hat{u}_i, z_i)$ rather than as a function $g_i(c_i, z_i)$ of consumption and income $(c_i, z_i)$. The relationship between the two expressions is as follows:

$$\hat{g}_i(\hat{u}_i, z_i) = g_i(\hat{u}_i + v_i(z_i), z_i), \quad \forall \hat{u}_i \in \mathbb{R}, \forall z_i \in Z.$$  

(10)

The theorem has the following important corollary: (The straightforward proof is in the Appendix.)

**Corollary 2.** The theorem has the following important corollary:

**Proposition 3.** Let $g$ and $\hat{g}$ be related as in (10). Then welfare weights $g$ are structurally utilitarian if and only if $\forall i \in I, \forall \hat{u}_i \in \mathbb{R}, \forall z_i \in Z, \frac{\partial}{\partial z_i} \hat{g}_i(\hat{u}_i, z_i) = 0$.

We now come to the theorem that shows that when welfare weights are structurally utilitarian, they correspond to a global social ranking. Say that $W : \mathcal{S} \to \mathbb{R}$ is a generalized utilitarian social welfare function if there exists a real-valued function $F_i(u_i) = F(u_i, x_i, y_i)$, which is (i) smooth in $u_i$ and smooth in $(u_i, x_i, y_i)$ unless $(x_i, y_i)$ are discrete and (ii) strictly increasing in $u_i$, such that $W(T) = \int F_i(U_i(c_i(T), z_i(T)))$ d$t$.

It follows from the envelope theorem that for all $(T^\theta)_{\theta \in \Theta}$ in $T$ and $\theta_0 \in \Theta$,

$$\frac{d}{d\theta} \bigg|_{\theta = \theta_0} W(T^\theta) = - \int F_i'(U_i(c_i(\theta_0), z_i(\theta_0))) \frac{\partial}{\partial c_i} U_i(c_i(\theta_0), z_i(\theta_0)) \frac{\partial}{\partial \theta} T_i(z_i(\theta_0), \theta) \, dt.$$  

(11)

It follows that $F_i'(U_i(c_i, z_i)) \frac{\partial}{\partial c_i} U_i(c_i, z_i)$ are the social welfare weights arising from a generalized utilitarian social welfare function. Formally, say that a system of welfare weights $g$ arise from a generalized utilitarian social welfare function if there exists $F_i(u_i) = F(u_i, x_i, y_i)$ satisfying properties (i) and (ii) above such that for all $i, c_i$, and $z_i$, $g_i(c_i, z_i) = F_i'(U_i(c_i, z_i)) \frac{\partial}{\partial c_i} U_i(c_i, z_i)$.

**Theorem 1.** Welfare weights $g$ are structurally utilitarian if and only if they arise from a generalized utilitarian social welfare function.

The theorem has the following important corollary:

**Corollary 2.** If welfare weights $g$ are structurally utilitarian, then there exists a generalized utilitarian social welfare function $W$ from which the welfare weights can be derived in the sense that for all $(T^\theta)_{\theta \in \Theta}$ in $T$ and $\theta_0 \in \Theta$, $\frac{d}{d\theta} \bigg|_{\theta = \theta_0} W(T^\theta) = - \int g_i(T^{\theta_0}) \frac{\partial}{\partial \theta_0} T_i(z_i(T^{\theta_0}), \theta) \, d$t$, so that the welfare weights correspond to a consistent social ranking.

Both the theorem and the corollary are proved in the Appendix. It should be clear that if welfare weights arise from a social welfare function, then it is not possible to use them to construct a social preference cycle. A proof sketch of Theorem 1 is as follows. First, if welfare

---

\(^{10}\)Note that I build smoothness into the definition of a generalized utilitarian social welfare function because I assumed similar smoothness properties on welfare weights. If the smoothness requirements on welfare weights were relaxed somewhat, one could correspondingly weaken the smoothness requirements for a generalized utilitarian social welfare function and still prove a corresponding version of Theorem 1 below.

\(^{11}\)One might wonder why the social welfare function in the above theorem is additively separable; the answer is that $i$’s welfare weight is assumed to depend only on $i$’s consumption, income, and characteristics, and not on the distribution of these in society.
weights are a function of \( \hat{v} \). These weights are structurally utilitarian because, given quasilinearity, both \( U_i(c_i, z_i) \) and \( \frac{\partial}{\partial c_i} U_i(c_i, z_i) \) are uniquely determined by \( c_i - v_i(z_i) \). Going in the other direction, by Proposition 3, structural utilitarianism is equivalent to the requirement that welfare weights are a function of \( \hat{v} \). Then define the utility function

\[
F_U(c_i, z_i) = \frac{\partial}{\partial c_i} U_i(c_i, z_i).
\]

It follows that there exists a strictly increasing function \( g \) such that

\[
F_i(c_i, z_i) = g_i(c_i - v_i(z_i)) = \hat{g}_i(v_i(z_i)).
\]

Then define the utility function \( W_i(c_i, z_i) = w_i(c_i - v_i(z_i)) \). Observe that the utility function \( W_i(c_i, z_i) \) is ordinally equivalent to \( U_i(c_i, z_i) \) in the sense that the two represent the same preferences over consumption and income. It follows that there exists a strictly increasing function \( F_i \) such that \( W_i(c_i, z_i) = F_i(U_i(c_i, z_i)) \). Since \( g_i(c_i, z_i) = g(c_i, z_i, x_i, y_i) \), there exists some function \( F \) such that \( F_i(u_i) = F(u_i, x_i, y_i) \). By construction, \( g_i(c_i, z_i) = \hat{g}_i(c_i - v_i(z_i)) = w'_i(c_i - v_i(z_i)) = \frac{\partial}{\partial c_i} W_i(c_i, z_i) = F'_i(U_i(c_i, z_i)) \frac{\partial}{\partial c_i} U_i(c_i, z_i) \), which is what we need to show.

5 A simple version of the main theorem

5.1 The special case when taxes can be completely individualized

I now prove a simplified version of my main result. A stronger version is found in Section 6.

Consider the special case in which taxes can be completely individualized so that each agent \( i \) faces an individualized tax schedule \( T^i_\theta \) that can differ from the tax schedule faced by other agents. In our framework, this is possible if each agent’s observable characteristics uniquely identify them: formally, for all \( i, j \in I, i \neq j \Rightarrow x_i \neq x_j \). I assume that the map \( i \mapsto x_i \) is smooth, that there are no unobservable characteristics \( y_i \), and that the functions \( u(\cdot), (z_i, x_i) \mapsto v(z_i, x_i), (c_i, z_i, x_i) \mapsto g(c_i, z_i, x_i) \) are smooth. This case is not interesting from an optimal tax perspective because we can simply set the marginal tax rate equal to zero for each agent, so that all agents earn the efficient level of income and we can meet the revenue requirement and achieve any redistribution we wish via individualized lumpsum taxes. However, the assumption of completely individualized taxes does allow us to illustrate the problems with welfare weights in a simple way.

Theorem 2. Suppose that taxes can be completely individualized. If welfare weights \( g \) are not structurally utilitarian, then they are inconsistent in the sense there exist tax policies \( T_0, T_1, T_2, T_3 \), each of which raises the same revenue, and such that welfare weights imply a social preference cycle of the form: \( T_0 \prec^g T_1 \sim^g T_2 \prec^g T_3 \sim^g T_0 \).

A proof sketch follows. Assume that welfare weights are not structurally utilitarian. Then it is possible to construct a completely individualized family \( (T^0_\theta) \) of tax policies such that for some set \( S \) of agents, where both \( S \) and the set of agents not in \( S \) have positive measure, we have that

1. for agents not in \( S \), taxes are completely unchanged as \( \theta \) varies;
2. for agents in \( S \), the optimal response \( (c_i, z_i) \) to taxes changes as \( \theta \) changes in such a way that the aggregate welfare weight on \( S \), \( g_S = \int_S g_i(c_i, z_i) \, d\mu_i \), changes but the utility of each agent \( i \) is unchanged, so that agents in \( S \) are indifferent about the value of \( \theta \).
That it is possible to construct a family with the second property follows from the assumption that welfare weights are not structurally utilitarian. The characterization (9) of structural utilitarianism implies that, for some agent \(i\), it is possible to vary \((c_i, z_i)\) in such a way that utility \(U_i(c_i, z_i)\) does not change, but the welfare weight \(g_i(c_i, z_i)\) changes. By the smoothness of welfare weights and utility functions, this holds for all agents in a neighborhood \(S\) of \(i\), and we may choose the neighborhood so that \(g_i(c_i, z_i)\) changes in the same direction for all agents in \(S\) as \(\theta\) changes, and hence the aggregate welfare weight \(g_S\) changes as well. These changes can be brought about as optimal responses to a linear tax individualized policy (for agents in \(S\)),

\[
T_{\theta i}(z_i) = \tau_i(\theta) z_i - \kappa_i(\theta),
\]

where the marginal tax rate \(\tau_i(\theta)\) controls the choice pretax income \(z_i\) and consumption \(c_i\) is brought to desired level by the lumpsum tax \(\kappa_i(\theta)\). In the above construction, all agents are indifferent as \(\theta\) changes. So it follows from part 1 of Proposition 2 – Pareto in difference along paths – that, letting \(\theta\) vary from \(\theta_0\) to \(\theta_1\), welfare weights will imply that the resulting change is socially indifferent:

\[
T^{\theta_0} \sim^g T^{\theta_1}. \tag{12}
\]

Let \(O\) be a positive measure set of agents that is disjoint from \(S\), and such that the set of agents outside of both \(S\) and \(O\) has positive measure. By our assumptions, the aggregate welfare weight \(g_S\) on agents in \(S\) changes as \(\theta\) varies between \(\theta_0\) and \(\theta_1\), while the aggregate welfare weight \(g_O = \int_O g_i(c_i, z_i) \, di\) on agents in \(O\) does not change. It follows that the social marginal rate of substitution \(g_S/g_O\) of consumption of agents in \(S\) for consumption of agents in \(O\) changes as \(\theta\) moves from \(\theta_0\) to \(\theta_1\). Assume without loss of generality that \(g_S\) increases as \(\theta\) increases. It follows that there exists some pair of payments \(t_S\) and \(t_O\), such that, for sufficiently small \(\epsilon > 0\), increasing taxes for agents in \(S\) by \(\epsilon t_S\) lumpsum, while reducing the taxes of agents in \(O\) by \(\epsilon t_O\) lumpsum is desirable at \(\theta_0\) and undesirable at \(\theta_1\). Formally, if we define \(T^{\theta,\epsilon}\) by:

\[
T^{\theta,\epsilon}_i(z_i) = \begin{cases} 
T^\theta_i(z_i) + \epsilon t_S, & \text{if } i \in S; \\
T^\theta_i(z_i) - \epsilon t_O, & \text{if } i \in O; \\
T^\theta_i(z_i), & \text{otherwise}. 
\end{cases} \tag{13}
\]

It then follows that if \(t_S\) and \(t_O\) are chosen as described above, then for sufficiently small \(\epsilon > 0\),

\[
T^{\theta_0,\epsilon} \sim^g T^{\theta_1,\epsilon}, \quad T^{\theta_0} \sim^g T^{\theta_0,\epsilon}, \quad T^{\theta_1} \sim^g T^{\theta_1,\epsilon}. \tag{14}
\]

Formally this part of the argument appeals to Proposition 1 – the local improvement principle. \(T^{\theta,\epsilon}\) differs from \(T^\theta\) for each agent \(i\) at most by a change in the lumpsum payment that is independent of \(\theta\). Because utility is quasilinear, \(T^{\theta,\epsilon}\) then inherits from \(T^\theta\) the property that each agent is indifferent at \(\theta\) changes, so that again by Pareto indifference along paths (Proposition 2),

\[
T^{\theta_0,\epsilon} \sim^g T^{\theta_1,\epsilon}. \tag{15}
\]
Putting (12), (14), and (15) together, we have that for sufficiently small $\epsilon > 0$,

$$T^{\theta_0} \prec^g T^{\theta_0,\epsilon} \sim T^{\theta_1,\epsilon} \prec^g T^{\theta_1} \sim^g T^{\theta_0}.$$  

(16)

So on the assumption that welfare weights are not structurally utilitarian, we have constructed a social preference cycle.

The last step is to show that revenue can be held fixed across the tax policies in the cycle. This requires a modification of the tax policies $T^{\theta}$ and $T^{\theta,\epsilon}$. Observe that $T^{\theta} = T^{\theta,\epsilon}$ when $\epsilon = 0$, so we can identify $T^{\theta}$ and $T^{\theta,0}$. Now consider a positive measure set of agents $Q$, which is disjoint from both $S$ and $O$. We modify the tax policies $T^{\theta,\epsilon}$ only for agents $i$ in $Q$, and otherwise these policies are not altered. We assume that for $i \in Q$, $T^{\theta,\epsilon}_i (z_i) = \bar{\tau} (\theta, \epsilon) z_i + \bar{\kappa}_i (\theta, \epsilon)$, where $\bar{\tau} (\theta, \epsilon)$ is a marginal tax rate, common to agents in $Q$, and $\bar{\kappa}_i (\theta, \epsilon)$ is a lumpsum tax. We may assume, that for each agent $i \in Q$, the lumpsum tax $\bar{\kappa}_i (\theta, \epsilon)$ is chosen so as to offset any utility change as the marginal tax rate $\bar{\tau} (\theta, \epsilon)$ changes, so that agents in $Q$ are indifferent among tax policies $T^{\theta,\epsilon}$ as $\theta$ and $\epsilon$ vary. Note however that if the marginal tax rate changes, and the lumpsum tax adjusts to keep agents’ utility constant, this will change the revenue raised by the tax policy. We may then also assume that $\bar{\tau} (\theta, \epsilon)$ (which determines $\kappa_i (\theta, \epsilon)$ for each $i \in Q$ up to a constant) is chosen so that the change in revenue among agents in $Q$ just offsets any change in revenue among agents in $S$ and $O$ as $\theta$ and $\epsilon$ change. In this way, we keep revenue constant as we create the social preference cycle. The above arguments establishing the cycle are unaltered because agents in $Q$ are indifferent as $\theta$ and $\epsilon$ change. A formal version of the proof of Theorem 2 is in the Appendix.

### 5.2 A detailed example: libertarian weights

I now present a detailed example. The argument is parallel to that in the previous section, although some of the details differ. In particular, in this example, I no longer assume that taxes can be completely individualized. Instead, I assume that agents have a single observable binary characteristic $x_i$ that takes values $A$ and $B$. For $i \in [0, \frac{1}{2}]$, $x_i = A$ and for $i \in (\frac{1}{2}, 1]$, $x_i = B$, so half of the population has each characteristic. I assume it is possible to condition taxes on the characteristic, but the characteristic is not relevant to payoffs or welfare weights. In particular, all types share the same cost of of earning income $v (z_i, A) = v (z_i, B) = v (z_i) = \frac{1}{2} z_i^2$. I assume that welfare weights are libertarian and identical across agents, so that, for all $i \in [0, 1]$, welfare weights are of the form $g_i (c_i, z_i) = \tilde{g} (t_i)$, where $\tilde{g}$ is increasing in the tax $t_i = z_i - c_i$ paid by each agent.

**Proposition 4.** In the model of the preceding paragraph, welfare weights are inconsistent in the sense there exist tax policies $T_0, T_1, T_2, T_3$, each of which raises the same revenue, and such that welfare weights imply a social preference cycle of the form: $T_0 \prec^g T_1 \sim^g T_2 \prec^g T_3 \sim^g T_0$.

This result resembles Theorem 2. Libertarian weights are not structurally utilitarian. This can be seen in the argument below, in which we construct a tax policy such that utility is held fixed but the total tax paid by specific agents, and hence also their libertarian welfare weight, varies.
I now establish the proposition. Consider linear taxes of the form $T(z) = \kappa + \tau z$, where $\tau$ is the marginal tax rate and $\kappa$ is a lumpsum payment. If an agent with cost function $v(z) = \frac{1}{2}z^2$ faces such a tax schedule, the agent will solve the problem $\max_z [z (1 - \tau) - \frac{1}{2}z^2 - \kappa]$. If we define $z(\tau)$ as the optimal solution to the agent’s problem when the agent faces a marginal tax rate of $\tau$, the first order condition for the agent’s optimization problem implies that $z(\tau) = 1 - \tau$. Observe that, because utility is quasilinear, the optimal income $z(\tau)$ does not depend on the lumpsum tax. Recall that $\hat{U}_i(c_i, z_i) = c_i - v_i(z_i)$ is a utility representation that is ordinally equivalent to $U_i(c_i, z_i) = u(c_i, -v_i(z_i))$. Define $\kappa(\tau)$ to solve:

$$
\frac{z(\tau)(1 - \tau)}{\text{income post marginal tax}} - \frac{\kappa(\tau)}{\text{lumpsum}} - \frac{v(z(\tau))}{\text{cost of earning income}} = 0.
$$

That is, $\kappa(\tau)$ is the lumpsum tax that makes agents’ utility equal to zero — using the utility representation $\hat{U}_i(c_i, z_i)$ — when facing marginal tax rate $\tau$. Given our assumptions, $\kappa(\tau) = \frac{1}{2} (1 - \tau)^2$.

Consider now the doubly parameterized family of tax policies $(T^{\tau, \epsilon})$ — parameterized by $\tau$ and $\epsilon$, where we assume that $\tau \in [0, 1]$.\(^{12}\)

$$
T^{\tau, \epsilon}_i(z_i) = \begin{cases} 
\tau z_i + \kappa(\tau) + \epsilon, & \text{if } x_i = A, \\
\left(\sqrt{1 - \tau^2}\right) z_i + \kappa\left(\sqrt{1 - \tau^2}\right) - \epsilon, & \text{if } x_i = B.
\end{cases}
$$

Observe first that $\epsilon$ just parameterizes a transfer from agents with characteristic $A$ to agents with characteristic $B$; since utility is quasilinear, such a transfer does not lead to a behavioral response, and hence, because there is an equal mass of type $A$ and type $B$ agents, the transfer is revenue neutral. Agents with characteristic $A$ face a marginal tax rate of $\tau$, and agents with characteristic $B$ face a marginal tax rate of $\sqrt{1 - \tau^2}$. As $\tau$ rises from 0 to 1, the marginal tax rate of type $A$ agents rises from 0 to 1 while the marginal tax rate of type $B$ agents falls from 1 to 0. Moreover, as $\tau$ rises from 0 to 1, the per agent revenue raised from type $A$ agents falls from $\frac{1}{2} + \epsilon$ to $0 + \epsilon$, and, the per agent revenue raised from type $B$ agents rises from $0 - \epsilon$ to $\frac{1}{2} - \epsilon$. The formula $\sqrt{1 - \tau^2}$ was chosen for type $B$ agents’ marginal tax rate because this is the formula required for the revenue effects from type $A$ and type $B$ agents to exactly offset one another so that the total revenue of the tax policy remains equal to $\frac{1}{2}$ for all $\tau$ and $\epsilon$.\(^{13}\)

When $\epsilon = 0$, observe that the lumpsum tax $\kappa(\tau)$ is chosen so as to keep type $A$ agents’ utility

\[12\] We have $\kappa(\tau) = z(\tau)(1 - \tau) - v(z(\tau)) = (1 - \tau)(1 - \tau) - \frac{1}{4} \left((1 - \tau)^2\right) = \frac{1}{4} (1 - \tau)^2$.

\[13\] Note that, when $\tau = 0$, $z_i(T^{\tau, \epsilon}) = 0$ if $x_i = B$ and, when $\tau = 1$, $z_i(T^{\tau, \epsilon}) = 0$ if $x_i = A$, so strictly speaking $T^{0, \epsilon}, T^{1, \epsilon} \notin \mathcal{F}$ as $\mathcal{F}$ requires $z_i(T) > 0, \forall i$ (see Section 3.1). However, the requirement $z_i(T) > 0$ was only for convenience to simplify other arguments. In any event, that $z_i(T)$ is sometimes zero is inessential to the argument in this section, and if, for some small $\delta > 0$, we replaced $T^{0, \epsilon}$ by $T^{\delta, \epsilon}$ and $T^{1, \epsilon}$ by $T^{1 - \delta, \epsilon}$, the argument would be essentially unchanged and $T^{\delta, \epsilon}, T^{1 - \delta, \epsilon} \in \mathcal{F}$.

\[14\] This is verified by the calculation: $R(T^{\tau, \epsilon}) = \frac{1}{2} \left[z(\tau) + \kappa(\tau) + \epsilon\right] + \frac{1}{2} \left[z\left(\sqrt{1 - \tau^2}\right) \sqrt{1 - \tau^2} + \kappa(\tau) - \epsilon\right] = \frac{1}{2} \left[(1 - \tau^2) + \frac{1}{2} (1 - \tau)^2 \right] + \frac{1}{2} \left[(1 - \sqrt{1 - \tau^2}) \sqrt{1 - \tau^2} + \frac{1}{2} (1 - \sqrt{1 - \tau^2})^2\right] = \frac{1}{2} \left[\left(\frac{1}{2} (1 - \tau) (1 + \tau)\right) + \frac{1}{2} \left(\frac{1}{2} (1 - \sqrt{1 - \tau^2}) (1 + \sqrt{1 - \tau^2})\right)\right] = \frac{1}{2} (1 - \tau^2) + \frac{1}{4} \tau^2 = \frac{1}{2}$.
equal to zero as \( \tau \) varies. So type \( A \) agents are indifferent among all tax policies of the form \( T^{\tau,0} \). Likewise the lumpsum tax \( \kappa \left( \sqrt{1 - \tau^2} \right) \) makes type \( B \) agents indifferent among all tax policies of the form \( T^{\tau,0} \). Because utility is quasilinear, these indifference conditions continue to hold if, in addition, there is a fixed transfer from type \( A \) to type \( B \) agents. So for any fixed \( \epsilon \), all agents are indifferent among the tax policies \( T^{\tau, \epsilon} \) as \( \tau \) varies. So by part \( 1 \) of Proposition \( 2 \) – Pareto indifference along paths – it follows that varying \( \tau \) from 0 to 1 is socially indifferent:

\[ \forall \epsilon, \quad T^{0, \epsilon} \sim^g T^{1, \epsilon}. \]  

(17)

For any fixed \( \tau \), let us write \( T^{\tau, \epsilon}_i (z_i, \epsilon) = T^{\tau, \epsilon}_i (z_i) \). Under \( T^{0,0} \), the marginal tax rate for type \( A \) agents is 0, but the lumpsum tax is \( \kappa(0) = \frac{1}{2} (1 - 0)^2 = \frac{1}{2} \), so that the total per person tax for type \( A \) agents is \( \frac{1}{2} \). In contrast, the marginal tax rate for type \( B \) agents is 1, inducing no labor supply and zero income, and the lumpsum tax is \( \kappa(1) = \frac{1}{2} (1 - 1)^2 = 0 \), so that, the total per person tax for type \( B \) agents is 0. Since the libertarian weights \( \tilde{g}(t) \) are increasing in taxes paid \( t \), we have

\[
\int_0^1 g_i \left( T^{0,0}_i \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} T^{0,0}_i (z_i (T^{0,0}_i), \epsilon) \, di = \int_0^{\frac{1}{2}} \left[ \tilde{g} \left( \frac{1}{2} \right) \times 1 \right] \, di + \int_{\frac{1}{2}}^1 \left[ \tilde{g}(0) \times (-1) \right] \, di \\
= \frac{1}{2} \tilde{g} \left( \frac{1}{2} \right) - \frac{1}{2} \tilde{g}(0) > 0.
\]  

(18)

At \( T^{1,0} \), the situation is reversed so that type \( A \) agents pay a per person tax of 0, while type \( B \) agents pay a per person tax of \( \frac{1}{2} \). So we have:

\[
\int_0^1 g_i \left( T^{1,0}_i \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} T^{1,0}_i (z_i (T^{1,0}_i), \epsilon) \, di = \frac{1}{2} \tilde{g}(0) - \frac{1}{2} \tilde{g} \left( \frac{1}{2} \right) < 0.
\]  

(19)

It follows from (18), (19), and Proposition 1 – the local improvement principle – that

\[
T^{0,0} \prec^g T^{0, \epsilon}, \quad T^{1,0} \prec^g T^{1, \epsilon}, \quad \text{for sufficiently small } \epsilon > 0.
\]  

(20)

In other words, transferring a dollar from type \( A \) to type \( B \) agents is bad at \( T^{0,0} \) according to libertarian weights because, at \( T^{0,0} \), agents type \( A \) agents pay more tax than type \( B \) agents, and good at \( T^{1,0} \) because, at \( T^{1,0} \), type \( A \) agents pay less in tax than type \( B \) agents. Putting (17) and (20) together, for sufficiently small \( \epsilon > 0 \), we have

\[
T^{1,0} \prec^g T^{1, \epsilon} \sim^g T^{0, \epsilon} \prec^g T^{0,0} \sim^g T^{1,0}.
\]  

(21)

This establishes that the libertarian welfare weights imply a cycle.

\[^{15}\text{See footnote} \, 13\]
6 The main theorem without individualized taxes

In this section, I show that it is possible to generate social preference cycles when all agents face the same tax schedule. The argument then becomes more complicated but its overall structure is similar. One of the reasons that the proof becomes more complicated is that if taxes are not individualized, it will no longer be possible to hold all agents indifferent as we modify taxes in a nontrivial way. So the proof of the theorem in the general case no longer appeals to the Pareto indifference principle inherent in the welfare weights approach (Proposition 2). The step in the preceding argument in which all agents are kept indifferent as the parameter $\theta$ varies is replaced by a step in which if benefits and costs to different agents are aggregated according to the system of social welfare weights $g$, then the change as $\theta$ varies is socially indifferent.

6.1 Additional assumptions

For the main result, I assume that there are no observable characteristics, but there is a single one-dimensional real valued unobservable characteristic $y$. Because there are no observable characteristics on which to condition taxes, I omit the subscript $i$ on taxes and write $T(z_i)$ rather than $T_i(z_i)$. This also simplifies the sets $\hat{T}$ defined in Section 3.1. The set $\hat{T}$ now becomes the set of tax policies $T$ such that (i) for each $i \in I$, the optimal income $z_i(T)$ exists, is unique, $z_i(T) > 0$, and $\frac{d^2}{dz_i^2} \hat{U}_T(z_i(T)) < 0$, and (ii) the map $z \mapsto T(z)$ is smooth. In the definition of $T$, the class of smooth parameterized families of tax policies $(T_\theta)$, in Section 3.1, the condition 2 simplifies to: the map $(z, \theta) \mapsto T(z, \theta)$ is smooth. I assume that the function $i \mapsto y_i$ assigning to each $i$ their characteristic $y_i$ is smooth, strictly increasing in $i$, and, more specifically, the derivative of $y_i$ with respect to $i$ is positive at all values of $i$ in $I = [0, 1]$. In this case we can write $v_i(z_i) = v(z_i, y_i)$ and $g_i(c_i, z_i) = g(c_i, z_i, y_i)$. Moreover, I assume that a higher value of $y$ corresponds to the ability to earn income at a lower cost, so that $\forall z, \forall y, \frac{\partial^2}{\partial y \partial z} v(z, y) < 0$. This means that, for any tax policy $T$ in $\hat{T}$, agents with a higher index $i$ – hence a higher value of $y_i$ – earn higher income.

6.2 Statement of the theorem

Theorem 3. Under the supplementary assumptions of Section 6.1 if welfare weights weights $g$ are not structurally utilitarian, then there exist tax policies $T_0, T_1, T_2, T_3$, each of which raises the same revenue, and such that welfare weights imply a social preference cycle of the form $T_0 \prec^g T_1 \sim^g T_2 \prec^g T_3 \sim^g T_0$.

Together, Theorems 1 and 3 characterizes the exact property on welfare weights – structural utilitarianism – that is required for welfare weights to be consistent. If welfare weights are structurally utilitarian, they are compatible with a social welfare function, and hence a consistent social preference, and if welfare weights are not structurally utilitarian, they imply a social preference cycle. This means that to acquire a consistent method of evaluating tax policies from welfare weights, welfare weights must be quite similar to traditional welfare weights, and the promise of the GSMWW
approach that one can represent very general values with a generalized welfare weights approach
is not fulfilled. To really represent broader values, we need to seek more general approaches that
differ more fundamentally from the traditional utilitarian approach.

6.3 Proof sketch

Here I sketch the proof of the main theorem; the missing details can be found in the Appendix.
Like in the proof of the simpler version of the theorem in Section 5.1, we construct a doubly
parameterized family of tax policies, \((T^{\theta,e})_{\theta \in \Theta, e \in E}\), where \(\Theta = [\underline{\theta}, \overline{\theta}]\) and \(E = [\underline{e}, \overline{e}]\). Say that
\((T^{\theta,e}) \in T_2\) if, for all \(e \in E\), \((T^{\theta,e})_{\theta \in \Theta} \in T\), and, for all \(\theta \in \Theta\), \((T^{\theta,e})_{e \in E} \in T\). Heuristically, we
can think of \(e\) as parameterizing a redistribution from some a set of higher income agents
\(S\) to a set of lower income agents \(O\) – as \(e\) rises, taxes on agents in \(S\) rise while those in \(O\) fall. The specific
construction of \(T^{\theta,e}\) in the Appendix bears out this interpretation (see the proof of Lemma 3), and,
in this way, the argument resembles the argument in Section 5.1.

6.3.1 Sufficient conditions for a social preference cycle

Now suppose that we construct such a family \((T^{\theta,e})\) with the following two properties:

1. Indifference to \(\theta\). Holding fixed \(e\), the value of \(\theta\) is socially indifferent:

   \[
   \forall e \in E, \forall \theta' \in \Theta, \int g_i (\theta', e) \left[ \frac{\partial}{\partial \theta} \right] T (z_i (\theta', \epsilon), \theta, \epsilon) \, di = 0. \tag{22}
   \]

2. Changing desirability of redistribution \(e\). There exist \(\theta_0 \in (\underline{\theta}, \overline{\theta})\) and \(e_0 \in (\underline{e}, \overline{e})\) such that
   at \((\theta_0, e_0)\), as \(\theta\) crosses \(\theta_0\), a change in \(e\) goes from being undesirable to being desirable:

   \[
   \left. \int g_i (\theta_0, e_0) \left[ \frac{\partial}{\partial e} \right] T (z_i (\theta_0, e_0), \theta_0, \epsilon) \, di \right|_{\epsilon = e_0} = 0, \tag{23}
   \]

   \[
   \left. \frac{d}{d\theta} \right|_{\theta = \theta_0} \int g_i (\theta, e_0) \left[ \frac{\partial}{\partial e} \right] T (z_i (\theta, e_0), \theta, \epsilon) \, di < 0. \tag{24}
   \]

The following lemma shows that, if we can construct a family \((T^{\theta,e})\) with the above properties,
that is sufficient to construct a social preference cycle.

**Lemma 1.** Suppose that welfare weights \(g\) are such that there exists a family \((T^{\theta,e}) \in T_2\) that satisfies \(22\) and \(24\). Then there exist parameter values \(\theta_-, \theta_+ \in \Theta\) and \(e_0, e_+ \in E\) for which there exists a social preference cycle of the form

\[
T^{\theta_+, e_0} \prec g T^{\theta_+, e_+} \sim g T^{\theta_-, e_+} \prec g T^{\theta_-, e_0} \sim g T^{\theta_+, e_0}.
\]
Proof. Suppose there is a family \((T^{\theta, \epsilon})\) satisfying (22)–(24). It follows from (23) and (24) that for \(\theta_- \in \Theta\) such that \(\theta_- < \theta_0\) and \(\theta_-\) is sufficiently close to \(\theta_0\),

\[
\int g_i (\theta_-, \epsilon_0) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} T (z_i (\theta_-, \epsilon_0), \theta_-, \epsilon) \, d\epsilon > 0,
\]

and for \(\theta_+ \in \Theta\) such that \(\theta_+ > \theta_0\) and \(\theta_+\) is sufficiently close to \(\theta_0\),

\[
\int g_i (\theta_+, \epsilon_0) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} T (z_i (\theta_+, \epsilon_0), \theta_+, \epsilon) \, d\epsilon < 0.
\]

It follows from (23), (27), and the local improvement principle (Proposition 1) that for \(\epsilon_+ \in E\) such that \(\epsilon_+ > \epsilon_0\) and \(\epsilon_+\) sufficiently close to \(\epsilon_0\), \(T^{\theta_-, \epsilon_0} \succ T^{\theta_-, \epsilon_+}\) and \(T^{\theta_+, \epsilon_0} \succ T^{\theta_+, \epsilon_+}\). It follows from (22) and the global indifference principle (Section 3.2) that \(T^{\theta_-, \epsilon_0} \sim T^{\theta_+, \epsilon_0}\) and \(T^{\theta_-, \epsilon_+} \sim T^{\theta_+, \epsilon_+}\). Putting the just derived relations together, we derive the cycle (25). \(\square\)

6.3.2 Non-structurally utilitarian weights allow a family \((T^{\theta, \epsilon})\) satisfying the sufficient conditions for a social preference cycle

I now show that the sufficient conditions for a social preference cycle (22)–(24) are jointly satisfiable if (and only if) welfare weights are not structurally utilitarian. It is convenient to define \(\hat{U}_i (T) = \hat{U}_i (c_i (T), z_i (T))\). I begin by stating a fairly immediate corollary of Proposition 3 which is proved in the Appendix:

**Corollary 3.** If \(g\) is not structurally utilitarian, then there exists tax policy \(T \in \hat{T}\) for which there exist agents \(i_a, i_b \in (0, 1)\) with \(i_a < i_b\) such that either

\[
\forall i \in (i_a, i_b), \quad \frac{\partial}{\partial z_{i}} \hat{g}_i \left( \hat{U}_i (T), z_i (T) \right) < 0 \quad (28)
\]

or

\[
\forall i \in (i_a, i_b), \quad \frac{\partial}{\partial z_{i}} \hat{g}_i \left( \hat{U}_i (T), z_i (T) \right) > 0. \quad (29)
\]

Next I show that in the presence of condition (22), condition (24) takes a more convenient form.

**Lemma 2.** Assume that \((T^{\theta, \epsilon}) \in T_2\) satisfies (22). Then (24) holds if and only if

\[
\int \frac{\partial}{\partial z_i} \hat{g}_i \left( \hat{U}_i (\theta_0, \epsilon_0), z_i (\theta_0, \epsilon_0) \right) \left[ \frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} z_i (\theta, \epsilon_0) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} T (z_i (\theta_0, \epsilon_0), \theta_0, \epsilon) \right. \\
- \left. \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} z_i (\theta_0, \epsilon) \frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} T (z_i (\theta_0, \epsilon_0), \theta, \epsilon) \right] \, d\epsilon < 0. \quad (30)
\]

Since, by Proposition 3 for structurally utilitarian weights, \(\frac{\partial}{\partial z_i} \hat{g}_i \left( \hat{U}_i (\theta_0, \epsilon_0), z_i (\theta_0, \epsilon_0) \right) = 0\) everywhere, the integral in the left-hand side of (30) is always equal to zero when welfare weights are
structurally utilitarian. Hence, it follows immediately from Lemma 2 that a necessary condition for (22)-(24) to be satisfied is for welfare weights not to be structurally utilitarian. However, what we need to show here is that not being structurally utilitarian is a sufficient condition for the ability to construct a family of tax policies for which (22)-(24) to hold.

Proof outline of Lemma 2 The key is to show that, when expanded, the expression in the left-hand side of (24) and the ε-derivative of the expression on the left-hand side of (22), evaluated at \((\theta_0, \epsilon_0)\), have overlapping terms. In particular, below I will define terms \(A, B, C\) such that:

\[
\frac{d}{d\epsilon} \int_{\epsilon=0} g_i (\theta_0, \epsilon) \frac{\partial}{\partial \theta} T (z_i (\theta_0, \epsilon), \theta, \epsilon) \, d\epsilon = A + C, \tag{31}
\]

\[
\frac{d}{d\theta} \int_{\epsilon=0} g_i (\theta_0, \epsilon) \frac{\partial}{\partial \epsilon} T (z_i (\theta_0, \epsilon), \theta, \epsilon) \, d\epsilon = B + C. \tag{32}
\]

The terms \(A\) and \(B\) are as follows:

\[
A = \int \frac{\partial}{\partial z_i} \hat{g}_i (\hat{U}_i (\theta_0, \epsilon_0), z_i (\theta_0, \epsilon_0)) \frac{\partial}{\partial \epsilon} z_i (\theta_0, \epsilon) \frac{\partial}{\partial \theta} \hat{T} (\epsilon_0) \, d\epsilon, \tag{33}
\]

\[
B = \int \frac{\partial}{\partial z_i} \hat{g}_i (\hat{U}_i (\theta_0, \epsilon_0), z_i (\theta_0, \epsilon_0)) \frac{\partial}{\partial \theta} z_i (\theta_0, \epsilon) \frac{\partial}{\partial \epsilon} \hat{T} (\epsilon_0) \, d\epsilon. \tag{34}
\]

The term \(C\), as well as the derivation of (31) and (32), are in the Appendix. Note that (22) implies that the left-hand side of (31) is equal to zero, which implies that the right-hand side is equal to zero as well. It follows that \(C = -A\). So \(B + C = B - A\). It follows that the left-hand side of (34) is less than zero—which is what (24) says—if and only if \(B - A < 0\). But \(B - A < 0\) is equivalent to (30). This completes the proof of Lemma 2. \(\square\)

The next Lemma shows that in order to able to construct a family \((T^{\theta, \epsilon})\) that satisfies (22), (23), and (30), it is sufficient to find a tax policy \(T\) and \(i_a, i_b\) for which (29) holds. (The Appendix presents an analogous lemma – Lemma A.2 – corresponding to condition (28).)

**Lemma 3.** Let \(T \in \bar{T}\) and let \(i_a, i_b \in (0, 1)\) be such that \(i_a < i_b\). Then there exists a family \((T^{\theta, \epsilon}) \in T_2\) with \(T^{\theta_0, \epsilon_0} = T\) for some interior parameter values \(\theta_0, \epsilon_0\) and that satisfies (22), (23), and (30), and

\[
\frac{\partial}{\partial \theta} z_i (\theta_0, \epsilon_0) \frac{\partial}{\partial \epsilon} T (\theta_0, \epsilon_0, \theta_0, \epsilon) \begin{cases} < 0, & \text{if } i \in (i_a, i_b), \\ = 0, & \text{if } i \notin (i_a, i_b). \end{cases} \tag{35}
\]

The lemma is proven in the Appendix. This lemma does not depend on any assumptions on welfare weights, but just on the broad flexibility that is available in constructing tax policies.

The following lemma puts together the previous results derived in this section.

**Lemma 4.** If \(g\) is not structurally utilitarian, then there exists a family of tax policies \((T^{\theta, \epsilon}) \in T_2\) satisfying (22)-(24).
Proof. First assume that \( g \) is not structurally utilitarian. It then follows from Corollary \( 3 \) that there exists a tax policy \( T \in \hat{T} \) and \( i_a, i_b \in (0,1) \) such that \( i_a < i_b \) and either \( 28 \) or \( 29 \) hold. First assume that \( 29 \) holds.

It follows from Lemma \( 3 \) that there exists a parameterized family of tax policies \( (T^{\varphi, \epsilon})_{\varphi \in \Theta, \epsilon \in E} \) with \( T^{\varphi_0, \epsilon_0} = T \) satisfying \( 22 \), \( 23 \), and \( 35 \), where, in \( 35 \), \( i_a \) and \( i_b \) are chosen to be the same values for which \( 29 \) holds. Moreover, \( 29 \) and \( 35 \) together imply \( 30 \). So in this case, we can construct \( (T^{\varphi, \epsilon}) \in T_2 \) satisfying \( 22 \), \( 23 \), and \( 30 \). A similar argument – invoking a variant of Lemma \( 3 \) (Lemma \( A.2 \) in Section \( A.10.4 \) of the Appendix) shows that, when \( 28 \) rather than \( 29 \) holds, we can still construct \( (T^{\varphi, \epsilon}) \in T_2 \) satisfying \( 22 \), \( 23 \), and \( 30 \). It now follows from Lemma \( 2 \) that whenever welfare weights are not structurally utilitarian, it is possible to construct a tax policy satisfying \( 22 \)-\( 24 \). □

6.3.3 Holding revenue constant

The construction of the previous section can be extended so that the family \( (T^{\varphi, \epsilon}) \) is be chosen so that revenue is held constant, as stated by the following lemma.

Lemma 5. If \( g \) is not structurally utilitarian, then there exists a constant revenue family of tax policies \( (T^{\varphi, \epsilon}) \in T_2 \) satisfying \( 22 \)-\( 24 \).

Lemma 5 is a strengthening of Lemma \( 4 \) that differs from Lemma \( 4 \) only in that family \( (T^{\varphi, \epsilon}) \) is required to be a constant revenue family in the sense that all tax policies \( T^{\varphi, \epsilon} \) raise the same revenue. I have separated this additional requirement into a separate lemma because the argument that revenue can be held constant appeals to different principles than the proof of the other properties. The basic idea is similar to that described in Section \( 5.1 \) for holding revenue constant. In particular, once we construct a family \( (T^{\varphi, \epsilon}) \) satisfying \( 22 \)-\( 24 \), as we know we can do from Lemma \( 4 \), we consider a positive measure set \( Q \) of agents at a different income level than agents in \( S \) and \( O \), and vary the revenue raised from agents in \( Q \) as \( \varphi \) and \( \epsilon \) vary exactly so as to offset revenue changes elsewhere in the tax schedule in such a way that there is no detectable welfare change in \( Q \) according to welfare weights; this is analogous to moving along a social indifference curve (for agents in \( Q \)) along which revenue varies. The details are in the Appendix.

6.3.4 Putting it all together

Putting Lemmas \( 4 \) and \( 5 \) together yields Theorem \( 3 \), the main result.

6.4 An application: Poverty alleviation

I now present an application to illustrate the main result. Maintain all of the assumptions of Section \( 6.1 \). Let \( \bar{c} \) be the poverty line; that is, \( \bar{c} \) is the level of consumption below which agents are considered to be poor. Now consider welfare weights which capture the goal of poverty alleviation by concentrating weight on agents beneath the poverty line. Saez and Stantacheva presented such
an example. I modify their example slightly to make welfare weights smooth. Suppose that \( g_i(c_i, z_i) = \tilde{g}(c_i) \), where \( \tilde{g}(c_i) \) is decreasing in \( c_i \) until \( c_i \) gets to \( \hat{c} \) and then remains constant at the value \( g \) thereafter, where \( g > 0 \). I assume that \( g > 0 \) to be in conformity with my prior assumptions but we may assume that \( g \) is arbitrarily close to zero. So agents below the poverty line have a higher welfare weight than agents above the poverty line, the welfare weight is greater the further below the poverty line the agent is, and constant for agents above the poverty line.

Now consider a doubly parameterized family of tax policies \((T^{\theta, \epsilon})\) of the form \( T(z, \theta, \epsilon) = T(z, \theta, \epsilon) = Tf(z) + (\theta - \epsilon)z + \alpha \epsilon - \kappa(\theta, \epsilon) \) where \( f(z) \) is a smooth function and, for some \( \theta_0 \), \( \kappa(\theta_0, \epsilon) = 0, \forall \epsilon \). Assume that there exists \( \epsilon_0 \) and income level \( \bar{z} \) (within the income distribution), such that, when facing tax schedule \( T^{\theta_0, \epsilon_0} \), all agents earn positive income, all agents earning income \( \bar{z} \) or above are strictly above the poverty line, and a positive measure of agents with income below \( \bar{z} \) are beneath the poverty line. I assume that \( f(z) = 0 \) for all \( z \) with \( z \leq \bar{z} \), and \( f(z) > 0 \) for all \( z \) with \( z > \bar{z} \), so that the \( \theta f(z) \) term specifies taxes that only apply to agents above the poverty line when \((\theta, \epsilon)\) is close to \((\theta_0, \epsilon_0)\). Noting that the optimal income for \( i \), \( z_i(\theta, \epsilon) \), is independent of \( \alpha \), assume that \( \alpha \) is chosen so that \( \int g(\theta_0, \epsilon_0) [z_i(\theta_0, \epsilon_0) - \alpha] \) = 0, which says that, at \( T^{\theta_0, \epsilon_0} \), the positive welfare effect of increasing \( \epsilon \) due to decreasing marginal tax rates through the term \(-\epsilon z\) is just offset by the negative welfare effect of the increase in the lumpsum tax \( \alpha \epsilon \). (Note that, by our assumptions, \( \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} \kappa(\theta_0, \epsilon) = 0.\) Finally, we assume that \( \kappa(\theta, \epsilon) \) satisfies the following set of differential equations (note that \( g_i(\theta, \epsilon) \) depends on \( \kappa(\theta, \epsilon) \)):

\[
\frac{\partial}{\partial \theta} \bigg|_{\theta=\theta'} \kappa(\theta, \epsilon) = \int \frac{g_i(\theta', \epsilon)}{g_j(\theta', \epsilon)} \bigg[ z_i(\theta', \epsilon) + f(z_i(\theta', \epsilon)) \bigg] \, di, \quad \forall \theta', \forall \epsilon. \tag{36}
\]

Rearranging terms, one can see that (36) says that for any fixed value of \( \epsilon \), when changing \( \theta \), the welfare effect due to increasing marginal tax rates through the term \( \theta f(z) + \theta z \) is just offset by the welfare effect of the change in the lumpsum tax \( \kappa(\theta, \epsilon) \). Note that the differential equations (36) and the conditions \( \kappa(\theta_0, \epsilon) = 0, \forall \epsilon \) uniquely determine \( \kappa(\theta, \epsilon) \).

**Proposition 5.** With poverty alleviation welfare weights, if \((T^{\theta, \epsilon})\) has the properties assumed in this section, \((T^{\theta, \epsilon})\) satisfies (22)-(24), the sufficient conditions for a preference cycle in Lemma 7.

The proof is in the Appendix. Condition (22) corresponds to \( \int g(\theta_0, \epsilon_0) [z_i(\theta_0, \epsilon_0) - \alpha] \) = 0 (and \( \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} \kappa(\theta_0, \epsilon) = 0 \), and (23) corresponds to (36). The key calculation that drives the argument is that:

\[
\frac{d}{d\theta} \bigg|_{\theta=\theta_0} \int g_i(\theta, \epsilon_0) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} T(z_i(\theta, \epsilon_0), \theta, \epsilon) = \int \frac{g_i(c_i(\theta_0, \epsilon_0))}{g_j(c_j(\theta_0, \epsilon_0))} \frac{v_i'(z_i(\theta_0, \epsilon_0))}{v_j'(z_j(\theta_0, \epsilon_0))} \, di \times \int \frac{g_i(c_i(\theta_0, \epsilon_0))}{g_j(c_j(\theta_0, \epsilon_0))} f(z_i(\theta_0, \epsilon_0)) \, di < 0, \tag{37}
\]

which establishes (24). It follows from Proposition 5 that, in the poverty alleviation example, with
tax policies as described above, we can construct a social preference cycle exactly as in the proof of Lemma 1 (see Section 6.3.1 above). We have not worried about holding revenue constant, but Lemma 5 tells us that we can modify the construction of \((T^{d,e})\) so as to hold revenue constant as well. Of course, the reason we could construct a cycle is that poverty reduction welfare weights are not structurally utilitarian. In particular, by increasing both consumption \(c_i\) and income \(z_i\) so as to hold total utility \(u(c_i - v_i(z_i))\) fixed, it is possible to bring an agent above the poverty line, and, in this way, we can change their welfare weight; this is not consistent with structural utilitarianism. In general welfare weights that respond to changes in consumption but do not take into account labor supply costs will not be structurally utilitarian, and hence will lead to social preference cycles.

7 Discussion

The motivation for generalized social marginal welfare weights was as a means of addressing the omission of broader values in economic analysis. I have argued in this paper that this solution does not work, because generalized welfare weights, once they stray too far from traditional utilitarian weights, are inconsistent. In this closing section, I will discuss some related literature, and how the current contribution differs, as well as ways forward on the problem of incorporating broader normative values in economic analysis.

7.1 The Pareto principle and broader values: related literature

Saez and Stantcheva (2016) write “if the weights are nonnegative, then our theory respects the Pareto principle in the sense that, around the local optimum, there is no Pareto improving small reform.” It may appear that Saez and Stantcheva have uncovered a way of incorporating broader values into economic analysis compatibly with the Pareto principle. Several authors, including Sen (1970, 1979a, 1979b) and Kaplow and Shavell (2001, 2009), have argued that incorporating broader moral considerations into economic evaluation is inconsistent with the Pareto principle. Different authors have interpreted this conflict in different ways. Sen interprets this as an argument against insisting on the Pareto principle, whereas Kaplow and Shavell (2001) interpret it as an argument against including non-welfarist considerations in normative economic evaluation. As they say, one philosopher’s modus ponens is another philosopher’s modus tollens. Fleurbaey, Tungodden and Chang (2003) are critical of Kaplow and Shavell (2001), and take a more positive view incorporating broader values compatibly with the Pareto principle. In discussing the Saez and Stantcheva approach critically, Fleurbaey and Maniquet (2018), who also take a more positive view of incorporating broader values, compatibly with Pareto, write,

... the social welfare function approach has been introduced by Bergson (1938) and Samuelson (1947) not out of a taste for elegance, but because it is the only way to define social preferences that are both transitive and Paretian. Therefore, a method that

---

17See also Weymark (2017).
directly weights tax changes at the various earning levels is compatible with transitive
and Paretian social preferences, and then extendable to the study of nonlocal reforms,
only if it relies on the classical framework of the social welfare function. (p. 1059)

This informal passage is closely related to the results developed formally in the current paper.

7.2 The contribution of this paper

The results of Sen and Kaplow and Shavell do not directly apply to the generalized social welfare
weights approach of Saez and Stantcheva (2016) for two reasons: (1) the social welfare weights
approach only claims to provide criteria for local improvements and optima, and not a global
objective, and (2) the sense in which the generalized social welfare weights approach is “Paretian”
according to Saez and Stantcheva is just that local optima are locally Pareto efficient, which is
weaker than the standard Pareto principles such as Pareto indifference or weak Pareto, which are
the principles discussed by Sen and Kaplow and Shavell. The current paper does bridge these gaps
to some extent by showing formally how the local comparisons implied by welfare weights imply
global comparisons and also by establishing Pareto principles implied by welfare weights which are
closer to the standard Pareto principles (see my Proposition 2).

Fleurbaey and Maniquet (2018) only discuss the potential intransitivity of welfare weights briefly
and they do not present a formal result characterizing when generalized social welfare weights
are consistent. Nor do they provide a methodology for collecting the local judgements of the
generalized social welfare weights into implicit global rankings. In this paper, I do both of these
things. I show how to collect the local judgements of generalized social welfare weights into global
social judgements (see Section 3.2) and that the precise property that is necessary and sufficient
for welfare weights to be consistent is structural utilitarianism (see Theorems 1 and 3). Unlike
Fleurbaey and Maniquet, I also construct specific examples of cases in which generalized welfare
weights are inconsistent. Moreover, my result is stronger than the point made by Fleurbaey and
Maniquet in another way. I show that when welfare weights are not structurally utilitarian, they
are not consistent with any social welfare function, Paretian or not. Notice, in this regard, that
Theorem 3 does not mention any Pareto principle; it simply says that if welfare weights are not
structurally utilitarian, then they are inconsistent.

7.3 Two ways forward

I now highlight two ways forward if broader values are to be incorporated into normative economic
analysis and specifically optimal tax. Fleurbaey and Maniquet (2018) write that “the classical
social welfare function framework is more flexible than commonly thought, and can accommodate
a very large set of nonutilitarian values. More specifically, fairness concepts can help solve the
interpersonal comparison difficulties that the utilitarian approach faces when agents have different
preferences by providing useful selections of suitable individual utility indexes,” and their paper
shows that Paretian social welfare functions can capture a broad set of values in an optimal tax
In the setting of the current paper, Theorem 1 shows that structurally utilitarian welfare weights are compatible with a generalized utilitarian social welfare function of the form \( \int F(U_i(c_i, z_i), x_i, y_i) \, di \). We may think of the function \( F(u_i, x_i, y_i) \) as potentially reweighting utilities \( u_i \) on the basis of certain moral considerations which are responsive to the characteristics \((x_i, y_i)\).

Not all values can be captured with Paretian approaches. The second way forward embraces this point. Consider libertarianism as an example. Suppose that one thinks that people are entitled to their pre-tax incomes and that in some way taxation is like theft. This view is not faithfully rendered as saying that additional income to people who have been taxed more should be given additional weight in comparison to those who have been given less; rather it is the view that it is wrong to tax, or at least, if not absolutely wrong, that it is bad to tax, and that this bad is tolerated, to the extent that it is, because of the other important purposes of taxation. On a rights-based version of libertarianism, taxing people is bad not because it reduces their utility but because it violates their entitlements. Imagine there is a function \( s(t_i) \) for each agent \( i \), that measures how bad it is to violate \( i \)'s entitlements. We might then minimize the non-Paretian social welfare function \( W(T) = -\int s(T_i(z_i(T))) \, di \) subject to a revenue requirement. Such an approach will not be Paretian, even in the sense of Proposition 2 and so it follows from Corollary 1 that this approach cannot be captured by welfare weights. Alternatively we may trade off rights based concerns as captured by \( s(t_i) \) against utilitarian concerns. Or we may want to go farther, and consider more thoroughly procedural approaches that do not appeal to a social objective (or even a local social objective). Whatever the right approach, it seems unlikely that we can capture the richness of broader ethical values by means of conservative modifications, such as by modifications of welfare weights, in a way that strongly preserves the structure of traditional optimal tax theory; we should expect that incorporation of broader values will require a more thorough change in the way that we normatively evaluate taxes and other economic policies.

References

Berg, K. and Piacquadio, P. G. (2020), ‘The equal-sacrifice social welfare function with an application to optimal income taxation’.

Bergson, A. (1938), ‘A reformulation of certain aspects of welfare economics’, The Quarterly Journal of Economics 52(2), 310–334.

\(^{18}\) Other work representing broader values with Paretian social welfare functions includes Fleurbaey and Maniquet (2011), Piacquadio (2017) and Berg and Piacquadio (2020).

\(^{19}\) Fleurbaey and Maniquet (2018) recognize this, writing “we highlight another way in which at least some fairness principles can remain compatible with the Pareto principle ... Not all fairness principles fall in this category, obviously, and the socialist and libertarian principles mentioned two paragraphs earlier provide examples of non-Paretian approaches.” (p. 1040) For criticisms of the Pareto principle, see Sen (1979b), Mongin (1997/2016) and Sher (2020).

\(^{20}\) For approaches to libertarian taxation, see Nozick (1974), Feldstein (1976), Young (1987), Weinzierl (2014), and Vallentyne (2018). For an approach to non-welfarist optimal taxation, see Kanbur, Pirttilä and Tuomala (2006).
Besley, T. and Coate, S. (1992), ‘Workfare versus welfare: Incentive arguments for work requirements in poverty-alleviation programs’, *The American Economic Review* 82(1), 249–261.

Feldstein, M. (1976), ‘On the theory of tax reform’, *Journal of public economics* 6(1-2), 77–104.

Fleurbaey, M. and Maniquet, F. (2011), *A theory of fairness and social welfare*, Vol. 48, Cambridge University Press.

Fleurbaey, M. and Maniquet, F. (2018), ‘Optimal income taxation theory and principles of fairness’, *Journal of Economic Literature* 56(3), 1029–79.

Fleurbaey, M., Tungodden, B. and Chang, H. F. (2003), ‘Any non-welfarist method of policy assessment violates the pareto principle: A comment’, *Journal of Political Economy* 111(6), 1382–1385.

Hartman, P. (1982), *Ordinary Differential Equations*, second edn, Berkhauser.

Kanbur, R., Keen, M. and Tuomala, M. (1994), ‘Optimal non-linear income taxation for the alleviation of income-poverty’, *European Economic Review* 38(8), 1613–1632.

Kanbur, R., Pirttilä, J. and Tuomala, M. (2006), ‘Non-welfarist optimal taxation and behavioural public economics’, *Journal of Economic Surveys* 20(5), 849–868.

Kaplow, L. and Shavell, S. (2001), ‘Any non-welfarist method of policy assessment violates the pareto principle’, *Journal of Political Economy* 109(2), 281–286.

Kaplow, L. and Shavell, S. (2009), *Fairness versus welfare*, Harvard university press.

Mirrlees, J. A. (1971), ‘An exploration in the theory of optimum income taxation’, *The review of economic studies* 38(2), 175–208.

Mongin, P. (1997/2016), ‘Spurious unanimity and the pareto principle’, *Economics & Philosophy* 32(3), 511–532.

Nozick, R. (1974), *Anarchy, state, and utopia*, Vol. 5038, New York: Basic Books.

Piacquadio, P. G. (2017), ‘A fairness justification of utilitarianism’, *Econometrica* 85(4), 1261–1276.

Saez, E. and Stantcheva, S. (2016), ‘Generalized social marginal welfare weights for optimal tax theory’, *American Economic Review* 106(1), 24–45.

Samuelson, P. A. (1947), *Foundations of economic analysis*, Harvard University Press.

Sen, A. (1970), ‘The impossibility of a paretian liberal’, *The Journal of Political Economy* 78(1), pp. 152–157.

Sen, A. (1979a), ‘Personal utilities and public judgements: or what’s wrong with welfare economics’, *The economic journal* 89(355), 537–558.
Sen, A. (1979b), ‘Utilitarianism and welfarism’, *The Journal of Philosophy* **76**(9), 463–489.

Sher, I. (2020), ‘How perspective-based aggregation undermines the pareto principle’, *Politics, Philosophy & Economics* **19**(2), 182–205.

Vallentyne, P. (2018), Libertarianism and taxation, *in* M. O’Neill and S. Orr, eds, ‘Taxation: philosophical perspectives’, Oxford University Press, pp. 98–110.

Weinzierl, M. (2014), ‘The promise of positive optimal taxation: normative diversity and a role for equal sacrifice’, *Journal of Public Economics* **118**, 128–142.

Weinzierl, M. (2017), ‘Popular acceptance of inequality due to innate brute luck and support for classical benefit-based taxation’, *Journal of Public Economics* **155**, 54–63.

Weymark, J. A. (2017), ‘Conundrums for nonconsequentialists’, *Social Choice and Welfare* **48**(2), 269–294.

Young, H. P. (1987), ‘Progressive taxation and the equal sacrifice principle’, *Journal of Public Economics* **32**(2), 203–214.
Appendix

A Proofs of results stated in main text

A.1 Proof of Proposition 1

Assume that $g$, $(T^{\theta})$ and $\theta_0$ are as in the hypothesis of the proposition. Now, first assume that $\int g_i(\theta_0) \frac{\partial}{\partial \theta_0} \bigg|_{\theta=\theta_0} T_i(z_i(\theta_0), \theta) \, d\theta < 0$. It follows from the smoothness of welfare weights, utility functions and parameterized families of tax policies that if $\theta_1$ is such that $\theta_1 > \theta_0$ and $\theta_1$ is sufficiently close to $\theta_0$, then for all $\theta' \in [\theta_0, \theta_1]$, $\int g_i(T^{\theta'}) \frac{\partial}{\partial \theta} \bigg|_{\theta=\theta'} T_i(z_i(T^{\theta'}), \theta) \, d\theta < 0$. This establishes the first claim of Proposition 1.

Next assume that $\int g_i(\theta_0) \frac{\partial}{\partial \theta_0} \bigg|_{\theta=\theta_0} T_i(z_i(\theta_0), \theta) \, d\theta > 0$. Now define the parameterized family of tax policies, $(\hat{T}^{\theta})_{\theta \in [-\overline{\theta}, -\overline{\theta}]}$ by $\hat{T}^{\theta} = T^{\theta}$, $\forall \theta \in [-\overline{\theta}, -\overline{\theta}]$, and, using notation analogous to that introduced in Section 3.1 let $\hat{T}_i(z, \theta) = \hat{T}^{\theta}_i(z)$. Then we have:

$$\int g_i(\hat{T}^{-\theta_0}) \frac{\partial}{\partial \theta} \bigg|_{\theta=-\theta_0} \hat{T}_i(z_i(\hat{T}^{-\theta_0}), \theta) \, d\theta = \int g_i(T^{\theta_0}) \times \left(- \frac{\partial}{\partial \theta} \bigg|_{\theta=\theta_0} T_i(z_i(T^{\theta_0}), \theta) \right) \, d\theta = -\int g_i(T^{\theta_0}) \frac{\partial}{\partial \theta} \bigg|_{\theta=\theta_0} T_i(z_i(T^{\theta_0}), \theta) \, d\theta < 0,$$

where the inequality follows from the assumption made at the beginning of the paragraph. It follows from the smoothness of welfare weights, utility functions and parameterized families of tax policies that if $-\theta_1 \in (-\overline{\theta}, -\overline{\theta})$ is sufficiently close to $-\theta_0$, then for all $\theta' \in [-\theta_1, -\theta_0]$, $\int g_i(T^{\theta'}) \frac{\partial}{\partial \theta} \bigg|_{\theta=\theta'} \hat{T}_i(z_i(T^{\theta'}), \theta) \, d\theta < 0$. So the global improvement principle implies that, for all $-\theta' \in (-\theta_1, -\theta_0)$, $\hat{T}^{-\theta'} \succ_R \hat{T}^{-\theta_0}$. So for all $\theta' \in (\theta_0, \theta_1), T^{\theta_0} \succ_R T^{\theta'}$. This establishes the second claim of Proposition 1. □

A.2 Proof of Proposition 2

First assume that all agents are indifferent as $\theta$ varies in the interval $[\theta_0, \theta_1]$. Then, for all $\theta' \in [\theta_0, \theta_1]$ and agents $i$, $\frac{\partial}{\partial \theta} \bigg|_{\theta=\theta'} U_i(T^{\theta'}) = 0$. Hence, by (6), for all $\theta' \in [\theta_0, \theta_1]$ and agents $i$, $\frac{\partial}{\partial \theta} \bigg|_{\theta=\theta'} T_i(z_i(T^{\theta'}), \theta) = 0$. So, for all $\theta' \in [\theta_0, \theta_1]$, $\int g_i(T) \frac{\partial}{\partial \theta} \bigg|_{\theta=\theta'} T_i(z_i(T^{\theta'}), \theta) \, d\theta = 0$. So by the global indifference principle (in Section 3.2), $T^{\theta_0} \sim_R T^{\theta_1}$. This establishes Pareto indifference along paths. Weak Pareto along paths is similar, appealing again to (6), and using the global improvement principle (also in Section 3.2) instead of the global indifference principle. □

A.3 Definitions for and Proof of Corollary 1

Consider a social welfare function $W : \mathcal{T} \to \mathbb{R}$ defined on tax policies. Say the social welfare function is well-behaved if for all $(T^{\theta})_{\theta \in \Theta}$ in $\mathcal{T}$ and $\theta_0 \in \Theta$, the derivative $\frac{d}{d\theta} \bigg|_{\theta=\theta_0} W(T^{\theta})$ exists.
Say that a social welfare function $W$ is **Paretian along paths** if for all $(T^\theta)_{\theta \in \Theta}$ in $T$ and all $\theta_0, \theta_1 \in \Theta$ with $\theta_0 < \theta_1$, $W$ satisfies the following properties:

1. **Pareto indifference along a path.** Suppose that all agents are indifferent among all tax policies $T^\theta$ for $\theta \in [\theta_0, \theta_1]$. Then $W(T^{\theta_0}) = W(T^{\theta_1})$.

2. **Weak Pareto along paths.** Suppose that, for all $\hat{\theta} \in [\theta_0, \theta_1]$ and all agents $i$, $\frac{d}{d\theta} U_i\left(\hat{\theta}\right) > 0$. Then $W(T^{\theta_0}) < W(T^{\theta_1})$.

Say that a system of welfare weights $g$ implements social welfare function $W$ if $W$ is well-behaved and for all $(T^\theta)_{\theta \in \Theta}$ in $T$ and all $\theta' \in \Theta$,

$$
\frac{d}{d\theta}\bigg|_{\theta=\theta'} W\left(T^\theta\right) > 0 \Leftrightarrow \int g_i\left(T^{\theta'}\right) \frac{\partial}{\partial \theta} T\left(z_i\left(T^{\theta'}\right), \theta\right) di < 0 \quad \text{and} \quad (A.1)
$$

$$
\frac{d}{d\theta}\bigg|_{\theta=\theta'} W\left(T^\theta\right) = 0 \Leftrightarrow \int g_i\left(T^{\theta'}\right) \frac{\partial}{\partial \theta} T\left(z_i\left(T^{\theta'}\right), \theta\right) di = 0. \quad \text{(A.2)}
$$

The first condition says that increasing $\theta$ is good according to the social welfare function $W$ and this is detected by the $\theta$-derivative of the expression $W\left(T^\theta\right)$ if and only if increasing $\theta$ is desirable according to welfare weights $g$. The second condition says that the $\theta$-derivative of does not detect any change in social welfare if and only if welfare weights do not detect any change in social welfare.

Having made the terms in the corollary precise, I now prove the corollary. Assume that the system of welfare weights $g$ implements social welfare function $W$. Let $(T^\theta)_{\theta \in \Theta}$ in $T$ and let $\theta_0, \theta_1 \in \Theta$ with $\theta_0 < \theta_1$, and suppose that all agents are indifferent among all tax policies $T^{\theta'}$ for $\theta' \in [\theta_0, \theta_1]$. Then arguing as in the proof of Proposition 2, it follows that $\int g_i\left(T^{\theta'}\right) \frac{\partial}{\partial \theta} T\left(z_i\left(T^{\theta'}\right), \theta\right) di = 0$. So by (A.2), $\frac{d}{d\theta}\bigg|_{\theta=\theta'} W\left(T^\theta\right) = 0$, for all $\theta' \in [\theta_0, \theta_1]$. So $W(T^{\theta_0}) = W(T^{\theta_1})$. So any social welfare function implemented by $g$ satisfies Pareto indifference along paths. The argument that any social welfare function $W$ implemented by welfare weights satisfies Weak Pareto along paths, proceeds similarly, using (A.1), in the place of (A.2), to derive $\frac{d}{d\theta}\bigg|_{\theta=\theta'} W\left(T^\theta\right) > 0$, for all $\theta' \in [\theta_0, \theta_1]$, and hence $W(T^{\theta_0}) < W(T^{\theta_1})$. □

### A.4 Proof of Proposition 3

It is convenient to prove a stronger version of Proposition 3 which adds a third equivalent condition – condition 2 in Proposition A.1 below – to conditions 1 and 2. Recall that we have assumed that $g_i(c_i, z_i)$ is a smooth function of $(c_i, z_i)$.

**Proposition A.1.** Let $g$ and $\hat{g}$ be related as in (17). Then the following conditions are equivalent:

1. $g$ is structurally utilitarian.

2. $\forall i \in I, \forall \hat{u}_i \in \mathbb{R}, \forall z_i, z'_i \in Z, \; \hat{g}_i(\hat{u}_i, z_i) = \hat{g}_i(\hat{u}_i, z'_i)$.

3. $\forall i \in I, \forall \hat{u}_i \in \mathbb{R}, \forall z_i \in Z, \frac{\partial}{\partial \hat{u}_i} \hat{g}_i(\hat{u}_i, z_i) = 0$. 

27
Proof. First I argue that condition 1 of the proposition implies condition 2. Assume that \( g \) is structurally utilitarian. Now choose \( i \in I, z_i, z_i' \in Z \), and \( \hat{u}_i \in \mathbb{R} \). Define \( c_i = \hat{u}_i + v_i(z_i) \) and \( c_i' = \hat{u}_i + v_i(z_i') \). Then observe that

\[
c_i - v_i(z_i) = \hat{u}_i = c_i' - v_i(z_i') . \tag{A.3}
\]

Then \( \hat{g}_i(\hat{u}_i, z_i) = g_i(c_i, z_i) = g_i(c_i', z_i') = \hat{g}_i(\hat{u}_i, z_i') \), where the first and last equalities follow from (10), and the middle equality follows from (A.3) and the assumption that \( g \) is structurally utilitarian. It follows that condition 2 of the proposition holds.

Next I argue that condition 2 implies condition 1. So assume condition 2. Choose \( i \in I, c_i, c_i' \in \mathbb{R}, z_i, z_i' \in Z \) such that \( \hat{u}_i = c_i - v_i(z_i) = c_i' - v_i(z_i') \). It follows that \( g_i(c_i, z_i) = \hat{g}_i(\hat{u}_i, z_i) = \hat{g}_i(\hat{u}_i, z_i') = g_i(c_i', z_i') \), where the first and last equalities follow from (10), and the middle equality follows from condition 2 of the proposition. This establishes condition 1.

Finally, consider the equivalence of conditions 2 and 3. First observe that our smoothness assumptions imply that condition 2 implies: \( \forall i \in I, \forall \hat{u}_i \in \mathbb{R}, \forall z_i \in Z, \frac{\partial}{\partial c_i} \hat{g}_i(\hat{u}_i, z_i) = 0 \). Going in the other direction, the equivalence now follows from the fundamental theorem of calculus.

### A.5 Proof of Theorem 1

First assume welfare weights arise from a generalized utilitarian social welfare function, meaning that they are of the form \( g_i(c_i, z_i) = F_i'(U_i(c_i, z_i)) \frac{\partial}{\partial c_i} U_i(c_i, z_i) \). These weights are structurally utilitarian because, if, for all \( c_i, c_i', z_i, z_i' \), if \( c_i - v_i(z_i) = c_i' - v_i(z_i') \), then \( U_i(c_i, z_i) = u(c_i - v_i(z_i)) = u(c_i' - v_i(z_i')) = U_i(c_i', z_i') \) and \( \frac{\partial}{\partial c_i} U_i(c_i, z_i) = u'(c_i - v_i(z_i)) = u'(c_i' - v_i(z_i')) = \frac{\partial}{\partial c_i} U_i(c_i', z_i') \). So if \( c_i - v_i(z_i) = c_i' - v_i(z_i') \), then \( g_i(c_i, z_i) = g_i(c_i', z_i') \).

Going in the other direction, by Proposition 3 structural utilitarianism is equivalent to the requirement that, holding fixed agent characteristics \((x_i, y_i)\), welfare weights are a function of \( \hat{u}_i = c_i - v_i(z_i) \), so that, assuming structural utilitarianism, we can write \( g_i(c_i, z_i) = g(c_i, z_i, x_i, y_i) = \hat{g}(\hat{u}_i, x_i, y_i) \). Define the function \( w_i(\hat{u}_i) = w(\hat{u}_i, x_i, y_i) \) by \( w_i(\hat{u}_i^0) = \int_0^{\hat{u}_i^0} \hat{g}_i(\hat{u}_i) d\hat{u}_i \). Now define the Function \( F : \mathbb{R} \times X \times Y \to \mathbb{R} \) by \( F(v_i, x_i, y_i) = w(u^{-1}(v_i), x_i, y_i) \), where \( u^{-1}(-) \) is the inverse of \( u(-) \). If \( x_i \) and \( y_i \) are discrete, the smoothness of \( w \) and \( u \) imply that \( F \) is smooth. If \( x_i \) and \( y_i \) are discrete, \( w \) is smooth in its first argument and hence \( F \) is smooth in \( v_i \). Let \( F_i(v_i) = F(v_i, x_i, y_i) = W_i(c_i, z_i) = W_i(c_i, z_i) = \int_0^{\hat{u}_i^0} \hat{g}_i(\hat{u}_i) d\hat{u}_i \). We have \( W_i(c_i, z_i) = F_i(U_i(c_i, z_i)) = w_i(u^{-1}(u(c_i - v_i(z_i)))) = w_i(c_i - v_i(z_i)) \). Note that, from the above, we have \( g_i(c_i, z_i) = g_i(c_i - v_i(z_i)) = g_i(c_i - v_i(z_i)) = \frac{\partial}{\partial c_i} W_i(c_i, z_i) = F_i'(U_i(c_i, z_i)) \frac{\partial}{\partial c_i} U_i(c_i, z_i) \). So the weights arise from a generalized utilitarian social welfare function.

### A.6 Proof of Corollary 2

Suppose that welfare weights \( g \) are structurally utilitarian. It follows from Theorem 1 that welfare weights are of the form \( g_i(c_i, z_i) = F_i'(U_i(c_i, z_i)) \frac{\partial}{\partial c_i} U_i(c_i, z_i) \) for \( F_i(u_i) = F(u_i, x_i, y_i) \) for some \( F \). So for the social welfare function \( W(T) = -\int F_i(U_i(c_i(T), z_i(T))) d\hat{u}_i \), the envelope theorem im-
plies that, for all $(T^θ)_{θ ∈ Θ}$ in T and $θ_0 ∈ Θ$, \( \frac{∂}{∂θ} \Big|_{θ=θ_0} W (T^θ) = - \int g_ι (T^{θ_0}) \frac{∂}{∂θ} \Big|_{θ=θ_0} T_i (z_i (T^{θ_0}), θ) \) di.

\[ \square \]

### A.7 Proof of Theorem 2

What follows is a more formal version of the argument in the main text. Assume that welfare weights are not structurally utilitarian. It follows from Proposition 3 that there exists $j ∈ I$, $u^* ∈ R$, $z^* ∈ Z$, such that $\frac{∂}{∂z_i^*} \hat{g}_j (u^*, z^*) ≠ 0$. Smoothness of the primitives implies that we can choose $z^*$ so that $z^* > 0$. Assume that $\frac{∂}{∂z_i^*} \hat{g}_j (u^*, z^*) < 0$. (The argument would be similar if we assumed instead that $\frac{∂}{∂z_i^*} \hat{g}_j (u^*, z^*) > 0$.) Our smoothness assumptions then imply that there exists a non-degenerate closed interval of agents $S$, which is a proper subset of $[0, 1]$, such that, for all agents $i ∈ S$, $\frac{∂}{∂z_i^*} \hat{g}_i (u^*, z^*) < 0$.

Let $O$ and $Q$ be two other non-degenerate closed intervals contained in $[0, 1]$, such that $S, O,$ and $Q$ are pairwise disjoint. Now consider a doubly parameterized family of tax policies $(T^{θ, ε})_{θ ∈ Θ, ε ∈ E'}$, where $Θ = [\underline{θ}, \overline{θ}]$ for some $\underline{θ} < \overline{θ}$ and $E = [-\epsilon, \epsilon]$ for some $\epsilon > 0$, and which takes the following form:

\[
T^{θ, ε}_i (z_i) = \begin{cases} 
τ_i (θ) z_i + κ_i (θ) + ε t_S, & \text{if } i ∈ S, \\
-ε t_O, & \text{if } i ∈ O, \\
\tilde{τ} (θ, ε) z_i + \tilde{κ}_i (θ, ε), & \text{if } i ∈ Q, \\
0, & \text{otherwise.}
\end{cases}
\]  

(A.4)

Above $τ_i (θ)$ is a personalized marginal tax rate for agents in $i ∈ S$, and $\tilde{τ} (θ, ε)$ is a marginal tax rate which is not personalized on $Q$; both $τ_i (θ)$ and $\tilde{τ} (θ, ε)$ depend on parameter values. $κ_i (θ)$ and $\tilde{κ}_i (θ, ε)$ are personalized lumpsum taxes that depend on parameters. $t_S$ and $t_O$ are positive real numbers, so that $t_S$ and $-t_O$ are lumpsum taxes as well. I assume that the map $(i, θ) ↦ τ_i (θ)$ is smooth on the domain $S × Θ$ and that the map $(θ, ε) ↦ \tilde{τ} (θ, ε)$ is smooth on the domain $Θ × E$.

Moreover, I assume that there exists $θ_0 ∈ (\underline{θ}, \overline{θ})$ such that, for all $i ∈ Θ$, $τ_i (θ_0) = 1 - v^*_i (z^*)$ and, for all $θ ∈ Θ$, $τ_i' (θ) > 0$.

In what follows let $z_i (θ, ε)$ denote $i$’s optimal income in response to tax policy $T^{θ, ε}$ and let $\hat{U}_i (θ, ε) = \hat{U}_i (T^{θ, ε}) = z_i (θ, ε) - T^{θ, ε}_i (z_i (θ, ε)) - v_i (z_i (θ, ε))$ be $i$’s utility in response to $T^{θ, ε}$, using the representation that omits the outer utility function $u (⋅)$, and $g_i (θ, ε) = g_i (T^{θ, ε}) = \hat{g}_i (\hat{U}_i (θ, ε), z_i (θ, ε))$ be $i$’s welfare weight at $T^{θ, ε}$. When an agent $i$ in $S$ faces tax policy $T^{θ_0, 0}$, they will solve the problem $\max z_i (1 - τ_i (θ_0)) z_i - κ_i (θ_0) - v_i (z_i)$. It follows from the construction of $τ_i (θ_0)$ and the fact that $v_i (z_i)$ is strictly convex that $z_i = z^*$ uniquely satisfies the agent’s first order condition when $(θ, ε) = (θ_0, 0)$, namely, $(1 - τ (θ_0)) - v^*_i (z_i) = 0$. Because agents’ objective is strictly concave, it follows that $z_i = z^*$ is the unique optimum for all agents $i ∈ S$ when facing tax policy $T^{θ_0, 0}$, so that $z_i (θ_0, 0) = z^*$ for all $i ∈ S$. For all $i ∈ S$, define the function $κ_i (θ)$

\[ 21 \]By a non-degenerate closed interval, I mean a closed interval which is not equal to a single point.

\[ 22 \]Of course, it is possible that $\frac{∂}{∂z_i} \hat{g}_i (u^*, z^*) < 0$ for all $i ∈ [0, 1]$, but in this case there is also a closed interval $S$, which is a proper subset of $[0, 1]$, on which this property holds.

\[ 23 \]We allow for the possibility that $τ_i (θ_0) < 0$. 

29
in (A.4) to solve:

\[(1 - \tau_i(\theta)) z_i(\theta, 0) - v_i(z_i(\theta, 0)) - \kappa_i(\theta) = \hat{u}^*, \quad \forall \theta \in \Theta.\]

That is, the lumpsum tax $\kappa_i(\theta)$ is chosen so as the keep the agents’ (in $S$) utility fixed at $\hat{u}^*$ when the agent faces tax policies of the form $T^{q_0}$ as $\theta$ changes – where we measure utility via the representation $\hat{U}_i(T^{q_0})$ that excludes the outer utility function $u(\cdot)$. Note that we can freely define $\kappa_i(\theta)$ in this way because the optimal income $z_i(\theta, 0)$ depends only on the marginal tax rate $\tau_i(\theta)$ and not on the lumpsum tax $\kappa_i(\theta)$. Note, moreover, that, for any $\epsilon \in E$, $\theta \in \Theta$, and $i \in S$, $i$'s utility, when facing $T^{q,\epsilon}$, is $\hat{U}_i(\theta, \epsilon) = \hat{u}^* - \epsilon t_S$, which does not depend on $\theta$. So, holding $\epsilon$ fixed, each agent $i \in S$ is indifferent as $\theta$ varies. Likewise, for all $i \in Q$, define $\kappa_i(\theta, \epsilon)$ to satisfy the following equation:

\[\left(1 - \bar{\tau}(\theta, \epsilon)\right) z_i(\theta, \epsilon) - v_i(z_i(\theta, \epsilon)) - \kappa_i(\theta, \epsilon) = 0, \quad \forall \theta \in \Theta, \forall \epsilon \in E. \quad \text{(A.5)}\]

That is, the lumpsum tax $\kappa_i(\theta, \epsilon)$ is selected to keep the utility $\hat{U}_i(\theta, \epsilon)$ of all agents $i \in Q$ equal to zero as $\theta$ and $\epsilon$ vary. Again, observe that $z_i(\theta, \epsilon)$ only depends on the marginal tax rate $\bar{\tau}(\theta, \epsilon)$ and not on the lumpsum tax $\kappa_i(\theta, \epsilon)$. Given the above, it follows by construction that, holding $\epsilon$ fixed, all agents are indifferent, as $\theta$ varies in $T^{q,\epsilon}$. So, it follows from part 1 of Proposition 2 – Pareto indifference along paths – that

\[T^{q_0,\epsilon} \sim g T^{q_1,\epsilon}, \quad \forall \epsilon \in E, \quad \text{(A.6)}\]

where $\theta_1$, satisfying $\theta_0 < \theta_1$, is a value of $\theta$ that we now select. In particular, it follows from the facts that $\frac{\partial}{\partial z_i} \hat{g}_i(z^*, \hat{u}^*) < 0$ and $z_i(\theta_0, 0) = z^*$ for all $i \in S$ and the smoothness of the primitives of the model that if we choose $\theta_1$ sufficiently close to $\theta_0$,

\[\frac{\partial}{\partial z_i} \hat{g}_i(\hat{U}_i(\theta, 0), z_i(\theta, 0)) = \frac{\partial}{\partial z_i} \hat{g}_i(\hat{u}^*, z_i(\theta, 0)) < 0, \quad \forall \theta \in [\theta_0, \theta_1], \forall i \in S. \quad \text{(A.7)}\]

So let us choose $\theta_1$ so that (A.7) is satisfied. Moreover, since $z_i(\theta_0, 0) = z^* > 0, \forall i \in S$, we may assume that $\theta_1$ is chosen sufficiently close to $\theta_0$ that, for all $i \in S$ and $\theta \in [\theta_0, \theta_1], z_i(\theta, 0) > 0$.

For any $\theta \in \Theta$, define $g_S(\theta, 0) = \int_S g_i(\theta, 0) \, d\xi$ and $g_O(\theta, 0) = \int_O g_i(\theta, 0) \, d\xi$. It follows from the fact that $\hat{U}_i(\theta, 0) = \hat{u}^*, \forall \theta \in \Theta, \forall i \in S$, (A.7), and the assumption that $\tau^*_i(\theta) > 0, \forall \theta \in \Theta, \forall i \in S$, which, given that $z_i(\theta, 0) > 0, \forall \theta \in [\theta_0, \theta_1], \forall i \in S$, implies that $\frac{\partial}{\partial \theta} g_S(\theta, 0) < 0, \forall \theta \in [\theta_0, \theta_1], \forall i \in S$, that

\[\frac{\partial}{\partial \theta} g_S(\theta, 0) > 0, \quad \forall \theta \in [\theta_0, \theta_1]. \quad \text{(A.8)}\]

Choose $\theta' \in (\theta_0, \theta_1)$ and suppose that the positive numbers $t_S$ and $t_O$ in (A.4) were selected to
satisfy
\[ g_S (\theta', 0) t_S = g_O (\theta', 0) t_O. \] (A.9)

Then, writing \( T (z_i, \theta, \epsilon) = T^{\theta, \epsilon} (z_i) \), we have:
\[
\int g_i (\theta_0, 0) \left. \frac{\partial}{\partial \epsilon} \right|_{\epsilon=0} T_i (z_i (\theta_0, 0), \theta_0, \epsilon) \, d\epsilon
= g_S (\theta_0, 0) t_S - g_O (\theta_0, 0) t_O + \int_Q \left[ \left. \frac{\partial}{\partial \epsilon} \right|_{\epsilon=0} \bar{\tau} (\theta_0, \epsilon) \right] z_i (\theta_0, 0) + \left. \frac{\partial}{\partial \epsilon} \right|_{\epsilon=0} \bar{\kappa}_i (\theta_0, \epsilon) \, d\epsilon
= g_S (\theta_0, 0) t_S - g_O (\theta_0, 0) t_O + \int_Q - \left. \frac{\partial}{\partial \epsilon} \right|_{\epsilon=0} \hat{U}_i (\theta_0, \epsilon) \, d\epsilon
= g_S (\theta_0, 0) t_S - g_O (\theta_0, 0) t_O < 0,
\] (A.10)

where the second equality follows from the envelope theorem, and the third equality follows from the fact that, by (A.5), the utility of all agents in \( Q \) is held fixed as \( \epsilon \) varies in \( T^{\theta_0, \epsilon} \), so that, for all \( i \in Q \), \( \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} \hat{U}_i (\theta_0, \epsilon) = 0 \). The inequality follows from (A.9), and the facts that \( g_O (\theta, 0) \) is constant in \( \theta \), that, by (A.8), \( g_S (\theta, 0) \) is increasing in \( \theta \), and that \( \theta_0 < \theta' \). Using similar arguments,
\[
\int g_i (\theta_1, 0) \left. \frac{\partial}{\partial \epsilon} \right|_{\epsilon=0} T_i (z_i (\theta_1, 0), \theta_1, \epsilon) \, d\epsilon = g_S (\theta_1, 0) t_S - g_O (\theta_1, 0) t_O > 0,
\] (A.11)

The reason that the the inequality in (A.11) points in the opposite direction of the inequality in (A.10) is that, whereas \( \theta_0 < \theta' \), \( \theta_1 > \theta' \). It follows from (A.10), (A.11), and the local improvement principle – Proposition 11 – that
\[
T^{\theta_0, 0} \prec^g T^{\theta_0, \epsilon}, \quad T^{\theta_1, 0} \prec^g T^{\theta_1, \epsilon}, \quad \text{for sufficiently small } \epsilon > 0.
\] (A.12)

Putting (A.6) and (A.12), together, we have that for sufficiently small \( \epsilon > 0 \),
\[
T^{\theta_0, 0} \prec^g T^{\theta_0, \epsilon} \sim^g T^{\theta_1, \epsilon} \prec^g T^{\theta_1, 0} \sim^g T^{\theta_0, 0}.
\] (A.13)

So, on the assumption that welfare weights are not structurally utilitarian, we have constructed a social preference cycle.

The last step is to show that revenue can be held fixed across the tax policies in the cycle. This is achieved via the selection of \( \bar{\tau} (\theta, \epsilon) \) in (A.4). For any marginal tax rate \( \tau \), write \( z_i (\tau) \) to be the income that \( i \) would earn, if \( i \) faces the tax policy \( T (z) = \tau z \), or, in other words, if \( i \) faces a constant marginal tax rate of \( \tau \). It follows that, for all \( i \in Q \), we can write \( z_i (\bar{\tau} (\theta, \epsilon)) = z_i (\theta, \epsilon) \) because every agent \( i \in Q \) faces the constant marginal tax rate \( \bar{\tau} (\theta, \epsilon) \) under tax policy \( T^{\theta, \epsilon} \). Let
\[ R_Q(\theta, \epsilon) \] be the revenue raised from agents in \( Q \) by tax policy \( T^{\theta, \epsilon} \). Then we have
\[
R_Q(\theta, \epsilon) = \int_Q T(z_i(\theta, \epsilon), \theta, \epsilon) \, d\theta = \int_Q [\check{\tau}(\theta, \epsilon) z_i(\theta, \epsilon) + \check{\kappa}_i(\theta, \epsilon)] \, d\theta \\
= \int_Q [\check{\tau}(\theta, \epsilon) z_i(\theta, \epsilon) + (1 - \check{\tau}(\theta, \epsilon)) z_i(\theta, \epsilon) - v_i(z_i(\theta, \epsilon))] \, d\theta \\
= \int_Q [z_i(\theta, \epsilon) - v_i(z_i(\theta, \epsilon))] \, d\theta,
\]
where the third equality follows from \((A.5)\). Next, for any marginal tax rate \( \tau \), define \( \check{R}_Q(\tau) \) by
\[
\check{R}_Q(\tau) = \int_Q [z_i(\tau) - v_i(z_i(\tau))] \, d\theta.
\]
Then it follows from \((A.14)\) and the fact that \( z_i(\check{\tau}(\theta, \epsilon)) = z_i(\theta, \epsilon) \) that \( \check{R}_Q(\check{\tau}(\theta, \epsilon)) = R_Q(\theta, \epsilon) \).

Since we assume that, in the absence of taxes, all agents earn positive income (see Section 2), there exists a positive marginal tax rate \( \tau_0 \), which is sufficiently small that, for all \( i \in Q \), \( z_i(\tau_0) > 0 \).

From agent \( i \)'s first order condition, when facing marginal tax rate \( \tau_0 \), we have that, for all \( i \in Q \),
\[
0 = (1 - \tau_0) - v_i'(z_i(\tau_0)) < 1 - v_i'(z_i(\tau_0)).
\]
Assume that \( \check{\tau}(\theta_0, 0) = \tau_0 \). Define \( R_{-Q}(\epsilon, \theta) = \int_{i \in Q \setminus Q} T_i(z_i(\theta, \epsilon), \theta, \epsilon) \, d\theta \) to be the revenue raised by tax policy \( T^{\theta, \epsilon} \) from all agents not in \( Q \). Now consider the condition
\[
\check{R}_Q(\check{\tau}(\theta, \epsilon)) + R_{-Q}(\theta, \epsilon) = \check{R}_Q(\tau_0) + R_{-Q}(\theta_0, 0).
\]
Observe that \( \check{R}_Q'(\check{\tau}(\theta, \epsilon)) = \int_Q z_i'(\tau_0) [1 - v_i'(z_i(\tau_0))] \, d\theta < 0 \).

It follows from the implicit function theorem that the function \( \check{\tau}(\theta, \epsilon) \) is uniquely determined in a neighborhood of \((\theta_0, 0)\) by \( \check{\tau}(\theta_0, 0) = \tau_0 \) and \((A.15)\). Redefining \( \check{\epsilon} \) to be sufficiently small and \( \check{\theta} \) and \( \check{\theta} \) to be sufficiently close to \( \theta_0 \) if necessary, and assuming that \( \theta_1 \) was chosen sufficiently close to \( \theta_0 \) so that \( \theta_0 < \theta_1 < \check{\theta} \) still holds, we may assume that we have thus defined \( \check{\tau}(\theta, \epsilon) \) on all of \( \Theta \times E \). Note now that \((A.15)\) implies that the revenue of \( T^{\theta, \epsilon} \) is held constant as \( \theta \) and \( \epsilon \) vary. This completes the proof. \( \square \)

A.8 Proof of Corollary 3

Assume that \( g \) is not structurally utilitarian. It follows from Proposition 3 that there exists an agent \( j \in (0, 1) \), \( z^* \in Z \) with \( z^* > 0 \) and \( \tilde{u}^* \in \mathbb{R} \) and such that
\[
\frac{\partial}{\partial z_j} g_j(\tilde{u}^*, z^*) \neq 0.
\]
\[ \tag{A.16} \]

\[24\] The assumption that, in the absence of taxes, all agents earn positive income, is not necessary for the proof. In the absence of this assumption, we could instead select \( \tau_0 \) to be a sufficiently small negative marginal tax rate that, for all \( i \in Q \), \( z_i(\tau_0) > 0 \). Then the proof would proceed in the same way as below except that \( \check{R}_Q'(\check{\tau}(\theta, \epsilon)) > 0 \) rather than \( \check{R}_Q'(\check{\tau}(\theta, \epsilon)) < 0 \). However what matters for the argument is only that \( \check{R}_Q'(\check{\tau}(\theta, \epsilon)) \neq 0 \).

\[25\] This inequality follows from the facts that, by our assumptions above imply that, for all \( i \in Q \) (i) \( z_i(\tau_0) > 0 \), so that \( z_i'(\tau_0) < 0 \), and that (ii) \( 1 - v_i'(z_i(\tau_0)) > 0 \).
We can assume that \( j \) is in the interior of \( I = [0,1] \) and \( z^* > 0 \) because of the smoothness of the primitives. Choose a smooth strictly convex tax policy \( T \) such that (i) \( T'(z^*) = 1 - v'_j(z^*) \), (ii) \( T'(0) \) is sufficiently small (or negative if \( 1 - v'_j(z^*) < 0 \)) such that all agents would earn a positive income in response to \( T - \) recall that in the absence of taxes, all agents earn a positive income (see Section 2), and (iii) \( \lim_{z \to \infty} T'(z) > 0 \). These assumptions, together with the strict convexity of \( v_i(z_i) \) and the assumption that \( v'_i(z_i) > 1 \) for sufficiently large \( z_i \) (see Section 2), imply that \( T \in \hat{\Theta} \). (See Section 6.1 for the requirements on \( \hat{\Theta} \).) It follows from property (i) that \( z_j(T) = z^* \).

By the appropriate choice of a lumpsum transfer in \( T \), we can ensure that \( \hat{U}_j(T) = \hat{u}^* \). (A.16) together with the smoothness of the primitives and of \( T \) now ensure that if we select a sufficiently small interval \((i_a, i_b)\) containing \( j \), then either (28) or (29) holds. □

### A.9 Omitted details from the proof of Lemma 2

Here I present the details of the proof of Lemma 2 that were omitted in the main text: the expression for the overlapping term \( C \) discussed in the text, and the proof of conditions (31)-(32). First, I present the expression for the term \( C \), which I will prove is the overlapping term below:

\[
C = \int \left( -\frac{\partial}{\partial t_i} \hat{g}_i \left( \hat{U}_i \left( \theta_0, \epsilon_0 \right), z_i \left( \theta_0, \epsilon_0 \right) \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} T \left( z_i \left( \theta_0, \epsilon_0 \right), \theta_0, \epsilon \right) \frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} T \left( z_i \left( \theta_0, \epsilon_0 \right), \theta, \epsilon_0 \right) \right. \\
+ \left. g_i \left( \theta_0, \epsilon_0 \right) \left[ -\frac{\partial^2}{\partial \theta \partial z_i} \bigg|_{\theta = \theta_0, z_i = z_i(\theta_0, \epsilon_0)} T \left( z_i, \theta, \epsilon_0 \right) + \frac{\partial^2}{\partial z_i^2} \bigg|_{z_i = z_i(\theta_0, \epsilon_0)} T \left( z_i, \theta_0, \epsilon_0 \right) \right] v''_i \left( z_i \left( \theta_0, \epsilon_0 \right) \right) \right) \, di.
\]

(A.17)

Next, I present some useful preliminary facts, which I use to establish (31)-(32). Observe that at \((\theta_0, \epsilon_0)\), agent \( i \)'s optimization problem is: \( \max_{z_i} \left[ z_i - v_i(z_i) - T \left( z_i, \theta_0, \epsilon_0 \right) \right] \). The first-order condition is: \( 1 - v'_i(z_i) - \frac{\partial}{\partial z_i} T \left( z_i, \theta_0, \epsilon_0 \right) = 0 \). Applying the implicit function theorem to the first-order condition, we have:

\[
\frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} z_i \left( \theta, \epsilon_0 \right) = -\frac{\partial^2}{\partial \theta \partial z_i} \bigg|_{\theta = \theta_0, z_i = z_i(\theta, \epsilon_0)} T \left( z_i, \theta, \epsilon_0 \right)
\]

(A.18)

\[
\frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} z_i \left( \theta_0, \epsilon \right) = -\frac{\partial^2}{\partial \epsilon \partial z_i} \bigg|_{\epsilon = \epsilon_0, z_i = z_i(\theta_0, \epsilon_0)} T \left( z_i, \theta_0, \epsilon \right)
\]

(A.19)

\[\text{As} \ (T^{\theta, \epsilon})_{\theta \in \Theta, \epsilon \in E} \in T_2, \text{ it follows that, for all} \ \theta \in \Theta \text{ and } \epsilon \in E, \ T^{\theta, \epsilon} \in \hat{\Theta}, \text{ which implies that the first-order condition uniquely characterizes agent } i \text{'s optimal income } z_i(T).\]
I am now ready to establish (31)-(32). First, I establish (31): 

\[
\frac{d}{d\epsilon} \bigg|_{\epsilon=0} \int g_i(\theta_0, \epsilon) \frac{\partial}{\partial \theta} \bigg| \frac{T(z_i(\theta_0, \epsilon), \theta, \epsilon)}{\theta=\theta_0} \, d\epsilon
\]

\[
= \int \bigg( \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} \left[ g_i(\theta_0, \epsilon) \frac{\partial}{\partial \theta} \bigg| \frac{T(z_i(\theta_0, \epsilon), \theta, \epsilon)}{\theta=\theta_0} \right] \bigg) \, d\epsilon \tag{A.20}
\]

\[
= \int \bigg( \left[ \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} g_i(\theta_0, \epsilon) \frac{\partial}{\partial \theta} \bigg| \frac{T(z_i(\theta_0, \epsilon_0), \theta, \epsilon_0)}{\theta=\theta_0} \right] \bigg) \, d\epsilon \tag{A.21}
\]

\[
+ \frac{d}{d\epsilon} \bigg|_{\epsilon=0} \frac{\partial}{\partial \theta} \bigg| \frac{T(z_i(\theta_0, \epsilon), \theta, \epsilon)}{\theta=\theta_0} \bigg) \bigg) \, d\epsilon \tag{A.22}
\]

\[
= \int \bigg( \left[ \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} g_i(\theta_0, \epsilon) \frac{\partial}{\partial \theta} \bigg| \frac{T(z_i(\theta_0, \epsilon_0), \theta, \epsilon_0)}{\theta=\theta_0} \right] \bigg) \, d\epsilon \tag{A.23}
\]

\[
= \int \bigg( \left[ \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} g_i(\theta_0, \epsilon) \frac{\partial}{\partial \theta} \bigg| \frac{T(z_i(\theta_0, \epsilon_0), \theta, \epsilon_0)}{\theta=\theta_0} \right] \bigg) \, d\epsilon \tag{A.24}
\]

\[
= A + C \tag{A.25}
\]
where (A.22) analyzes the term \( \frac{\partial}{\partial \epsilon} |_{\epsilon=\epsilon_0} g_i (\theta_0, \epsilon) \) and appeals to the fact that, by the envelope theorem, \( \frac{\partial}{\partial \epsilon} |_{\epsilon=\epsilon_0} \hat{U}_i (\theta_0, \epsilon) = - \frac{\partial}{\partial \epsilon} |_{\epsilon=\epsilon_0} T (z_i (\theta_0, \epsilon), \theta, \epsilon) \), and (A.24) follows from (A.19). A is defined by (33) and C is defined by (A.17). This establishes (31).

As the derivation of (32) is similar, I present it in an abbreviated form:

\[
\frac{d}{d\theta} \bigg|_{\theta=\theta_0} \int g_i (\theta, \epsilon_0) \frac{\partial}{\partial \epsilon} |_{\epsilon=\epsilon_0} T (z_i (\theta, \epsilon_0), \theta, \epsilon) \, di
\]

\[
= \int \left( \left[ - \frac{\partial}{\partial \hat{u}_i} \hat{g}_i \left( \hat{U}_i (\theta_0, \epsilon_0), z_i (\theta_0, \epsilon_0) \right) \frac{\partial}{\partial \theta} \bigg|_{\theta=\theta_0} T (z_i (\theta_0, \epsilon_0), \theta, \epsilon_0) \right.
\]

\[
+ \frac{\partial}{\partial z_i} \hat{g}_i \left( \hat{U}_i (\theta_0, \epsilon_0), z_i (\theta_0, \epsilon_0) \right) \frac{\partial}{\partial \theta} \bigg|_{\theta=\theta_0} z_i (\theta, \epsilon_0) \bigg] \frac{\partial}{\partial \epsilon} |_{\epsilon=\epsilon_0} T (z_i (\theta_0, \epsilon_0), \theta, \epsilon)
\]

\[
+ g_i (\theta_0, \epsilon_0) \left[ \frac{\partial^2}{\partial z_i \partial \epsilon} \bigg|_{z_i=z_i(\theta_0,\epsilon_0),\epsilon=\epsilon_0} T (z_i, \theta_0, \epsilon) \frac{\partial}{\partial \theta} \bigg|_{\theta=\theta_0} z_i (\theta, \epsilon_0) \right]
\]

\[
+ \frac{\partial^2}{\partial \theta \partial \epsilon} \bigg|_{\theta=\theta_0, \epsilon=\epsilon_0} T (z_i (\theta_0, \epsilon_0), \theta, \epsilon) \right) \bigg] \, di
\]

\[
= B + C.
\]

The justification is similar to the justification for (A.20)-(A.25), using (A.18) instead of (A.19). This establishes (32). □

## A.10 Proof of Lemma 3

The main argument proving Lemma 3 is presented in Section A.10.1. The proofs of a supporting lemma and some related material are presented in the subsequent subsections.

### A.10.1 Main argument

Choose \( T \in \mathcal{F} \). To establish the lemma, I construct a doubly parameterized family of tax policies \( (T^\theta, \epsilon)_{\theta \in \Theta, \epsilon \in E} \in T_2 \) satisfying (22), (23), and (35), and such that, for the \( \theta_0 \in (\hat{\theta}, \overline{\theta}) \), \( \epsilon_0 \in (\hat{\epsilon}, \overline{\epsilon}) \) that feature in the preceding conditions, \( T^{\theta_0, \epsilon_0} = T \).
Recall that the support of a function \( h \) with argument \( x \) is the closure of \( \{ x : h(x) \neq 0 \} \).

To construct \( (T^{\theta,\epsilon})_{\theta,\epsilon \in \Theta \times E} \), I consider four smooth tax reforms \( \mu_1, \mu_2, \eta_1, \eta_2 \). Let \( i_k, k = 1, 2, 3, 5 \) be elements of \((0, 1)\) be such that \( i_1 < i_2 < i_3 = i_4 < i_5 = i_6 \). The reader will notice that we have skipped \( i_4 \); this term will be introduced below (see Lemma A.1). If we let \( \zeta_k = z_{i_k}(T) \) for \( k = 1, 2, 3, 5 \), it follows from assumptions on \( v \) and \( i \mapsto y_i \) in Section 6.1 that \( \zeta_1 < \zeta_2 < \zeta_3 < \zeta_5 \).

I assume that \( \mu_1(z) = 0 \) when \( z \leq \zeta_3 \), \( \mu_1(z) \) is increasing in \( z \) on the interval \((\zeta_3, \zeta_5)\), and \( \mu_1(z) \) remains constant at some positive number thereafter. I assume that \( \mu_2(z) = 0 \) when \( z \leq \zeta_2 \), \( \mu_2(z) \) is increasing in \( z \) on the interval \((\zeta_2, \zeta_3)\) and \( \mu_2(z) = 1 \) when \( z \geq \zeta_3 \). Assume, moreover, that \( \mu_1 \) and \( \mu_2 \) are chosen such that:

\[
\int_0^1 g_i(T) \mu_1(z_i(T)) \, di = \int_0^1 g_i(T) \mu_2(z_i(T)) \, di. \tag{A.26}
\]

That is, both tax reforms \( \mu_1 \) and \( \mu_2 \) have the same marginal effect on social welfare, when benefits are weighted by welfare weights. The above assumptions imply the following lemma, which is proved in Section A.10.2 of the Appendix.

**Lemma A.1.** There exists \( i_4 \in (i_3, i_5) \) such that \( \mu_1(z_{i_4}(T)) = 1 \).

If we define \( \zeta_4 = z_{i_4}(T) \), it follows from the fact that \( i_3 < i_4 < i_5 \) that \( \zeta_3 < \zeta_4 < \zeta_5 \). So Lemma A.1 says that there is some income level \( \zeta_4 \), between \( \zeta_3 \) and \( \zeta_5 \), such that \( \mu_1(\zeta_4) = 1 \), and moreover income level \( \zeta_4 \) is chosen by some agent \( i_4 \) when facing tax policy \( T \).

Assume that \( \eta_1 \) has support \([\zeta_3, \zeta_5]\), and that \( \eta_1 \) is increasing on \((\zeta_3, \zeta_4)\) and decreasing on \((\zeta_4, \zeta_5)\), which implies that \( \eta_1(z) > 0, \forall z \in (\zeta_3, \zeta_5) \). Assume that the support of \( \eta_2 \) is \([\zeta_1, \zeta_2]\), that \( \eta_2(z) < 0, \forall z \in (\zeta_1, \zeta_2) \), and that

\[
\int_0^1 g_i(T) \eta_1(z_i(T)) \, di = - \int_0^1 g_i(T) \eta_2(z_i(T)) \, di. \tag{A.27}
\]

In other words the marginal welfare effect of reform \( \eta_1 \) is the negative of the marginal welfare effect of reform \( \eta_2 \), so that the two cancel out.

For any real numbers, \( \theta \) and \( \epsilon \), define \( \hat{T}^{\theta,\epsilon} \) by:

\[
\hat{T}^{\theta,\epsilon} = T + \theta \mu_1 + \epsilon (\eta_1 + \eta_2). \tag{A.28}
\]

It follows from the Picard-Lindelöf theorem and related results (see Section A.10.3) that there exist real numbers \( \bar{\theta}, \bar{\eta}, \xi, \tau \) such that \( \bar{\theta} < 0 < \bar{\eta}, \xi < 0 < \tau \), and such that we can define the real-valued function \( \zeta(\theta, \epsilon) \) on \( \Theta \times E \), where \( \Theta = [\bar{\theta}, \bar{\eta}] \) and \( E = [\xi, \tau] \), by

\[
\zeta(0, \epsilon) = 0, \quad \forall \epsilon \in E, \tag{A.29}
\]
\[
\int g_i \left( \bar{T}^{\theta, \epsilon} - \zeta (\theta, \epsilon) \mu_2 \right) \\
\times \left[ \mu_1 \left( z_i \left( \bar{T}^{\theta, \epsilon} - \zeta (\theta, \epsilon) \mu_2 \right) \right) - \frac{\partial}{\partial \theta} \zeta (\theta, \epsilon) \mu_2 \left( z_i \left( \bar{T}^{\theta, \epsilon} - \zeta (\theta, \epsilon) \mu_2 \right) \right) \right] \, di = 0, \tag{A.30}
\]
\forall \theta \in \Theta, \forall \epsilon \in E.

Next, for all \( \theta \in \Theta \) and \( \epsilon \in E \), define
\[
T^{\theta, \epsilon} = T + [\theta \times \mu_1] - [\zeta (\theta, \epsilon) \times \mu_2] + [\epsilon \times (\eta_1 + \eta_2)]
= \bar{T}^{\theta, \epsilon} - \zeta (\theta, \epsilon) \mu_2. \tag{A.31}
\]

It is straightforward to establish that if \( \theta, \bar{\theta}, \zeta, \) and \( r \) are all chosen sufficiently close to 0, then for all \( \theta \in [\bar{\theta}, \bar{\theta}], \epsilon \in [\bar{\epsilon}, \bar{\epsilon}] \), \( T^{\theta, \epsilon} \in \hat{\mathcal{T}} \) (see Section A.10.3). Moreover, by Corollary 4.1 on p. 101 of Hartman (1982), \( (\theta, \epsilon) \mapsto \zeta (\theta, \epsilon) \) is smooth, and hence also \( (z, \theta, \epsilon) \mapsto T (z, \theta, \epsilon) \) is smooth, so that \( T^{\theta, \epsilon} \) satisfies (23).

Next I seek to establish that \( T^{\theta, \epsilon} \) satisfies (23). In the special case in which \( \theta = \theta_0 \) and \( \epsilon = \epsilon_0 \) (recall that \( \theta_0 = \epsilon_0 = 0 \)), the general statement in (A.30) reduces to
\[
\int g_i (T) \left[ \mu_1 (z_i (T)) - \frac{\partial}{\partial \theta} \zeta (\theta, \epsilon_0) \mu_2 (z_i (T)) \right] \, di = 0.
\]
Solving for \( \frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} \zeta (\theta, \epsilon_0) \) from the above equation, it follows that

\[
\frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} \zeta (\theta, \epsilon_0) = \frac{\int g_i (T) \mu_1 (z_i (T)) \, di}{\int g_i (T) \mu_2 (z_i (T)) \, di} = 1, \tag{A.35}
\]

where the second equality follows from (A.26).

Consider the type \( i \) agent’s optimization problem when facing tax policy \( T^{\theta, \epsilon} \)–that is, of choosing \( z \) so as to maximize \( z - v_i (z) - T^{\theta, \epsilon} (z) \). It follows from the implicit function theorem applied to the first order condition for this optimization problem at \((\theta, \epsilon) = (\theta_0, \epsilon_0)\) that

\[
\frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} z_i (\theta, \epsilon_0) = -\frac{\mu_1' (z_i (T)) - \frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} \zeta (\theta, \epsilon_0) \mu_2' (z_i (T))}{T'' (z_i (T)) + v_i'' (z_i (T))},
\]

\[
\frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} z_i (\theta_0, \epsilon) = -\frac{\eta_1' (z_i (T)) + \eta_2' (z_i (T))}{T'' (z_i (T)) + v_i'' (z_i (T))}, \quad \forall i \in [0, 1],
\]

where the second equality for the term \( \frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} z_i (\theta, \epsilon_0) \) uses (A.35), and the equality for the term \( \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} z_i (\theta_0, \epsilon) \) uses the fact that \( \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} \zeta (\theta_0, \epsilon) = 0 \), which follows from (A.29) and the assumption that \( \theta_0 = 0 \). These equations simplify when \( i \in [i_3, i_5] \). In particular,

\[
\frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} z_i (\theta, \epsilon_0) = -\frac{\mu_1' (z_i (T))}{T'' (z_i (T)) + v_i'' (z_i (T))}, \quad \forall i \in [i_3, i_5], \tag{A.36}
\]

\[
\frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} z_i (\theta_0, \epsilon) = -\frac{\eta_1' (z_i (T))}{T'' (z_i (T)) + v_i'' (z_i (T))},
\]

This simplification is explained by the observations that (i) since \( \mu_2 (z_i (T)) = 1 \) when \( i \in [i_3, i_5] \), \( \mu_2' (z_i (T)) = 0 \) when \( i \in [i_3, i_5] \), and (ii) the support of \( \eta_2 \) is \([\hat{z}_1, \hat{z}_2] \), so that \( \eta_2' (z_i (T)) = 0 \) when \( i \in [i_3, i_5] \).

When \( (\theta, \epsilon) = (\theta_0, \epsilon_0) \) and \( i \in [i_3, i_5] \), (A.33)-(A.34) also simplify:

\[
\frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} T (z_i (\theta_0, \epsilon_0), \theta, \epsilon_0) = \mu_1 (z_i (T)) - 1, \quad \forall i \in [i_3, i_5], \tag{A.37}
\]

\[
\frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} T (z_i (\theta_0, \epsilon_0), \theta_0, \epsilon) = \eta_1 (z_i (T)),
\]

where the first equality uses (A.35) and the fact that \( \mu_2 (z) = 1 \) when \( z \in [\hat{z}_3, \hat{z}_5] \), and the second equality uses the fact that \( \eta_2 (z) = 0 \) when \( z \in [\hat{z}_3, \hat{z}_5] \).
Recalling that \( i_a = i_3 \) and \( i_b = i_5 \), it follows from (A.36) and (A.37) that

\[
\forall i \in (i_a, i_b),
\frac{\partial}{\partial \theta}\bigg|_{\theta=\theta_0} z_i(\theta, \epsilon_0) \frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=\epsilon_0} T(z_i(\theta_0, \epsilon_0), \theta_0, \epsilon) - \frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=\epsilon_0} z_i(\theta_0, \epsilon) \frac{\partial}{\partial \theta}\bigg|_{\theta=\theta_0} T(z_i(\theta_0, \epsilon_0), \theta, \epsilon_0)
\]

\[
\left( + \text{ on } (i_3, i_4), - \text{ on } (i_4, i_5) \right) - \mu_1(z_i(T)) \eta_1(z_i(T)) + \left( + \text{ on } (i_3, i_4), - \text{ on } (i_4, i_5) \right) - \left[ \mu_1(z_i(T)) - 1 \right]
\]

\[
= \frac{T''(z_i(T)) + v_i''(z_i(T))}{\eta_1(z_i(T))} < 0.
\]

(A.38)

where the signs are derived from the assumptions we made above about \( \eta_1 \) and \( \mu_1 \) — in particular note that \( \mu_1(z) > 0 \) is increasing on \((\hat{z}_3, \hat{z}_5)\) and, by Lemma (A.1), \( \mu_1(\hat{z}_4) = 1 \) — as well as the fact that because \( T \in \mathcal{F}_2 \), and \( z_i(T) > 0, \forall i \in I \), it follows that \( 0 > \frac{d^2}{dz_i^2} \hat{T}_i^T(z_i(T)) = -v_i''(z_i(T)) + T''(z_i(T)), \forall i \in I \) (see Section 6.1).

Next observe that:

- The support of \( i \rightarrow \frac{\partial}{\partial \theta}\bigg|_{\theta=\theta_0} z_i(\theta, \epsilon_0) \) is \([i_2, i_5]\).
- The support of \( i \rightarrow \frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=\epsilon_0} T(z_i(\theta_0, \epsilon_0), \theta_0, \epsilon) \) is \([i_1, i_2] \cup [i_3, i_5]\).

Recalling that \( i_a = i_3 \) and \( i_b = i_5 \), it follows that

\[
\frac{\partial}{\partial \theta}\bigg|_{\theta=\theta_0} z_i(\theta, \epsilon_0) \frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=\epsilon_0} T(z_i(\theta_0, \epsilon_0), \theta_0, \epsilon) = 0, \ \forall i \notin (i_a, i_b).
\]

(A.39)

To understand why the above expression is equal to zero when \( i \in \{i_2, i_3, i_5\} \), note that the expressions \( \frac{\partial}{\partial \theta}\bigg|_{\theta=\theta_0} z_i(\theta, \epsilon_0) \) and \( \frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=\epsilon_0} T(z_i(\theta_0, \epsilon_0), \theta_0, \epsilon) \) are equal to zero on the boundaries of their supports.

- The support of \( i \rightarrow \frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=\epsilon_0} z_i(\theta_0, \epsilon) \) is contained in \([i_1, i_2] \cup [i_3, i_5]\).
- The support of \( i \rightarrow \frac{\partial}{\partial \theta}\bigg|_{\theta=\theta_0} T(z_i(\theta_0, \epsilon_0), \theta, \epsilon_0) \) is \([i_2, 1]\).

It follows that

\[
\frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=\epsilon_0} z_i(\theta_0, \epsilon) \frac{\partial}{\partial \theta}\bigg|_{\theta=\theta_0} T(z_i(\theta_0, \epsilon_0), \theta, \epsilon_0) = 0, \ \forall i \notin (i_a, i_b).
\]

(A.40)

Again, the above condition uses the fact that \( \frac{\partial}{\partial \epsilon}\bigg|_{\epsilon=\epsilon_0} z_i(\theta_0, \epsilon) \) and \( \frac{\partial}{\partial \theta}\bigg|_{\theta=\theta_0} T(z_i(\theta_0, \epsilon_0), \theta, \epsilon_0) \) are equal to zero on the boundaries of their supports. Together (A.39), (A.40), and the inequality established in (A.38) are equivalent to (35). We have now established that the family \( (T^{\theta, \epsilon}) \) satisfies all of the conditions required by the lemma. \( \square \)
A.10.2 Proof of Lemma A.1

Assume, for contradiction, that, for all \( i \in (i_3, i_5) \), \( \mu_1 (z_i (T)) \neq 1 \). Then, since the function \( i \mapsto \mu_1 (z_i (T)) \) is smooth (this follows from the assumed smoothness of relevant functions and the implicit function theorem), \( \mu_1 (z_{i_3} (T)) = 0 \), and \( i \mapsto \mu_1 (z_i (T)) \) is a constant function on \([i_5, 1]\), it follows from the intermediate value theorem that \( \mu_1 (z_i (T)) < 1, \forall i \in [i_3, 1] \). So

\[
\int_0^1 g_i (T) \mu_1 (z_i (T)) \, di = \int_0^1 g_i (T) \mu_1 (z_i (T)) \, di < \int_0^1 g_i (T) \, di
\]

where the first equality follows from the fact that the support of \( \mu_1 \) is \([\hat{z}_3, \hat{z}]\); the first inequality from the our conclusion that \( \mu_1 (z_i (T)) < 1, \forall i \in [i_3, 1] \) and the fact that \( g_i (T) > 0, \forall i \in [0, 1] \); the second equality form the fact that \( \mu_2 (z) = 1 \) for all \( z \in [\hat{z}_3, \hat{z}] \), and the last inequality from the fact that the \( \mu_2 \) is nonnegative everywhere and \( \mu_2 (z) > 0 \) for \( z \in (\hat{z}_2, \hat{z}_3) \). However, (A.41) contradicts (A.26). So the assumption that \( \mu_1 (z_i (T)) \) is never equal to 1 on \((i_3, i_5)\) leads to a contradiction, completing the proof. □

A.10.3 Elaboration of claims relating to (A.29)-(A.30).

This section discusses in more detail claims made in the paragraph of the proof of Lemma 3 that contains the conditions (A.29)-(A.30). Lemma A.2 implies that it is possible to choose \( \theta^* > 0, \epsilon^* > 0, \zeta^* > 0 \) sufficiently small that, for all \( \theta, \epsilon, \zeta \) with \( |\theta| \leq \theta^*, |\epsilon| \leq \epsilon^*, |\zeta| \leq \zeta^* \), \( \tilde{T}^{\theta, \epsilon} - \zeta \mu_2 \in \hat{\mathcal{F}} \). Let \( \tau = \epsilon^* \) and \( \zeta = \zeta^* \). For \( \epsilon \in [-\tau, \tau] \), define \( F_\epsilon (\theta, \zeta) : \mathbb{R}^2 \to \mathbb{R} \) by

\[
F_\epsilon (\theta, \zeta) = \frac{\int g_i \left( \tilde{T}^{\theta, \epsilon} - \zeta \mu_2 \right) \mu_1 \left( z_i \left( \tilde{T}^{\theta, \epsilon} - \zeta \mu_2 \right) \right) \, di}{\int g_i \left( \tilde{T}^{\theta, \epsilon} - \zeta \mu_2 \right) \mu_2 \left( z_i \left( \tilde{T}^{\theta, \epsilon} - \zeta \mu_2 \right) \right) \, di}.
\]

Choose \( M \geq |F_\epsilon (\theta, \zeta)|, \forall \epsilon \in [-\tau, \tau], \forall \theta \in [-\theta^*, \theta^*], \forall \zeta \in [-\zeta^*, \zeta^*] \). Let \( \overline{\theta} = \min \left\{ \theta^*, \frac{\zeta^*}{M} \right\} \). The smoothness of \( F_\epsilon (\theta, \zeta) \) implies that, in particular, \( F_\epsilon (\theta, \zeta) \) is continuous in \( \theta \) and uniformly Lipschitz continuous in \( \zeta \)\(^27\). It now follows from the Picard-Lindelöf theorem (see Theorem 1.1 on p. 8 of Hartman (1982)) that, for all \( \epsilon \in [-\tau, \tau] \), there exists a unique function \( \zeta_\epsilon (\theta) : [-\overline{\theta}, \overline{\theta}] \to [-\overline{\zeta}, \overline{\zeta}] \) satisfying \( \zeta_\epsilon (0) = 0 \) and

\[
\frac{d}{d\theta} \zeta_\epsilon (\theta) = F_\epsilon (\theta, \zeta_\epsilon (\theta)), \quad (A.42)
\]

for all \( \theta \in [-\overline{\theta}, \overline{\theta}] \). We can write \( \zeta (\theta, \epsilon) = \zeta_\epsilon (\theta) \) and \( \theta = -\overline{\theta} \xi = -\tau \). As (A.42) is equivalent to (A.30), this establishes the existence of the function \( \zeta (\epsilon, \theta) \) in the paragraph containing equations (A.29)-(A.30). The smoothness of \( (\theta, \epsilon) \mapsto \zeta (\theta, \epsilon) \), and hence also of \( (z, \theta, \epsilon) \mapsto T (z, \theta, \epsilon) \), follows \(^{27}\)Uniform Lipschitz continuity of \( F_\epsilon \) in \( \zeta \) means that there exists constant \( K \) satisfying \( |F_\epsilon (\theta, \zeta) - F_\epsilon (\theta', \zeta)| \leq K |\theta - \theta'|, \forall \theta, \theta' \in [-\overline{\theta}, \overline{\theta}], \forall \zeta \in [-\overline{\zeta}, \overline{\zeta}] \).
from the smoothness of \((\theta, \epsilon, \zeta) \mapsto F \epsilon (\theta, \zeta)\) and Corollary 4.1 on p. 101 of Hartman (1982). That \(\hat{T}^\theta, \epsilon \in \hat{T}\) follows from the fact that \(\zeta (\theta, \epsilon) \in [-\xi, \xi], \forall \theta \in [-\theta, \theta], \forall \epsilon \in [-\tau, \tau]\) and the conditions on parameters established above under which \(\hat{T}^\theta, \epsilon - \zeta \mu_2 \in \hat{T}\).

### A.10.4 A Variant of Lemma 3

This section discusses the proof of a variant of Lemma 3. I appeal to this variant in the proof of Lemma 4.

**Lemma A.2.** Let \(T \in \hat{T}\) be a tax policy in response to which all agents select a positive level of income and let \(i_a, i_b \in (0, 1)\) be such that \(i_a < i_b\). Then there exists a family \((T^\theta, \epsilon) \in T_2\) with \(T^{\theta_0, \epsilon_0} = T\) for some interior parameter values \(\theta_0, \epsilon_0\) and that satisfies (22), (23), and

\[
\frac{\partial}{\partial \theta} z_i (\theta, \epsilon_0) \frac{\partial}{\partial \epsilon} T (z_i (\theta_0, \epsilon_0), \theta, \epsilon) \begin{cases} > 0, & \text{if } i \in (i_a, i_b), \\ = 0, & \text{if } i \notin (i_a, i_b). \end{cases}
\]

(A.43)

This lemma differs from Lemma 3 only in that the inequality in \((A.43)\) points in the opposite direction to \((35)\). If one modifies the construction in the proof of Lemma 3 only by assuming that \(\eta_1\) is decreasing (rather than increasing) on \((\hat{z}_3, \hat{z}_4)\) and increasing (rather than decreasing) on \((\hat{z}_4, \hat{z}_5)\), so that \(\eta_1 (z) < 0\) (rather than \(\eta_1 (x) > 0\)) on \((\hat{z}_3, \hat{z}_5)\), and correspondingly if one assumes that \(\eta_2 (z) > 0\) on \((\hat{z}_1, \hat{z}_2)\) (rather than \(\eta_2 (z) < 0\)), then one flips the inequality in \((35)\), and so attains \((A.43)\). □

### A.11 Proof of Lemma 5

Assume that welfare weights \(g\) are not structurally utilitarian. By Lemma 4 in this case, we may choose \((T^\theta, \epsilon)_{\theta \in \Theta, \epsilon \in E} \in T_2\) satisfying (22)-(24). The construction of this family is presented in the proofs of Lemmas 3 and 4. Let us consider again the construction of \((T^\theta, \epsilon)\). First, by Corollary 3 since \(g\) is not structurally utilitarian we select \(T \in \hat{T}\) such that for some such that for some \(i_a, i_b \in (0, 1)\) with \(i_a < i_b\), either condition (28) or (29) is satisfied. An examination of the construction of the proof of Corollary 3 shows that it is possible to select \(T\) such that

\[T' (z_0 (T)) \neq 0.\]

(A.44)

We did not previously assume property (A.44) but let us assume henceforth that (A.44) is satisfied. Next we use tax policy \(T\) and \(i_a, i_b\) with the above properties to construct a family of tax polices, \((T^\theta, \epsilon)\), as in the proof of Lemma 3 of the form \(T^\theta, \epsilon = T + [\theta \times \mu_1] - [\gamma (\theta, \epsilon) \times \mu_2] + [\epsilon \times (\eta_1 + \eta_2)]\) (see (A.31)). The proof of Lemma 4 shows that such a family satisfies (22)-(24). It follows from their construction in the proof of Lemma 3 that the supports of the functions \(\mu_1, \mu_2, \eta_1, \)
and $\eta_2$ are all contained in the set $[\hat{z}_1, +\infty)$, where $\hat{z}_1 = z_{i_1}(T)$ was defined in the beginning of the proof of Lemma 3. As $0 < i_1$, it follows from assumptions in Section 6.1 that

$$z_0(T) < z_{i_1}(T) = \hat{z}_1.$$  \hfill (A.45)

It follows that

$$T^{\theta, \epsilon}(z) = T(z), \quad \forall z \in [0, \hat{z}_1], \forall \theta \in \Theta, \forall \epsilon \in E.$$  \hfill (A.46)

So $T^{\theta, \epsilon}(z) = T(z, \theta, \epsilon)$ does not depend on $\theta$ or $\epsilon$ for $z$ below $\hat{z}_1$. Recall that in the construction of $(T^{\theta, \epsilon})$, we assumed that $\theta_0 = 0$ and $\epsilon_0 = 0$, so that, by (A.31), $T^{\theta_0, \epsilon_0} = T$.

**Lemma A.3.** There exists a family of tax reforms $(\Delta T^\xi)_{\xi \in \Xi}$ where $\Xi = [\xi, \xi']$ for real numbers $\xi, \xi'$ satisfying $\xi < 0 < \xi'$, and such that $\Delta T^0 \equiv 0$, the support of $\Delta T^\xi$ is contained in $[0, \hat{z}_1]$ for all $\xi \in \Xi$, the map $(z, \xi) \mapsto \Delta T^\xi(z)$ is smooth, and for some sets $\Theta' = [\theta', \theta''] \subseteq \Theta$, $E' = [\epsilon', \epsilon''] \subseteq E$, with $\theta' < 0 < \theta''$ and $\epsilon' < 0 < \epsilon''$,

$$\int g_i \left( T^{\theta, \epsilon} + \Delta T^\xi \right) \frac{\partial}{\partial \xi} \Delta T \left( z_i \left( T^{\theta, \epsilon} + \Delta T^\xi \right) , \xi \right) \mathrm{d}i = 0, \quad \forall \theta \in \Theta', \forall \epsilon \in E', \forall \epsilon' \in \Xi,$$

$$\frac{\mathrm{d}}{\mathrm{d} \xi} R \left( T^{\theta, \epsilon} + \Delta T^\xi \right) \neq 0, \quad \forall \theta \in \Theta', \forall \epsilon \in E', \forall \epsilon' \in \Xi,$$

where in (A.47) we use the notation $\Delta T(z_i, \xi) = \Delta T^\xi(z_i)$. Moreover, $(\Delta T^\xi)_{\xi \in \Xi}$ can be constructed so that $T^{\theta, \epsilon} + \Delta T^\xi \in \mathcal{F}$, $\forall \xi \in \Xi$, $\forall \theta \in \Theta'$, $\forall \epsilon \in E'$.

To understand this lemma, first recall that $T^{\theta_0, \epsilon_0} = T$, and note that, by construction, all tax policies $T^{\theta, \epsilon}$ are equal to $T$ on the interval $[0, \hat{z}_1]$, which contains the support of all tax reforms $\Delta T^\xi$. Lemma A.3 says that the family of reforms $(\Delta T^\xi)$ is such that varying $\xi$ in $T^{\theta, \epsilon} + \Delta T^\xi$ has no effect on welfare according to welfare weights (see (A.47)), but does have an effect on revenue (see (A.48)). Obviously, if $T^{\theta, \epsilon}$ were an optimal tax policy, it would not be possible to do this. However note that $T$, which coincides with all policies $T^{\theta, \epsilon}$ at the bottom of the income distribution, is such that marginal tax rate at at the income $z_0(T)$ at the bottom of the income distribution is non-zero, and, moreover, since $T \in \mathcal{F}$, $z_0(T) > 0$ (see Section 6.1), and hence, none of the tax policies $T^{\theta, \epsilon}$ are optimal. As shown by Saez and Stantcheva (2016), (see Section A.2 of their Appendix), at an optimal tax policy in the generalized social welfare weights framework, the marginal tax rate for the bottom earner is zero if the bottom earner has a positive income. Lemma A.3 is proven in Section B.1 of the Appendix.

So let us assume that a family $(\Delta T^\xi)$ with the properties in Lemma A.3 is chosen. Noting that $\Delta T^0 \equiv 0$, it follows from (A.48) and the implicit function theorem that there exists $\theta'', \epsilon'' \in \Theta'$ with $\theta'' < 0 < \theta''$ and $\epsilon'' < 0 < \epsilon''$ and a function $\xi : \Theta'' \times E'' \to \Xi$, where $\Theta'' = [\theta'', \theta''']$.\hfill (A.47)
and $E'' = [\xi'', \xi''']$, satisfying:

$$
\hat{\xi} (\theta_0, \epsilon_0) = 0, \quad (A.49)
$$

$$
R \left( T^{\theta, \epsilon} + \Delta T^{\hat{\xi} (\theta, \epsilon)} \right) = R \left( T^{\theta_0, \epsilon_0} \right), \quad \forall \theta \in \Theta', \forall \epsilon \in E'', \quad (A.50)
$$

where, in (A.49), $\Delta T^{\hat{\xi} (\theta, \epsilon)}$ is $\Delta T^{\hat{\xi}}$ evaluated at $\xi = \hat{\xi} (\theta, \epsilon)$. Because the other functions occurring in (A.50) are smooth, it follows that $\hat{\xi} (\theta, \epsilon)$ is smooth.

Define the doubly parameterized family of tax policies $\hat{T}^{\theta, \epsilon}$ by

$$
\hat{T}^{\theta, \epsilon} = T^{\theta, \epsilon} + \Delta T^{\hat{\xi} (\theta, \epsilon)}, \quad \forall \theta \in \Theta'', \forall \epsilon \in E''. \quad (A.51)
$$

It follows from Lemma A.3, the fact that $(T^{\theta, \epsilon}) \in T^2$, and the fact that $\hat{\xi} (\theta, \epsilon)$ is smooth, that $\hat{T}^{\theta, \epsilon}$ satisfies (22)-(24).

Lemma A.4. $(\hat{T}^{\theta, \epsilon})_{\theta \in \Theta'', \epsilon \in E''} \in T^2$ satisfies (22)-(24).

It is straightforward to verify that $(\hat{T}^{\theta, \epsilon})_{\theta \in \Theta'', \epsilon \in E''}$ inherits the properties (22)-(24) from $(T^{\theta, \epsilon})_{\theta \in \Theta, \epsilon \in E'}$. Moreover it follows from (A.51) and (A.50) that $(\hat{T}^{\theta, \epsilon})_{\theta \in \Theta'', \epsilon \in E''}$ has constant revenue. Thus, we have constructed a family of tax policies with the desired properties, which completes the proof. □

A.12 Proof of Proposition 5

In the poverty alleviation model of Section 6.4, condition (36) is equivalent to condition (22). Given that $\kappa (\theta_0, \epsilon) = 0, \forall \epsilon$, it follows that $\frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} \kappa (\theta_0, \epsilon) = 0$, which implies that the assumption that

$$
\int g (\theta_0, \epsilon_0) \left[ z_i (\theta_0, \epsilon_0) - \alpha \right] d \bar{i} = 0 \quad (A.52)
$$

is equivalent to (23).

I now establish some facts that will be useful for establishing (24). First, using (36), we have

$$
\frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} \kappa (\theta, \epsilon_0) = \int \frac{g_i (\theta_0, \epsilon_0)}{\int g_j (\theta_0, \epsilon_0) d \bar{j}} \left[ z_i (\theta_0, \epsilon_0) + f (z_i (\theta_0, \epsilon_0)) \right] d \bar{i} \nonumber
$$

$$
= \int \frac{g_i (\theta_0, \epsilon_0)}{\int g_j (\theta_0, \epsilon_0) d \bar{j}} z_i (\theta_0, \epsilon_0) d \bar{i} + \int \frac{g_i (\theta_0, \epsilon_0)}{\int g_j (\theta_0, \epsilon_0) d \bar{j}} f (z_i (\theta_0, \epsilon_0)) d \bar{i} \nonumber
$$

$$
= \alpha + \beta \int \frac{g_i (\theta_0, \epsilon_0)}{\int g_j (\theta_0, \epsilon_0) d \bar{j}} f (z_i (\theta_0, \epsilon_0)) d \bar{i}, \quad (A.53)
$$

where the third equality follows from (A.52), and $\beta$ is a label for the last integral. It follows from the assumptions of Section 6.4 that $\beta > 0$. Let $\bar{i}$ be the unique agent satisfying $z_i (\theta_0, \epsilon_0) = \bar{z}$. 43
That such a \( \tilde{\gamma} \) exists and is unique follows from Lemma C.1. It also follows from Lemma C.1 and the assumptions in Section 6.4 that all agents in the interval \([0, \tilde{i}]\) earn an income less than \( \tilde{z} \) when facing tax policy \( T^{\theta_0, \epsilon_0} \). It follows from assumptions on \( f \) in Section 6.4 that, for all incomes \( z \) earned by agents in the interval \([0, \tilde{i}]\), when facing tax policy \( T^{\theta_0, \epsilon_0} \), \( f'(z_1(\theta_0, \epsilon_0)) = 0 \) and \( f''(z_1(\theta_0, \epsilon_0)) = 0 \). Taking this into account, and applying the implicit function theorem to the first order condition for agents' optimization problem when facing tax policy \( T^{\theta_0, \epsilon_0} \), we have

\[
\frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} z_i(\theta, \epsilon_0) = -\frac{1}{v_i''(z_i(\theta_0, \epsilon_0))}, \quad \forall i \in [0, \tilde{i}]. \tag{A.54}
\]

Again, using the fact that \( f'(z_1(\theta_0, \epsilon_0)) = 0, \forall i \in [0, \tilde{i}] \) and (A.53), and the fact that \( \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} \kappa(\theta_0, \epsilon) = 0 \), we have

\[
\frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} T(z_i(\theta_0, \epsilon_0), \theta, \epsilon) = z_i(\theta_0, \epsilon_0) - (\alpha + \beta), \quad \forall i \in [0, \tilde{i}]. \tag{A.55}
\]

Using (A.54) and (A.55), we have that, for all \( i \in [0, \tilde{i}] \),

\[
\frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} z_i(\theta, \epsilon_0) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} T(z_i(\theta_0, \epsilon_0), \theta, \epsilon) - \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} z_i(\theta_0, \epsilon) \frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} T(z_i(\theta_0, \epsilon_0), \theta, \epsilon)
= \left( -\frac{1}{v_i''(z_i(\theta_0, \epsilon_0))} \right) (-z_i(\theta_0, \epsilon_0) + \alpha) - \left( \frac{1}{v_i''(z_i(\theta_0, \epsilon_0))} \right) (z_i(\theta_0, \epsilon_0) - (\alpha + \beta))
= \frac{1}{v_i''(z_i(\theta_0, \epsilon_0))} \beta > 0.
\]

Next recall the relationship between the variables \( c_i, \hat{u}_i \) and \( z_i \) from Section 4: \( c_i = \hat{u}_i + v_i(z_i) \). It follows that \( \hat{g}_i(\hat{u}_i, z_i) = \hat{g}(\hat{u}_i + v_i(z_i)) \), and hence \( \frac{\partial}{\partial \hat{u}_i} \hat{g}_i(\hat{u}_i, z_i) = \hat{g}'(\hat{u}_i + v_i(z_i)) v_i'(z_i) = \hat{g}'(c_i) v_i'(z_i) \). It follows from the above that:

\[
\int \frac{\partial}{\partial z_i} \hat{g}_i(\hat{U}_i(\theta_0, \epsilon_0), z_i(\theta_0, \epsilon_0)) \left[ \frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} z_i(\theta, \epsilon_0) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} T(z_i(\theta_0, \epsilon_0), \theta, \epsilon) - \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} z_i(\theta_0, \epsilon) \frac{\partial}{\partial \theta} \bigg|_{\theta = \theta_0} T(z_i(\theta_0, \epsilon_0), \theta, \epsilon) \right] \, d\bar{i} \tag{A.56}
= \beta \int_0^{\tilde{i}} \hat{g}'(c_i(\theta_0, \epsilon_0)) \frac{v_i'(z_i(\theta_0, \epsilon_0))}{v_i''(z_i(\theta_0, \epsilon_0))} \, d\bar{i} < 0,
\]

where the upper bound of integration in the second integral follows from that fact that all \( i \in (\tilde{i}, 1] \) are above the poverty line when facing tax policy \( T^{\theta_0, \epsilon_0} \) and hence \( \hat{g}'(c_i(\theta_0, \epsilon_0)) = 0 \) for all such
agents. The inequality follows from the fact that \( v_i'(z_i) > 0 \) and \( v_i''(z_i) > 0 \), for all \( z_i \), \( \overline{g}'(c_i) \leq 0 \), for all \( c_i \), and, since a positive measure of agents in the interval \([0, \overline{v}]\) is beneath the poverty line at tax policy \( T_{\theta_0, \epsilon_0} \), \( \overline{g}'(c_i(\theta_0, \epsilon_0)) < 0 \) for a positive measure of agents in \([0, \overline{v}]\). It now follows from Lemma 2 that the family \((T_{\theta, \epsilon})\) in the poverty alleviation model of Section 6.4 satisfies \((24)\). (Note that the proof of Lemma 2 also establishes that the first integral in \( (A.56) \) is equal to \( \int g_i(\theta, \epsilon_0) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} T(z_i(\theta, \epsilon), \theta, \epsilon) \), which shows that \( (A.56) \) justifies condition \((57)\) in the main text.) □

**B Lemmas supporting Lemma 5**

### B.1 Proof of Lemma A.3

My proof of Lemma A.3 relies on a few other lemmas. The first lemma establishes the linearity of the function \( f(\Delta T) = \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} R(T + \epsilon \Delta T) \).

**Lemma B.1.** Let \( T \in \hat{\Theta} \). Let \( \Delta T_1 \) and \( \Delta T_2 \) be smooth tax reforms and let \( r_1 \) and \( r_2 \) be real numbers. Then \( \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} R(T + \epsilon (r_1 \Delta T_1 + r_2 \Delta T_2)) = r_1 \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} R(T + \epsilon \Delta T_1) + r_2 \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} R(T + \epsilon \Delta T_1) \).

**Lemma B.1** is proven in Section B.3. The second lemma is closely related to Lemma A.3.

**Lemma B.2.** Let \( T \in \hat{\Theta} \) (so that in particular \( z_0(T) > 0 \)), and suppose that \( T'(z_0(T)) \neq 0 \). Let \( z_* \) be such that \( z_0(T) < z_* \leq z_1(T) \). Then there exists a desirable revenue neutral tax reform \( \Delta T \) with support contained in \([0, z_*]\); formally, there exists a smooth tax reform \( \Delta T \) with support contained in \([0, z_*]\) such that \( \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} R(T + \epsilon \Delta T) = 0 \) and \( \int g_i(T) \Delta T(z_i(T)) \, di < 0 \). Moreover, there exist tax reforms \( \Delta T_1, \Delta T_2 \), with supports contained in \([0, z_*]\) such that \( \Delta T = \Delta T_1 - \Delta T_2, \Delta T_2(z) \geq 0, \forall z \), and \( \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=0} R(T + \epsilon \Delta T_2) \neq 0 \).

Recall that the set of agents is \( I = [0, 1] \), \( z_0(T) \) and \( z_1(T) \) are the optimal responses to \( T \) for agents 0 and 1 respectively. By the assumptions of Section 6.4, \( z_0(T) \) and \( z_1(T) \) are respectively the bottom and top of the income distributions earned in response to \( T \) (see also Lemma C.1).

**Lemma B.2** is proved in Section B.4. With these lemmas in place, I now proceed to prove Lemma A.3.

Recall from Section A.11 that \( T_{\theta_0, \epsilon_0} = T \) for a tax policy \( T \in \hat{\Theta} \) satisfying \( T'(z_0(T)) \neq 0 \). Recall also from Section A.11 that \( \hat{z}_1 = z_i(T) \), and moreover, \( z_0(T) < \hat{z}_1 < z_1(T) \). So letting \( \hat{z}_1 \) play the role of \( z_* \) in Lemma B.2, there exist tax reforms \( \Delta T_1, \Delta T_2 \), and \( \Delta T \), all with supports contained in \([0, \hat{z}_1]\) and satisfying the properties in Lemma B.2 in relation to the tax policy \( T = T_{\theta_0, \epsilon_0} \). Define the function \( F \)

\[
F(\xi, r) = \int g_i(T + \xi \Delta T_1 - r \Delta T_2) \Delta T_1(z_i(T + \xi \Delta T_1 - r \Delta T_2)) \, di \quad \int g_i(T + \xi \Delta T_1 - r \Delta T_2) \Delta T_2(z_i(T + \xi \Delta T_1 - r \Delta T_2)) \, di
\]

\[\text{(B.1)}\]

\(^{28}\) It follows from Lemma B.2 that if \( \xi \) and \( r \) are sufficiently close to 0, then \( T + \xi \Delta T_1 - r \Delta T_2 \in \hat{\Theta} \), and hence \( z_i(T + \xi \Delta T_1 - r \Delta T_2) \) is uniquely defined, and so \( g_i(T + \xi \Delta T_1 - r \Delta T_2) \) is also uniquely defined.
It follows from the properties stated in Lemma C.2 (which apply to \( T, \Delta T_1 \) and \( \Delta T_2 \)) and the smoothness of the relevant functions that if \( \xi \) and \( r \) are sufficiently close to zero, then the denominator in the above expression is nonzero\(^{29}\). Note that \( F \) is smooth in its arguments. It follows from the Picard-Lindelöf theorem that there exist real numbers \( \xi, \xi \in \Xi \) with \( \xi < 0 < \xi \), and a function \( s : [\xi, \xi] \to \mathbb{R} \) satisfying

\[
\begin{align*}
s (0) &= 0, \\
s' (\xi) &= F (\xi, s (\xi)) , \quad \forall \xi \in \Xi,
\end{align*}
\]

where \( \Xi = [\xi, \xi] \). Define the family of tax reforms \( (\Delta T^\xi)_{\xi \in \Xi} \) by the condition

\[
\Delta T^\xi = \xi \Delta T_1 - s (\xi) \Delta T_2 , \quad \forall \xi \in \Xi.
\]

Observe that \( \Delta T^0 \equiv 0 \) and, for all \( \xi \in \Xi \), the support of \( \Delta T^\xi \) is contained in \([0, \hat{z}_1] = [0, z_*]\) because the supports of \( \Delta T_1 \) and \( \Delta T_2 \) are contained in \([0, \hat{z}_1]\). It follows from the smoothness of the function \( F (\xi, r) \), and Corollary 4.1 on p. 101 of Hartman (1982) that the function \( s (\xi) \) is smooth, and hence also that the map \( (z, \xi) \mapsto \Delta T^\xi (z) \) is smooth. Lemma C.2 implies that it is possible to choose \( \xi \) and \( \xi \), and also \( \theta', \theta' \in \Theta, \xi', \xi' \in E \) with \( \theta' < 0 < \theta', \xi' < 0 < \xi' \) so that \( T^\theta, \epsilon + \Delta T^\xi = T + [\theta \times \mu_1] - [\xi (\theta, \epsilon) \times \mu_2] + [\epsilon \times (\eta_1 + \eta_2)] + \Delta T^\xi \in \tilde{\mathcal{H}}, \forall \theta \in \Theta' = [\theta', \theta'], \forall \epsilon \in E' = [\epsilon', \epsilon'], \forall \xi \in \Xi = [\xi, \xi] \). So let us assume that \( \xi, \xi, \theta', \theta', \xi', \xi' \), and \( \epsilon' \) are so chosen.

Next observe that

\[
\frac{\partial}{\partial \xi} \bigg|_{\xi = \xi'} \Delta T^\xi (z, \xi) = \Delta T_1 (z) - s' (\xi) \Delta T_2 (z) , \quad \forall \xi' \in \Xi, \forall z \in Z,
\]

where \( \Delta T (z, \xi) = \Delta T^\xi (z) \). Recalling that \( T^\theta, \epsilon = T \), and using (B.1), (B.4), and (B.5), it follows that (B.3) is equivalent to

\[
\int g_i \left( T + \Delta T^\xi \right) \frac{\partial}{\partial \xi} \bigg|_{\xi = \xi'} \Delta T (z, \xi) \, d i = 0 , \quad \forall \xi' \in \Xi.
\]

It follows that the family \( \Delta T^\xi \) satisfies (B.6).

Consider the tax reform \( \Delta T_1 - s' (0) \Delta T_2 \). This is just the tax reform \( \Delta T_1 - r \Delta T_2 \) in the special case in which \( r = s' (0) \). When facing the tax policies \( T + \epsilon (\Delta T_1 - s' (0) \Delta T_2) \) and \( T + \Delta T^\xi \), agent \( i \) faces, respectively, optimization problems \( \max_{z_i} \left[ \left. z_i - T (z_i) - \epsilon (\Delta T_1 (z_i) - s' (0) \Delta T_2 (z_i)) - v_i (z_i) \right| \right. \] and \( \max_{z_i} \left[ \left. z_i - T (z_i) - \Delta T^\xi (z_i) - v_i (z_i) \right| \right. \]. Note that, because \( T \in \tilde{\mathcal{H}} \), when \( \epsilon \) and \( \xi \) are sufficiently small, all agents select an interior income (see also Lemma C.2). Applying the implicit function

\(^{29}\)In particular, the facts that \( \Delta T_2 (z) \geq 0, \forall z \), and \( \frac{1}{\epsilon} \left. \right|_{\xi = 0} R (T + \epsilon \Delta T_2) \neq 0 \) imply that there exists a positive measure of agents \( i \) such that \( \Delta T_2 (z_i (T)) > 0 \), hence, invoking again \( \Delta T_2 (z) \geq 0, \forall z \), it follows that from the fact that welfare weights are positive \( \int g_i (T) \Delta T_2 (z_i (T)) \, d i > 0 \). That the denominator of (B.1) is positive now follows from our smoothness assumptions.
theorem to the agent’s first order conditions for these two problems, we have:

$$
\frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} z_i \left( T + \varepsilon \left( \Delta T_1 - s'(0) \Delta T_2 \right) \right) = -\frac{\Delta T_1 \left( z_i (T) \right) - s'(0) \Delta T_2 \left( z_i (T) \right)}{T''(z_i(T)) + v''(z_i(T))}
$$

\[= -\frac{\partial}{\partial \xi} \bigg|_{\xi=0} \Delta T \left( z_i (T), \xi \right) = \frac{d}{d\xi} \bigg|_{\xi=0} z_i \left( T + \Delta T^\xi \right), \]

where the second equality follows from (B.5). This, in turn, implies that

$$
\frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} R \left( T + \varepsilon \left( \Delta T_1 - s'(0) \Delta T_2 \right) \right)
$$

\[= \int \left[ \Delta T_1 \left( z_i (T) \right) - s'(0) \Delta T_2 \left( z_i (T) \right) \right] d\varepsilon + \int T' \left( z_i (T) \right) \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} z_i \left( T + \varepsilon \left( \Delta T_1 - s'(0) \Delta T_2 \right) \right) d\varepsilon
$$

\[= \int \frac{\partial}{\partial \xi} \bigg|_{\xi=0} \Delta T \left( z_i (T), \xi \right) d\varepsilon + \int T' \left( z_i (T) \right) \frac{d}{d\xi} \bigg|_{\xi=0} z_i \left( T + \Delta T^\xi \right) d\xi
$$

\[= \frac{d}{d\xi} \bigg|_{\xi=0} R \left( T + \Delta T^\xi \right), \]

where the second equality uses (B.5) and (B.7).

Next observe that, by the properties implied by Lemma B.2, 0 > \int g_i(T) \Delta T (z_i(T)) d\xi = \int g_i(T) \Delta T_1 (z_i(T)) d\xi - \int g_i(T) \Delta T_2 (z_i(T)) d\xi. So, since, again by the properties in Lemma B.2, \int g_i(T) \Delta T_2 (z_i(T)) d\xi > 0 (see footnote 29 of the Appendix), it follows that \( R(0,0) = \frac{\int g_i(T) \Delta T_1 (z_i(T)) d\xi}{\int g_i(T) \Delta T_2 (z_i(T)) d\xi} < 1 \). So, by (B.2) and (B.3), s'(0) < 1. It follows from Lemma B.1 and the properties of Lemma B.2 that

\[0 = \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} R \left( T + \varepsilon \Delta T \right) = \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} R \left( T + \varepsilon \left( \Delta T_1 - \Delta T_2 \right) \right)
\]

\[= \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} R \left( T + \varepsilon \Delta T_1 \right) - \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} R \left( T + \varepsilon \Delta T_2 \right)
\]

\[\neq \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} R \left( T + \varepsilon \left( \Delta T_1 - s'(0) \Delta T_2 \right) \right) = \frac{d}{d\xi} \bigg|_{\xi=0} R \left( T + \Delta T^\xi \right)
\]

where the non-equality \( \neq \) in the above derivation follows from the facts that s'(0) \neq 1 and \( \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} R \left( T + \varepsilon \Delta T_2 \right) \neq 0 \) (see Lemma B.2 for the latter). So, to summarize, \( \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} R \left( T + \Delta T^\xi \right) \neq 0 \). By our smoothness assumptions, if \( \bar{\xi} \) and \( \bar{\xi} \) in \( \Xi = [\bar{\xi}, \bar{\Xi}] \) are selected so as to be sufficiently close
to 0,

$$\frac{d}{d\xi} \bigg|_{\xi=\xi'} R \left( T + \Delta T^\xi \right) \neq 0, \quad \forall \xi' \in \Xi. \quad (B.9)$$

Let us assume that $\xi$ and $\xi'$ are so chosen.

Using the facts that, by the construction of $(T^{\theta,\epsilon})$, $T^{\theta,\epsilon}(z) + \Delta T^\xi(z) = T(z) + \Delta T^\xi(z), \forall z \in [0, \hat{z}_1], \forall \theta \in \Theta', \forall \epsilon \in E', \forall \xi \in \Xi$, and that, for all $\xi \in \Xi$, the support of $\Delta T^\xi$ is contained in $[0, \hat{z}_1]$ (see in particular Lemma [B.3] and (B.12)–(B.13) of Lemma [B.4]), it follows that, for all $\theta \in \Theta'$, for all $\epsilon \in E'$, and for all $\xi' \in \Xi$,

$$\int g_i \left( T + \Delta T^\xi \right) \frac{\partial}{\partial \xi} \bigg|_{\xi=\xi'} \Delta T \left( z_i \left( T + \Delta T^\xi' \right), \xi \right) \, di$$

$$= \int g_i \left( T^{\theta,\epsilon} + \Delta T^\xi \right) \frac{\partial}{\partial \xi} \bigg|_{\xi=\xi'} \Delta T \left( z_i \left( T^{\theta,\epsilon} + \Delta T^\xi' \right), \xi \right) \, di,$$

$$\frac{d}{d\xi} \bigg|_{\xi=\xi'} R \left( T + \Delta T^\xi \right) = \frac{d}{d\xi} \bigg|_{\xi=\xi'} R \left( T^{\theta,\epsilon} + \Delta T^\xi \right). \quad (B.10)$$

Conditions (A.47) and (A.48) follow from (B.6), (B.9), (B.10), and (B.11). We have now proven all the properties required by Lemma A.3. □

### B.2 Proof of Lemma A.4

The following lemmas collect some properties that follow fairly immediately from the above definitions, and which will be useful below.

**Lemma B.3.** For all $\theta \in \Theta', \epsilon \in E'$, and $\xi \in \Xi$,

$$z_i \left( T^{\theta,\epsilon} + \Delta T^\xi \right) \begin{cases} \in [0, \hat{z}_1], & \text{if } i \in [0, i_1], \\ \in [\hat{z}_1, +\infty), & \text{if } i \in [i_1, 1]. \end{cases} \quad (B.12)$$

**Lemma B.4.** For all $i \in [0, i_1], \theta \in \Theta', \epsilon \in E'$, and $\xi \in \Xi$,

$$z_i \left( T^{\theta_0,\epsilon_0} + \Delta T^\xi \right) = z_i \left( T^{\theta,\epsilon} + \Delta T^\xi \right), \quad (B.12)$$

$$g_i \left( T^{\theta_0,\epsilon_0} + \Delta T^\xi \right) = g_i \left( T^{\theta,\epsilon} + \Delta T^\xi \right). \quad (B.13)$$

For all $i \in [i_1, 1], \theta \in \Theta''$, and $\epsilon \in E''$,

$$z_i \left( \hat{T}_{\theta,\epsilon} \right) = z_i \left( T^{\theta,\epsilon} \right), \quad (B.14)$$

$$g_i \left( \hat{T}_{\theta,\epsilon} \right) = g_i \left( T^{\theta,\epsilon} \right). \quad (B.15)$$

---

30 The proof of Lemmas [B.3] and [B.4] depend on the fact that, for all $\xi \in \Xi$, the support of $\Delta T^\xi$ is contained in $[0, \hat{z}_1]$, but not the more detailed properties established in the current lemma, Lemma A.3
The lemmas are proven in Section B.5. We also state the following fact, which follows from the construction of $T^{\theta,\epsilon}$, and which will be useful in what follows:

**Fact 1.** $T^{\theta,\epsilon}(z)$ does not depend on $\theta$ and $\epsilon$ when $z \in [0, \hat{z}_1]$; that is $T^{\theta,\epsilon}(z) = T^{\theta_0,\epsilon_0}(z)$, $\forall z \in [0, \hat{z}_1], \forall \theta \in \Theta'$, $\forall \epsilon \in E''$.

We now proceed with the proof of Lemma B.2. Choose $\epsilon \in (\xi'', \xi')$ and $\theta' \in (\theta'', \theta')$. We have:

$$
\begin{align*}
A &= \int_0^{i_1} g_i \left( \hat{T}^{\theta',\epsilon} \right) \left[ \frac{\partial}{\partial \theta} \right]_{\theta=\theta'} T \left( z_i \left( \hat{T}^{\theta',\epsilon} \right), \theta, \epsilon \right) \\
&\quad + \frac{\partial}{\partial \xi} \left. \frac{\partial}{\partial \theta} \right|_{\theta=\theta'} \Delta T \left( z_i \left( \hat{T}^{\theta',\epsilon} \right), \xi \right) \left. \frac{\partial}{\partial \theta} \right|_{\theta=\theta'} \hat{\xi} \left( \theta, \epsilon \right) \\
&= \left[ \frac{\partial}{\partial \theta} \right]_{\theta=\theta'} \hat{\xi} \left( \theta, \epsilon \right) \int_0^{i_1} g_i \left( \hat{T}^{\theta',\epsilon} \right) \left. \frac{\partial}{\partial \xi} \right|_{\xi=\hat{\xi}(\theta',\epsilon)} \Delta T \left( z_i \left( \hat{T}^{\theta'^{\prime}}, \epsilon \right), \xi \right) \\
&= \left[ \frac{\partial}{\partial \theta} \right]_{\theta=\theta'} \hat{\xi} \left( \theta, \epsilon \right) \int_0^{i_1} g_i \left( T^{\theta',\epsilon} + \Delta T^{\xi(\theta',\epsilon)} \right) \left. \frac{\partial}{\partial \xi} \right|_{\xi=\hat{\xi}(\theta',\epsilon)} \Delta T^{\xi} \left( z_i \left( T^{\theta,\epsilon} + \Delta T^{\xi(\theta',\epsilon)} \right), \xi \right) \\
&= \left[ \frac{\partial}{\partial \theta} \right]_{\theta=\theta'} \hat{\xi} \left( \theta, \epsilon \right) \int_0^{i_1} g_i \left( T^{\theta_0,\epsilon_0} + \Delta T^{\xi(\theta',\epsilon)} \right) \left. \frac{\partial}{\partial \xi} \right|_{\xi=\hat{\xi}(\theta',\epsilon)} \Delta T^{\xi} \left( z_i \left( T^{\theta_0,\epsilon_0} + \Delta T^{\xi(\theta',\epsilon)} \right), \xi \right) \\
&= 0,
\end{align*}
$$

where the first equality follows from the definition (A.51) of $\hat{T}^{\theta,\epsilon}$; the second equality follows from Fact 1 and Lemma B.3 which imply that, when $i \in [0, i_1]$, $\frac{\partial}{\partial \theta} \left. T \right|_{\theta=\theta'} \left( z_i \left( T^{\theta',\epsilon} \right), \theta, \epsilon \right) = 0$; the third equality follows again from (A.51); the fourth equality follows from (B.12)-(B.13); the fifth equality follows from the fact that, for all $\xi \in \Xi$, the support of $\Delta T^{\xi}$ is contained in $[0, \hat{z}_1]$ and Lemma B.3 so that the integrand in the expression following the third equality is equal to zero when $i \in [i_1, 1]$; and the last equality follows from (A.47).
Next, observe that

\[ B = \int_{i_1}^{1} g_i \left( \hat{T}^{\theta',\epsilon} \right) \frac{\partial}{\partial \theta} \bigg|_{\theta = \theta'} T \left( z_i \left( \hat{T}^{\theta',\epsilon} \right), \theta, \epsilon \right) \, di \]

\[ = \int_{i_1}^{1} g_i \left( T^{\theta',\epsilon} \right) \frac{\partial}{\partial \theta} \bigg|_{\theta = \theta'} T \left( z_i \left( T^{\theta',\epsilon} \right), \theta, \epsilon \right) \, di \]

\[ = \int_{0}^{1} g_i \left( T^{\theta',\epsilon} \right) \frac{\partial}{\partial \theta} \bigg|_{\theta = \theta'} T \left( z_i \left( T^{\theta',\epsilon} \right), \theta, \epsilon \right) \, di \]

\[ = 0, \quad \text{(B.18)} \]

where the first equality follows from Lemma B.3, (A.51), and the fact that, for all \( \xi \in \Xi \), the support of \( \Delta T^\xi \) is contained in \([0, \hat{z}_1]\), so that \( T \left( z_i \left( \hat{T}^{\theta',\epsilon} \right), \theta, \epsilon \right) = \hat{T} \left( z_i \left( \hat{T}^{\theta',\epsilon} \right), \theta, \epsilon \right) \) when \( i \in [i_1, 1] \); the second equality follows from (B.14)-(B.15); the third equality follows from the fact that, by Fact 1, \( T \left( z_i \left( T^{\theta',\epsilon} \right), \theta, \epsilon \right) \) does not depend on \( \theta \) when \( i \in [0, i_1] \), so the integrand in the expression following the second equality is zero when \( i \in [0, i_1] \); and the last equality follows from the fact that \( (T^{\theta,\epsilon}) \) satisfies (22).

Putting together (B.16), (B.17), and (B.18), it follows that \( \left( \hat{T}^{\theta,\epsilon} \right)_{\theta \in \Theta^{\nu}, \epsilon \in E''} \) satisfies (22).

Next observe that:

\[ \int_{0}^{1} g_i \left( \hat{T}^{\theta_0,\epsilon_0} \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} \hat{T} \left( z_i \left( \hat{T}^{\theta_0,\epsilon_0} \right), \theta_0, \epsilon_0 \right) \, di \]

\[ = \int_{0}^{i_1} g_i \left( \hat{T}^{\theta_0,\epsilon_0} \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} \hat{T} \left( z_i \left( \hat{T}^{\theta_0,\epsilon_0} \right), \theta_0, \epsilon \right) \, di \]

\[ + \int_{i_1}^{1} g_i \left( \hat{T}^{\theta_0,\epsilon_0} \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon = \epsilon_0} \hat{T} \left( z_i \left( \hat{T}^{\theta_0,\epsilon_0} \right), \theta_0, \epsilon \right) \, di. \]

\[ \text{(B.19)} \]
Analyzing the first term:

\[
C = \int_0^{i_1} g_i \left( \hat{T}^{\theta_0, \epsilon_0} \right) \left[ \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} T \left( z_i \left( \hat{T}^{\theta_0, \epsilon_0} \right), \theta_0, \epsilon \right) \right.
\]

\[
+ \frac{\partial}{\partial \xi} \bigg|_{\xi=\hat{\xi}(\theta_0, \epsilon_0)} \Delta T^\xi \left( z_i \left( \hat{T}^{\theta_0, \epsilon_0} \right) \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} \hat{\xi} (\theta_0, \epsilon) \bigg] \, di
\]

\[
= \left[ \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} \hat{\xi} (\theta_0, \epsilon) \right] \int_0^{i_1} g_i \left( \hat{T}^{\theta_0, \epsilon_0} \right) \frac{\partial}{\partial \xi} \bigg|_{\xi=\hat{\xi}(\theta_0, \epsilon_0)} \Delta T^\xi \left( z_i \left( \hat{T}^{\theta_0, \epsilon_0} \right) , \xi \right) \, di
\]

\[
= \left[ \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} \hat{\xi} (\theta_0, \epsilon) \right] \int_0^1 g_i \left( T^{\theta_0, \epsilon_0} + \Delta T^\xi (\theta_0, \epsilon_0) \right) \frac{\partial}{\partial \xi} \bigg|_{\xi=\hat{\xi}(\theta_0, \epsilon_0)} \Delta T^\xi \left( z_i \left( T^{\theta_0, \epsilon_0} + \Delta T^\xi (\theta_0, \epsilon_0) \right) , \xi \right) \, di
\]

\[
= 0,
\]

(B.20)

where the first equality follows from \([A.51]\); the second equality from Fact \([1]\) and Lemma \([B.3]\) which imply that, when \(i \in [0, i_1]\), \(\frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} T \left( z_i \left( \hat{T}^{\theta_0, \epsilon_0} \right), \theta_0, \epsilon \right) = 0\); the third equality follows from \([A.51]\); the fourth equality follows from Lemma \([B.3]\) and the fact that, for all \(\xi \in \Xi\), the support of \(\Delta T^\xi\) is contained in \([0, \hat{z}_1]\), so that the integrand in the expression following the fourth equality is zero when \(i \in [i_1, 1]\); and the last equality follows from \([A.47]\).

Analyzing the second term:

\[
D = \int_{i_1}^{i_1} g_i \left( T^{\theta_0, \epsilon_0} \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} T \left( z_i \left( T^{\theta_0, \epsilon_0} \right), \theta_0, \epsilon \right) \, di
\]

\[
= \int_{i_1}^{i_1} g_i \left( T^{\theta_0, \epsilon_0} \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} T \left( z_i \left( T^{\theta_0, \epsilon_0} \right), \theta_0, \epsilon \right) \, di
\]

\[
= \int_0^{i_1} g_i \left( T^{\theta_0, \epsilon_0} \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} T \left( z_i \left( T^{\theta_0, \epsilon_0} \right), \theta_0, \epsilon \right) \, di
\]

\[
= 0,
\]

(B.21)

where the first equality follows from Lemma \([B.3]\) \([A.51]\), and the fact that, for all \(\xi \in \Xi\), the support of \(\Delta T^\xi\) is contained in \([0, \hat{z}_1]\), so that \(T \left( z_i \left( \hat{T}^{\theta_0, \epsilon_0} \right), \theta_0, \epsilon \right) = \hat{T} \left( z_i \left( \hat{T}^{\theta_0, \epsilon_0} \right), \theta_0, \epsilon \right)\) when \(i \in [i_1, 1]\); the second follows from the fact that, by \([A.49]\) and \(\Delta T^0 \equiv 0\), \(T^{\theta_0, \epsilon_0} = \hat{T}^{\theta_0, \epsilon_0}\); the third equality follows from the fact that, by Fact \([1]\) and Lemma \([B.3]\) \(\frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} T \left( z_i \left( T^{\theta_0, \epsilon_0} \right), \theta_0, \epsilon \right) = 0\) when \(i \in [0, i_1]\); and the last equality follows from the fact that \((T^{\theta, \epsilon})\) satisfies \([23]\).

Putting together \([B.19]\), \([B.20]\), and \([B.21]\), we see that \((\hat{T}^{\theta, \epsilon})\) satisfies \([23]\).
Next observe that:

\[
\frac{d}{d\theta} \bigg|_{\theta=\theta_0} \int_0^1 g_i \left( \hat{T}^{\theta,\epsilon_0} \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} \hat{T} \left( z_i \left( \hat{T}^{\theta,\epsilon_0} \right), \theta, \epsilon \right) \, d\epsilon
\]

\[
= \frac{d}{d\theta} \bigg|_{\theta=\theta_0} \int_0^1 g_i \left( \hat{T}^{\theta,\epsilon_0} \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} \hat{T} \left( z_i \left( \hat{T}^{\theta,\epsilon_0} \right), \theta, \epsilon \right) \, d\epsilon
\]

\[
+ \frac{d}{d\theta} \bigg|_{\theta=\theta_0} \int_1^{i_1} g_i \left( \hat{T}^{\theta,\epsilon_0} \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} \hat{T} \left( z_i \left( \hat{T}^{\theta,\epsilon_0} \right), \theta, \epsilon \right) \, d\epsilon.
\]

\[\text{(B.22)}\]

Analyzing the first term:

\[
E = \frac{d}{d\theta} \bigg|_{\theta=\theta_0} \int_0^1 g_i \left( \hat{T}^{\theta,\epsilon_0} \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} T \left( z_i \left( \hat{T}^{\theta,\epsilon_0} \right), \theta, \epsilon \right) \, d\epsilon
\]

\[
+ \frac{d}{d\theta} \bigg|_{\theta=\theta_0} \left[ \left( \frac{\partial}{\partial \epsilon} \xi \left( \theta, \epsilon \right) \right) \int_0^1 g_i \left( \hat{T}^{\theta,\epsilon_0} \right) \frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} \Delta T^\xi \left( z_i \left( \hat{T}^{\theta,\epsilon_0} \right), \xi \right) \, d\epsilon \right]
\]

\[
= \frac{d}{d\theta} \bigg|_{\theta=\theta_0} \left[ \left( \frac{\partial}{\partial \epsilon} \xi \left( \theta, \epsilon \right) \right) \int_0^1 g_i \left( T^{\theta,\epsilon_0} + \Delta T^\xi \left( \theta, \epsilon_0 \right) \right) \right.
\]

\[
\times \frac{\partial}{\partial \epsilon} \bigg|_{\xi=\xi \left( \theta, \epsilon_0 \right)} \Delta T^\xi \left( z_i \left( T^{\theta,\epsilon_0} + \Delta T^\xi \left( \theta, \epsilon_0 \right) \right), \xi \right) \, d\epsilon \right]
\]

\[
= \frac{d}{d\theta} \bigg|_{\theta=\theta_0} \left[ \left( \frac{\partial}{\partial \epsilon} \xi \left( \theta, \epsilon \right) \right) \int_0^1 g_i \left( T^{\theta,\epsilon_0} + \Delta T^\xi \left( \theta, \epsilon_0 \right) \right) \right.
\]

\[
\times \frac{\partial}{\partial \epsilon} \bigg|_{\xi=\xi \left( \theta, \epsilon_0 \right)} \Delta T^\xi \left( z_i \left( T^{\theta,\epsilon_0} + \Delta T^\xi \left( \theta, \epsilon_0 \right) \right), \xi \right) \, d\epsilon \right]
\]

\[
= 0,
\]

where the first equality follows from \((A.51)\); the second equality from Fact 1 and Lemma \([B.3]\) which imply that \(\frac{\partial}{\partial \epsilon} \bigg|_{\epsilon=\epsilon_0} T \left( z_i \left( \hat{T}^{\theta,\epsilon_0} \right), \theta, \epsilon \right) = 0, \forall \theta \in \Theta^\prime, \) when \(i \in \{0, i_1\}\); the third equality follows from \((A.51)\); the fourth equality follows from Lemma \([B.3]\) and the fact that, for all \(\xi \in \Xi,\) the support of \(\Delta T^\xi\) is contained in \([0, \hat{z}_i]\), so that the integrand in the expression following the fourth equality is zero, for all values of \(\theta\) in \(\Theta^\prime, \) when \(i \in \[i_1, 1\]\); and the last equality follows from \((A.47)\).
Analyzing the second term:

\[
F = \frac{d}{d\theta}|_{\theta=\theta_0} \int_{i_1}^{1} g_i \left( \hat{T}^{\theta,\epsilon_0} \right) \frac{\partial}{\partial \epsilon} \Bigg|_{\epsilon=\epsilon_0} T \left( z_i \left( \hat{T}^{\theta,\epsilon_0} \right), \theta, \epsilon \right) \, di \\
+ \frac{d}{d\theta}|_{\theta=\theta_0} \left[ \left( \frac{\partial}{\partial \epsilon} \hat{T}^{\theta,\epsilon_0} \right) \int_{i_1}^{1} g_i \left( \hat{T}^{\theta,\epsilon_0} \right) \frac{\partial}{\partial \xi} \Bigg|_{\xi=\xi(i_0,\epsilon_0)} \Delta T^\xi \left( z_i \left( \hat{T}^{\theta,\epsilon_0} \right), \xi \right) \, di \right]
\]

\[
= \frac{d}{d\theta}|_{\theta=\theta_0} \int_{i_1}^{1} g_i \left( T^{\theta,\epsilon_0} \right) \frac{\partial}{\partial \epsilon} \Bigg|_{\epsilon=\epsilon_0} T \left( z_i \left( T^{\theta,\epsilon_0} \right), \theta, \epsilon \right) \, di \\
= \frac{d}{d\theta}|_{\theta=\theta_0} \int_{0}^{1} g_i \left( T^{\theta,\epsilon_0} \right) \frac{\partial}{\partial \epsilon} \Bigg|_{\epsilon=\epsilon_0} T \left( z_i \left( T^{\theta,\epsilon_0} \right), \theta, \epsilon \right) \, di
\]

< 0,

where the first equality follows from (A.51); the second equality follows from Lemma [B.3] (A.51), and the fact that, for all \( \xi \in \Xi \), the support of \( \Delta T^\xi \) is contained in \([0, z_1] \), so that

\[
\frac{\partial}{\partial \epsilon}|_{\epsilon=\epsilon_0} \Delta T^\xi \left( z_i \left( \hat{T}^{\theta,\epsilon_0} \right), \xi \right) = 0, \forall \theta \in \Theta'', \text{ when } i \in [i_1, 1]; \text{ the third equality follows from Lemma [B.3] (A.51) and (B.14)-(B.15); the fourth equality follows from the fact that, by Fact 1 \( \frac{\partial}{\partial \epsilon}|_{\epsilon=\epsilon_0} T \left( z_i \left( \hat{T}^{\theta,\epsilon} \right), \theta, \epsilon \right) = 0, \forall \theta \in \Theta'', \text{ when } i \in [0, i_1]; \text{ and the inequality follows from the fact that } \left( \hat{T}^{\theta,\epsilon} \right) \text{ satisfies (24).}
\]

Putting together (B.22), (B.23), and (B.24), it follows that \( \left( \hat{T}^{\theta,\epsilon} \right)_{\theta \in \Theta'', \epsilon \in E''} \) satisfies (22)-(24), completing the proof of the lemma.

\[
\square
\]

### B.3 Proof of Lemma [B.1]

Let \( T \in \hat{T} \). Let \( \Delta T_1 \) and \( \Delta T_2 \) be smooth tax reforms and let \( r_1 \) and \( r_2 \) be real numbers. Then

\[
\frac{d}{dz}|_{z=0} \frac{z_i \left( T + \epsilon \left( r_1 \Delta T_1 + r_2 \Delta T_2 \right) \right)}{\Delta T_1 \left( z_i \left( T \right) \right) + \Delta T_2 \left( z_i \left( T \right) \right)} = - r_1 \frac{\Delta T_1 \left( z_i \left( T \right) \right) + \Delta T_2 \left( z_i \left( T \right) \right)}{\Delta T_1 \left( z_i \left( T \right) \right) + \Delta T_2 \left( z_i \left( T \right) \right)} + r_2 \frac{\Delta T_2 \left( z_i \left( T \right) \right)}{\Delta T_1 \left( z_i \left( T \right) \right) + \Delta T_2 \left( z_i \left( T \right) \right)}
\]

(B.25)

where the first and third equalities follow from applying the implicit function theorem to the first order conditions of agent \( i \)'s optimization problem when facing tax policies \( T + \epsilon \left( r_1 \Delta T_1 + r_2 \Delta T_2 \right) \), \( T + \epsilon \Delta T_1 \), and \( T + \epsilon \Delta T_2 \), and \( \Delta T_1 \left( z \right) \) and \( \Delta T_2 \left( z \right) \), are, respectively, the derivatives of \( \Delta T_1 \left( z \right) \)
and $\Delta T_2(z)$. Next, observe that

$$
\frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} R(T + \varepsilon (r_1 \Delta T_1 + r_2 \Delta T_2))
$$

$$
= \int [r_1 \Delta T_1(z_i(T)) + r_2 \Delta T_2(z_i(T))] \, di + \int T'(z_i(T)) \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} z_i(T + \varepsilon (r_1 \Delta T_1 + r_2 \Delta T_2)) \, di
$$

$$
= \int [r_1 \Delta T_1(z_i(T)) + r_2 \Delta T_2(z_i(T))] \, di
$$

$$
+ \int T'(z_i(T)) \left[ r_1 \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} z_i(T + \varepsilon \Delta T_1) + r_2 \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} z_i(T + \varepsilon \Delta T_2) \right] \, di
$$

$$
= r_1 \left[ \int \Delta T_1(z_i(T)) \, di + \int T'(z_i(T)) \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} z_i(T + \varepsilon \Delta T_1) \right]
$$

$$
+ r_2 \left[ \int \Delta T_2(z_i(T)) \, di + \int T'(z_i(T)) \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} z_i(T + \varepsilon \Delta T_2) \right]
$$

$$
= r_1 \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} R(T + \varepsilon \Delta T_1) + r_2 \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} R(T + \varepsilon \Delta T_2),
$$

where the second equality follows from (B.25). □

**B.4 Proof of Lemma B.2**

I begin by stating some useful background facts and then proceed to prove the lemma.

**B.4.1 Background facts**

Choose $T \in \mathcal{J}$, let $z_0 = z_0(T)$, and $z_1 = z_1(T)$. Define the function $\zeta : I \rightarrow Z$ by $\zeta(i) = z_i(T)$, and let $i = \zeta^{-1}$ be the inverse of $\zeta$ so that $i(z) = i$ if and only if $z_i(T) = z$. It follows from our assumptions in Section 6.1 that $\zeta(0) = z_0(T) > 0$ and $\zeta(i)$ is strictly increasing in $i$. Let $H$ be the cumulative distribution over incomes induced by tax policy $T$. Then, recalling that agents are uniformly distributed on the interval $I = [0, 1]$, it follows that $H(z) = 0$ for all $z \in Z$ such that $z < z_0$; $H(z) = i(z)$ for all $z \in Z$ with $z_0 \leq z \leq z_1$; and $H(z) = 1$ for all $z \in Z$ with $z_0 < z$. So if $h$ is the density corresponding to the cumulative distribution $H$, we have $h(z) = H'(z) = i'(z) = \frac{1}{\zeta'(i(z))}$ for all $z \in [z_0, z_1]$ and $h(z) = 0$ for all $z \in Z$ with $z \notin [z_0, z_1]$.

### Footnote

31 Strictly speaking, $h(z)$ is, respectively, the right- and left-derivative of $H(z)$ at $z = z_0$ and $z = z_1$, and we have $h(z_0) = \frac{1}{\zeta'(z_0)} = \frac{1}{\zeta'(0)}$ and $h(z_1) = \frac{1}{\zeta'(1)}$.  

54
Observe, using a change of variables, that:
\[
\frac{d}{dz} \bigg|_{z=0} R(T + \varepsilon \Delta T) = \int_0^1 \Delta T(z_i(T)) \, dz - \int_0^1 T'(z_i(T)) \frac{\Delta T'(z_i(T))}{T''(z_i(T)) + v''(z_i(T))} \, dz
\]
\[
= \int_{z_1}^{z_0} \Delta T(z) \, h(z) \, dz - \int_{z_0}^{z_1} T'(z) \frac{\Delta T'(z)}{T''(z) + v''(z_i(z))} \, h(z) \, dz
\]
\[
= \int_{z_0}^{z_1} \Delta T(z) \, h(z) \, dz - \int_{z_0}^{z_1} \Delta T'(z) \, k_T(z) \, dz
\]
where \(\Delta T'\) is the derivative of \(\Delta T\), \(v''(z)\) is \(v''(z)\) evaluated at \(i = i(z)\), and the second equality follows from applying the implicit function theorem to the first order condition for an agent’s optimization problem when facing tax policy \(T + \varepsilon \Delta T\) around \(\varepsilon = 0\) and
\[
k_T(z) = \frac{T'(z) \, h(z)}{T''(z) + v''(z_i(z))}, \quad \forall z \in [z_0, z_1]. \tag{B.27}
\]
Note that we include the subscript \(T\) in \(k_T\) to express the dependence of \(k_T\) on the tax policy \(T\) through the terms \(T'(z)\) and \(T''(z)\). It follows from the fact that \(\forall T \in T, \forall i \in I, \frac{\partial^2}{\partial z_i^2} \hat{T}(z_i(T)) < 0\) (see Section 6.1) that the denominator on the right hand side of (B.27) is positive for all \(z \in [z_0, z_1]\). Moreover the assumptions on \(v_i\) and \(y_i\) in Section 6.1 imply that \(\zeta'(i) > 0, \forall i \in I\), and hence that \(h(z) > 0, \forall z \in [z_0, z_1]\). It then follows that:
\[
T'(z_0) \neq 0 \Rightarrow k_T(z_0) \neq 0. \tag{B.28}
\]

### B.4.2 Main argument

Again, let \(z_0 = z_0(T)\) and \(z_1 = z_1(T)\). Choose \(z_\ast\) such that \(z_0 < z_\ast \leq z_1\). Let \(z_0 = z_0(T)\). As \(T \in T\), it follows from the assumptions in Section 6.1 that \(z_0 > 0\). Choose \(z_-\) so that \(0 < z_- < z_0\). Consider a smooth tax reform \(\Delta \hat{T}_1\) with \(\Delta \hat{T}_1(z) = 2, \forall z \in [0, z_-]\), \(\Delta \hat{T}_1(z) < 0, \forall z \in (z_\ast, z_\ast, \Delta \hat{T}_1(z_0) = 1, \text{ and } \Delta \hat{T}_1(z) = 0, \forall z \in [z_\ast, +\infty)\). So the smooth tax reform \(\Delta \hat{T}_1\) equal to 2 until \(z = z_-\), at which point it falls, passing through \(\Delta \hat{T}_1 = 1\) when \(z = z_0\), and reaching \(\Delta \hat{T}_1 = 0\) at \(z = z_\ast\) and remains at zero thereafter.

For each \(\gamma \in [1, +\infty)\), define \(z^\gamma\), \(z^\gamma_\ast\) by \(\gamma (z^\gamma - z_0) + z_0 = z_-\), \(\gamma (z^\gamma_\ast - z_0) + z_0 = z_\ast\). For \(\gamma \geq 1\), we have \(z^\gamma < z_0 < z^\gamma_\ast\); and \(z^\gamma \uparrow z_0\) and \(z^\gamma \downarrow z_0\) as \(\gamma \uparrow +\infty\). Define \(i^\gamma\) by the condition \(z_i^\gamma(T) = z^\gamma_i\); that such an \(i^\gamma\) exists follows from Lemma 6.1. Using assumptions in Section 6.1 we have \(i^\gamma \downarrow 0\) as \(\gamma \uparrow +\infty\).
Define the tax reform $\Delta T_1^\gamma$ by

$$
\Delta T_1^\gamma (z) = \begin{cases} 
2, & \text{if } z \in [0, z_-^\gamma], \\
\hat{\Delta} T_1 (\gamma (z - z_0) + z_0), & \text{if } z \in (z_-^\gamma, z_+^\gamma), \\
0, & \text{if } z \in [z_+^\gamma, +\infty). 
\end{cases}
$$

Using the properties of $\hat{\Delta} T_1$, it is straightforward to verify that, for all $\gamma \geq 1$, $\Delta T_1^\gamma$ is a smooth function of $z$. So $\Delta T_1^\gamma$ is similar to $\hat{\Delta} T_1$, except that in the former $z_-^\gamma$ and $z_+^\gamma$ play the roles of $z_-$ and $z_+$ in the latter. For $\gamma > 1$, $\Delta T_1^\gamma$ falls more steeply than $\hat{\Delta} T_1$ near $z = z_0$. Observe that, for all $\gamma \geq 1$, $[0, z_+^\gamma]$ is the support of both $\Delta T_1^\gamma$, so that the support of $\Delta T_1^\gamma$ is contained in $[0, z_+]$.

**Lemma B.5.** Assume, as above, that $T' (z_0) \neq 0$. Then $\lim_{\gamma \to \infty} \left. \frac{d}{d\gamma} \right|_{\gamma = 0} R \left( T + \varepsilon \Delta T_1^\gamma \right) \neq 0$.

Lemma B.5 is proven in Section B.4.3.

Since we are assuming that $T' (z_0) \neq 0$, it follows from Lemma B.5 that there exists a tax reform $\hat{\Delta} T_2$ with support contained in $[0, z_+]$ such that $\Delta T_2 (z) \geq 0, \forall z \in Z$, and

$$
\frac{d}{d\varepsilon} \left|_{\varepsilon = 0} R \left( T + \varepsilon \Delta T_2 \right) \right. \neq 0.
$$

(B.30)

In particular, we can choose $\Delta T_2 = \Delta \hat{T}^\gamma_1$ for some sufficiently large $\gamma_0$. However, for our purposes, it is not important whether $\Delta T_2 = \Delta \hat{T}^\gamma_1$ for some sufficiently large (fixed) $\gamma_0$; it matters only that is has the properties we have just ascribed to it.

It follows from Lemma B.1 and (B.30) that for each $\gamma > 1$, that there exists $r_\gamma$ such that

$$
\frac{d}{d\varepsilon} \left|_{\varepsilon = 0} R \left( T + \varepsilon \left( \Delta T_1^\gamma - r_\gamma \Delta \hat{T}_2 \right) \right) \right. = 0.
$$

(B.31)

It follows from Lemma B.1 (B.30), and (B.31) that

$$
r_\gamma = \frac{\frac{d}{d\varepsilon} \left|_{\varepsilon = 0} R \left( T + \varepsilon \Delta T_1^\gamma \right) \right.}{\frac{d}{d\varepsilon} \left|_{\varepsilon = 0} R \left( T + \varepsilon \Delta \hat{T}_2 \right) \right.}.
$$

(B.32)

The negative of the marginal welfare effect of a small tax reform in direction $\Delta T_1^\gamma - r_\gamma \Delta \hat{T}_2$ is

$$
W_\gamma = \int g_i (T) \Delta T_1^\gamma (z_i (T)) \, di - r_\gamma \int g_i (T) \Delta \hat{T}_2 (z_i (T)) \, di.
$$

Observe that because (i) the function $i \mapsto \Delta T_1^\gamma (z_i (T))$, whose domain is $[0, 1]$, has support $[0, i^\gamma]$ and $i^\gamma \downarrow 0$ as $\gamma$ approaches infinity and (ii) $\Delta T_1^\gamma (z)$ is bounded between 0 and 2, for all $z$, it follows that $\int g_i (T) \Delta T_1^\gamma (z_i (T)) \, di \to 0$ as $\gamma \to \infty$. Note that because $\Delta \hat{T}_2 (z) \geq 0, \forall z \in Z$, and $\Delta \hat{T}_2$ satisfies (B.30), there must be a positive measure set of agents $i$ such that $\Delta \hat{T}_2 (z_i (T)) > 0$. It
follows that \( \int g_1(T) \Delta \hat{T}_2(z_1(T)) \, dt > 0. \) Lemma \( \text{(B.5)} \) and \( \text{(B.32)} \) imply that \( \lim_{\gamma \to \infty} r_\gamma \neq 0 \). It now follows from the results of the previous paragraph that if \( \gamma \) is sufficiently large then \( \psi_\gamma \neq 0. \) Then choose such a sufficiently large \( \gamma. \) If \( \psi_\gamma < 0, \) then define \( \Delta T_1 = \Delta T_1^\gamma \) and \( \Delta T_2 = \gamma \Delta \hat{T}_2; \) and if \( \psi_\gamma > 0, \) define \( \Delta T_1 = -\Delta T_1^\gamma \) and \( \Delta T_2 = -\gamma \Delta \hat{T}_2. \) In either case define \( \Delta T = \Delta T_1 - \Delta T_2. \) In both cases, we have \( \int g_1(T) \Delta T(z_1(T)) \, dt < 0 \) and, appealing to Lemma \( \text{(B.1)} \) \( \text{(B.30)} \), \( \text{(B.31)}, \) \( \frac{d}{d\varepsilon}|_{\varepsilon=0} R(T + \varepsilon \Delta T) = 0 \) and \( \frac{d}{d\varepsilon}|_{\varepsilon=0} R(T + \varepsilon \Delta T_2) \neq 0. \) Finally note the support of \( \Delta T, \Delta T_1, \) and \( \Delta T_2 \) are all contained in \([0, z_4]. \) We have now established all of the properties required by Lemma \( \text{(B.2)} \). □

**B.4.3 Proof of Lemma \( \text{(B.5).} \)**

It follows from \( \text{(B.26)} \) and the fact that the support of \( \Delta T_1^\gamma \) is \([0, z_4^\gamma] \) that

\[
\frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} R(T + \varepsilon \Delta T_1^\gamma) = \int_{z_0}^{z_2^\gamma} \Delta T_1^\gamma (z) h(z) \, dz - \int_{z_0}^{z_2^\gamma} \frac{d}{dz} \Delta T_1^\gamma (z) k_T(z) \, dz, \tag{B.33}
\]

where \( \frac{d}{dz} \Delta T_1^\gamma (z) \) is the derivative of \( \Delta T_1^\gamma (z). \) Because \( z_4^\gamma \downarrow z_0 \) as \( \gamma \to \infty \) and \( \Delta T^\gamma (z) \) is bounded between 0 and 2 for all \( z, \)

\[
\lim_{\gamma \to \infty} \int_{z_0}^{z_2^\gamma} \Delta T_1^\gamma (z) h(z) \, dz = 0. \tag{B.34}
\]

Since \( \frac{d}{dz} \Delta T_1^\gamma (z) \leq 0, \forall z \in Z, \) it follows from the preceding that, if \( \gamma \) is sufficiently large, we have:

\[
\left( \max_{z \in [z_0, z_2^\gamma]} k_T(z) \right) \times \int_{z_0}^{z_2^\gamma} \frac{d}{dz} \Delta T_1^\gamma (z) \, dz \leq \int_{z_0}^{z_2^\gamma} \frac{d}{dz} \Delta T_1^\gamma (z) k_T(z) \, dz \leq \left( \min_{z \in [z_0, z_2^\gamma]} k_T(z) \right) \times \int_{z_0}^{z_2^\gamma} \frac{d}{dz} \Delta T_1^\gamma (z) \, dz. \tag{B.35}
\]

Next observe that

\[
\int_{z_0}^{z_2^\gamma} \frac{d}{dz} \Delta T_1^\gamma (z) \, dz = \int_{z_0}^{z_2^\gamma} \gamma \Delta \hat{T}_1'(\gamma [z - z_0] + z_0) \, dz = \int_{z_0}^{z_2^\gamma} \Delta \hat{T}_1'(\tilde{z}) \, d\tilde{z} = \Delta \hat{T}_1(z_4) - \hat{T}_1(z_0) = -1, \tag{B.36}
\]

where \( \Delta \hat{T}_1'(\tilde{z}) \) is the derivative of \( \Delta \hat{T}_1'(\tilde{z}) \) and the second equality uses the change of variables \( z \mapsto \tilde{z} = \gamma [z - z_0] + z_0. \) Next observe that as \( k \) is smooth and \( z^\gamma \downarrow z_0 \) and \( \gamma \to \infty, \)

\[
\lim_{\gamma \to \infty} \max_{z \in [z_0, z_2^\gamma]} k_T(z) = k_T(z_0) \quad \text{and} \quad \lim_{\gamma \to \infty} \min_{z \in [z_0, z_2^\gamma]} k_T(z) = k_T(z_0). \tag{B.37}
\]

\[\text{32}\text{Observe that, from } \text{(B.26), } \frac{d}{d\varepsilon}|_{\varepsilon=0} R(T + \varepsilon \Delta \hat{T}_2) = \int_{z_0}^{z_1} \Delta \hat{T}_2(z) h(z) \, dz - \int_{z_0}^{z_1} \Delta T'(z) k_T(z) \, dz, \text{ which is finite, so the denominator in } \text{(B.32)} \text{ is finite as well.} \]
It follows from (B.33), (B.34), (B.35), (B.36), and (B.37) that
\[
\lim_{\gamma \to \infty} \frac{d}{dz} \bigg|_{z=0} R(T + \varepsilon \Delta T^\gamma_1) = k_T(z_0).
\]

It follows from (B.28) and the assumption that $T'(z_0) \neq 0$ that $k_T(z_0) \neq 0$, which completes the proof. □

B.5 Proof of Lemmas B.3 and B.4

In this section, I prove Lemmas B.3 and B.4. For convenience, I state the following fact, which follows from assumptions in Section 6.1.

**Fact 2.** For all $T \in \hat{T}$ and, for all agents $i$, there exists a unique optimal income $z_i(T)$ for $i$ when facing $T$ and $z_i(T)$ is characterized by $i$’s first order condition in the sense that if $\frac{d}{dz} \hat{U}_i^T(z) = 0$, then $z = z_i(T)$.

To simplify notation, I will write $\bar{T}^{\theta,\epsilon,\xi} = T^\theta,\epsilon + \Delta T^\xi$. Fix some $\theta' \in \Theta'$, $\epsilon' \in \Theta'$, and $\epsilon' \in \Xi$. Recall that $\hat{z}_1 \in (z_0(T^{\theta_0,\epsilon_0}), z_1(T^{\theta_0,\epsilon_0}))$, and that $i_1$ is the unique agent in $I$ such that $z_{i_1}(T^{\theta_0,\epsilon_0}) = \hat{z}_1$. Let $I_0 := [0, i_1)$ and $I_1 := (i_1, 1]$.

Because, by construction of $T^{\theta,\epsilon}, T^{\theta_0,\epsilon_0}(z) = T^{\theta',\epsilon'}(z)$, for all $z \leq \hat{z}_1$, and both $T^{\theta_0,\epsilon_0}$ and $T^{\theta',\epsilon'}$ are smooth, $\frac{d}{dz} T^{\theta_0,\epsilon_0}(\hat{z}_1) = \frac{d}{dz} T^{\theta',\epsilon'}(\hat{z}_1)$. So, because $\hat{z}_1 = z_{i_1}(T^{\theta_0,\epsilon_0})$,
\[
\frac{d}{dz} \hat{U}_{i_1}^{T^{\theta_0,\epsilon_0}}(\hat{z}_1) = \frac{d}{dz} \hat{U}_{i_1}^{T^{\theta_0,\epsilon_0}}(\hat{z}_1) = 0.
\] (B.38)

Using the facts that, by construction of $T^{\theta,\epsilon}, T^{\theta,\epsilon}(z)$ does not depend on $\theta$ or $\epsilon$ when $z \leq \hat{z}_1$, and that the support of $\Delta T^\xi$ is contained in $[0, \hat{z}_1]$, it follows that
\[
\forall z \leq \hat{z}_1, \quad \bar{T}^{\theta',\epsilon',\xi'}(z) = \bar{T}^{\theta_0,\epsilon_0,\xi'}(z),
\](B.39)
\[
\forall z \geq \hat{z}_1, \quad \bar{T}^{\theta',\epsilon',\xi'}(z) = \bar{T}^{\theta',\epsilon'}(z).
\](B.40)

Using (B.40) and the smoothness of the tax policies $\bar{T}^{\theta',\epsilon',\xi'}$ and $\bar{T}^{\theta',\epsilon'}$, it follows that
\[
\frac{d}{dz} \hat{U}_{i_1}^{\bar{T}^{\theta',\epsilon',\xi'}}(\hat{z}_1) = \frac{d}{dz} \hat{U}_{i_1}^{\bar{T}^{\theta',\epsilon'}}(\hat{z}_1).
\](B.41)

It follows from Fact 2, (B.41), and (B.38), and the fact that $T^{\theta_0,\epsilon_0}$ and $\bar{T}^{\theta',\epsilon',\xi'}$ belong to $\hat{T}$ (the latter was established in Section B.1) that
\[
\hat{z}_1 = z_{i_1}(T^{\theta_0,\epsilon_0}) = z_{i_1}(\bar{T}^{\theta',\epsilon',\xi'})
\](B.42)

The following fact follows from our assumptions in Section 6.1.

**Fact 3.** For all $T \in \hat{T}$, the map $i \mapsto z_i(T)$ is strictly increasing in $i$.  

58
Proof. Since $T_{\theta'} \in \hat{\mathcal{T}}$, \eqref{eq:B.42} and Fact \ref{fact:B.2} together establish Lemma \ref{lem:B.3}.

It follows from \eqref{eq:B.39} and Lemma \ref{lem:B.3} that, for all $i \in [0, i_1]$, \[ \frac{d}{d z} \hat{U}_i \left( z_i \left( T_{\theta', \xi'} \right) \right) = 0. \] So, using the fact that $T_{\theta'}$ and $T_{\theta, \xi}$ belong to $\hat{\mathcal{T}}$, it follows from Fact \ref{fact:B.2} that \eqref{eq:B.12} holds. Similarly, it follows from \eqref{eq:B.40} and Lemma \ref{lem:B.3} that, for all $i \in [i_1, 1]$, \[ \frac{d}{d z} \hat{U}_i \left( z_i \left( T_{\theta', \xi'} \right) \right) = 0. \] So, using the fact that $T_{\theta'}$ and $T_{\theta, \xi}$ belong to $\hat{\mathcal{T}}$, it follows from Fact \ref{fact:B.2} that \eqref{eq:B.14} holds. If follows immediately from the definition of $g_i(T)$ (see Sections \ref{sec:2.1}, \ref{sec:2.2}, \ref{sec:12}, and \ref{sec:14}) that \eqref{eq:B.13} and \eqref{eq:B.15} hold. We have now established Lemma \ref{lem:B.4}.

\eqref{eq:B.10}-\eqref{eq:B.11} follow from \eqref{eq:B.12}-\eqref{eq:B.13} in Lemma \ref{lem:B.3} (note that, in \eqref{eq:B.10}-\eqref{eq:B.11}, $T = T_{\theta, \xi}$), and the fact that the support of $\Delta T_{\xi}$ is contained in $[0, \hat{z}_1]$. \hfill \QED

C Additional lemmas

**Lemma C.1.** For all $T \in \hat{\mathcal{T}}$, $\{z_i(T) : i \in I\} = [z_0(T), z_1(T)]$, and the map $i \mapsto z_i(T)$ is strictly increasing.

Proof. Let $\hat{T} \in \hat{\mathcal{T}}$. It follows from our assumptions (see Section \ref{sec:6.1}) that $z_i(T)$ is characterized by the first order condition $1 - T' \left( z_i(T) \right) - v'_i \left( z_i(T) \right) = 0$. The smoothness of $T$ and $\left( z, i \right) \mapsto v_i(z)$ imply that the function $i \mapsto z_i(T)$ is smooth. It follows from the facts that (i) $v_i(z) = v(z, y_i) \forall z, \forall y$, (ii) $\frac{d^2}{d z d y} v(z, y) < 0, \forall z, \forall y$, and (iii) $\frac{d}{d t} y_i > 0, \forall i$, that the map $i \mapsto z_i(T)$ is strictly increasing (again see Section \ref{sec:6.1} for the preceding assumptions). Since $i \mapsto z_i(T)$ is continuous and strictly increasing on $I = [0, 1]$, $\{z_i(T) : i \in I\} = [z_0(T), z_1(T)]$. \hfill \QED

**Lemma C.2.** Let $T \in \hat{\mathcal{T}}$. For $j = 1, \ldots, n$, let $\Theta_j = [-\theta_j, \theta_j] \subseteq \mathbb{R}$, where $\theta_j > 0$, and let $\hat{\Theta} = \times_{j=1}^n \Theta_j$. Write $\hat{\theta} = (\theta_1, \ldots, \theta_j, \ldots, \theta_n)$. Let \[ (\Delta T^\hat{\theta})_{\hat{\theta} \in \hat{\Theta}} \] be a family of tax reforms such that the map $(z, \hat{\theta}) \mapsto \Delta T^\hat{\theta}(z)$ is smooth and $\Delta T^{(0,0,0,\ldots,0)} \equiv 0$. Then there exist $\theta_j^* \in \Theta_j$ with $\theta_j^* > 0$ for $j = 1, \ldots, n$ such that, for all $\hat{\theta} = (\theta_1, \ldots, \theta_j, \ldots, \theta_n) \in \hat{\Theta}$, if $|\theta_j| \leq \theta_j^*$ for $j = 1, \ldots, n$, then $T + \Delta T^\hat{\theta} \in \hat{\mathcal{T}}$.

Proof. Since $T \in \hat{\mathcal{T}}$, it follows that, for all agents $i$, $\frac{d^2}{d z^2} \hat{U}_i \left( z_i(T) \right) < 0$ (see Section \ref{sec:6.1}). Also, since $T \in \hat{\mathcal{T}}$, $z_i(T) > 0$, for all agents $i$. By the smoothness of the primitives and $T$, it follows that there is a neighborhood $N_i$ of the income $z_i(T)$ such that, for all $z_i \in N_i$, $\frac{d^2}{d z^2} \hat{U}_i \left( z_i \right) < 0$ and $z_i > 0$. For each $i$, let $\delta_i = \sup \left\{ \delta > 0 : z_i(T) - \delta > 0, \forall z_i \in (z_i(T) - \delta, z_i(T) + \delta), \frac{d^2}{d z^2} \hat{U}_i \left( z_i \right) < 0 \right\}$. We have $\delta_i > 0, \forall i$, and, moreover, the smoothness of the primitives and of $T$ implies that $i \mapsto \delta_i$ is smooth. Since a continuous function attains its minimum on a compact set, it follows that $\delta^* = \min \{ \delta_i : i \in [0, 1] \}$ exists and $\delta^* > 0$. For each $i$, define the neighborhood $N_i' = (z_i(T) - \frac{1}{2} \delta^*, z_i(T) + \frac{1}{2} \delta^*)$ and let $N_i'$ be the closure of $N_i'$. For each $\hat{\theta} \in \hat{\Theta}$, define $T^\hat{\theta} = T + \Delta T^\hat{\theta}$. Define $\gamma_i \left( z_i(T) \right) = \min_{z_i \in N_i'} \hat{U}_i \left( z_i \right)$ and $\gamma_i = \min_{i \in [0, 1]} \gamma_i$. As $T^{(0,0,\ldots,0)} = T + \Delta T^{(0,0,\ldots,0)} = T$, and, as $T \in \hat{\mathcal{T}}$, $\hat{U}_i \left( z_i \right)$ has a unique maximizer $z_i(T)$, it follows that, \footnote{Observe that $\Delta T^{(0,0,\ldots,0)} \equiv 0$, so that, setting $\xi = 0$, Lemma \ref{lem:B.3} implies that, when $i \in [i_1, 1]$, $z_i \left( T^\hat{\theta} \right) \geq \hat{z}_1$.}
for all $i$, $\gamma_i^{(0,0,\ldots,0)} > 0$, and hence, again because a continuous function attains its minimum on a compact set, $\gamma^{(0,0,\ldots,0)} > 0$. It follows from our smoothness assumptions that there exist $\theta'_j \in \Theta_j$ with $\theta'_j > 0$ for $j = 1, \ldots, n$ such for all $\bar{\theta} = (\theta_1, \ldots, \theta_j, \ldots, \theta_n) \in \bar{\Theta}$, if $|\theta_j| \leq \theta'_j$ for $j = 1, \ldots, n$, $\gamma^{\bar{\theta}} > 0$, so that, for all such $\bar{\theta}$, $\hat{U}_i^{T^{\bar{\theta}}}(z_i)$ does not have any maximizers $z_i$ outside of $N'_i$. Note that we have: $\forall i \in I, \forall z_i \in \bar{N}'_i, \max_{z_i \in \bar{N}'_i} \frac{d^2}{dz_i^2} \hat{U}_i^{T^{\bar{\theta}}}(z_i) < 0$. So $\max_{z_i \in \bar{N}'_i, |\theta_j|^2 \leq \theta''_j} \frac{d^2}{dz_i^2} \hat{U}_i^{T^{\bar{\theta}}}(z_i) < 0$. As the map $\bar{\theta} \mapsto \max_{z_i \in \bar{N}'_i} \frac{d^2}{dz_i^2} \hat{U}_i^{T^{\bar{\theta}}}(z_i)$ is continuous, it follows that there exist $\theta''_j \in \Theta_j$ with $\theta''_j > 0$ for $j = 1, \ldots, n$ such for all $\bar{\theta} = (\theta_1, \ldots, \theta_j, \ldots, \theta_n) \in \bar{\Theta}$, if $|\theta_j| \leq \theta''_j$ for $j = 1, \ldots, n$, then, for all agents $i$ and all $z_i \in \bar{N}'_i$, $\frac{d^2}{dz_i^2} \hat{U}_i^{T^{\bar{\theta}}}(z_i) < 0$, so that $\hat{U}_i^{T^{\bar{\theta}}}(z_i)$ is strictly convex on $\bar{N}'_i$, implying that $\hat{U}_i^{T^{\bar{\theta}}}(z_i)$ has a unique maximizer on $\bar{N}'_i$. It follows that if $\theta^*_j = \min \{\theta'_j, \theta''_j\}$ for $j = 1, \ldots, n$, then, then for all $\bar{\theta} = (\theta_1, \ldots, \theta_j, \ldots, \theta_n) \in \bar{\Theta}$, if $|\theta_j| \leq \theta^*_j$ for $j = 1, \ldots, n$, then for, all agents $i$, $\hat{U}_i^{T^{\bar{\theta}}}(z_i)$ has a unique maximizer $z_i(\bar{\theta}) > 0$, and, moreover, $\frac{d^2}{dz_i^2} \hat{U}_i^{T^{\bar{\theta}}}(z_i(\bar{\theta})) < 0$, so that $T^{\bar{\theta}} \in \mathcal{F}$. □