HCEO WORKING PAPER SERIES

Generalized Social Marginal Welfare Weights Imply Inconsistent Comparisons of Tax Policies

Itai Sher

Working Paper

2021-009

03/2021

HUMAN CAPITAL AND
ECONOMIC OPPORTUNITY
GLOBAL WORKING GROUP

The University of Chicago
1126 E. 59th Street Box 107
Chicago IL 60637

www.hceconomics.org

Generalized Social Marginal Welfare Weights Imply Inconsistent Comparisons of Tax Policies

Itai Sher*
University of Massachusetts Amherst

February 15, 2021

Abstract

This paper concerns Saez and Stantcheva's (2016) *generalized social marginal welfare weights* (GSMWW), which are used to aggregate losses and gains due to the tax system, while incorporating non-utilitarian ethical considerations. That approach evaluates local changes in tax policy without appealing to a global social objective. However, I argue that local comparisons between different tax systems implicitly entail global comparisons. Moreover, whenever welfare weights are not of a utilitarian kind, these implied global comparisons are inconsistent. Part of the motivation for the GSMWW approach is that it provides a way to incorporate broader ethical judgements into the evaluation of the tax system while preserving the Pareto principle. I suggest that the problems with the approach ought to spark a reconsideration of Pareto if one wants to represent broader values in formal policy analysis.

1 Introduction

This paper explores an approach to optimal tax proposed by Saez and Stantcheva (2016), the *generalized social marginal welfare weights* (GSMWW) approach. The purpose of the approach is to incorporate broader values into evaluation of optimal income tax. The approach evaluates tax policies *locally* by assigning welfare weights to individuals on the basis of various morally relevant factors.

It is instructive to contrast the GSMWW approach with the the standard approach which involves maximization of a utilitarian objective $\int u_i \mathrm{d}i$ subject to a revenue requirement. Small tax reforms ΔT–in other words, small local changes in the tax policy–can be evaluated by weighing the effects of those changes on different people ΔT_i by their marginal utility u_i'; bearing in mind that the ΔT_i's are additional payments, a tax reform is desirable if $\int u_i' \Delta T_i \mathrm{d}i < 0$. A necessary condition for a tax policy to be optimal optimal is that no locally revenue neutral tax reform is desirable; that is, for all locally revenue neutral ΔT, $\int u_i' \Delta T_i \mathrm{d}i = 0$.

*email: isher@umass.edu. I am grateful for helpful comments and discussions with Maya Eden, Louis Perrault, Paolo Piacquadio, Peter Sher, and Matt Weinzierl and to seminar audiences at UC Riverside and at the Welfare Economics and Economic Policy virtual seminar.

Instead of positing a social objective and deriving weights, the GMSWW approach directly posits welfare weights g_i without an objective. These weights need not depend only on marginal utility; they can depend on other moral considerations, such as fairness, libertarian values, equality of opportunity, poverty alleviation. Parallelling the utilitarian formulas, a tax reform is desirable if $\int g_i \Delta T_i \mathrm{d}i < 0$, and and a tax policy T is optimal if, at that policy, for all locally revenue neutral ΔT, $\int g_i \Delta T_i \mathrm{d}i = 0$; that is, at an optimal policy, there is no small revenue neutral desirable reform. As the local optimality condition under the GMSWW approach has a similar structure to the local optimality condition under the standard utilitarian approach, it can be used to derive a familiar formula

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

for the optimal marginal tax rates, where the term $\bar{G}(z)$ represents a normative judgement embodied in generalized welfare weights, and $\alpha(z)$ and $e(z)$ are empirical terms. This generalizes a standard optima tax formula (Saez 2001) in which standard welfare weights play the role of the generalized weights above. This gives the impression that there is a neat separation between between the positive and normative ingredients that go into the determination of optimal policy, and that one can essentially use tools and results of standard optimal tax to accommodate broader ethical values. I shall argue that this is not correct: The decomposition of positive and normative ingredients does not take the form (1) and to incorporate broader values, we must depart more radically from the standard optimal tax framework.

One attractive feature of the GSMWW approach is that it preserves the Pareto efficiency of the standard utilitarian approach. In particular, Saez and Stantcheva (2016) establish that any tax policy that is locally optimal according to their approach is also locally Pareto optimal.

In this paper, I criticize the GSMWW approach. I want to emphasize that I greatly admire the approach. Saez and Stantcheva (2016) is an inspiring attempts to incorporate broader values into welfare economics. However, the approach ultimately runs into problems, and the problems it faces are instructive.

In contemplating the apparent compatibility, within the GSMWW approach, between the Pareto principle and broader values, one might be initially puzzled in light of results from Sen (1970, 1979) and Kaplow and Shavell (2001) to the effect that there is a conflict between social evaluations that incorporate broader non-welfarist values and the Pareto principle. How are Saez and Stantcheva (2016) able to incorporate these broader values without running afoul of Pareto?

Saez and Stantcheva (2016) write "In our approach ... there is no social welfare objective primitive that the government maximizes." (p. 24) The current paper contends that it is the lack of such an objective that allows for efficiency to co-exist with broader values. It is important to think about what it means that the GSMWW approach does not correspond to maximizing any objective. The GSMWW approach provides information about whether locally one tax policy is better than another. Suppose that there is no global objective that is consistent with all of those

local comparisons. Then I would argue that one should conclude that the comparisons implied by GSMWW are not coherent. One way of seeing this is that it would imply that anyone who did have a coherent global ranking of tax policies would disagree with the judgements of GSMWW for some comparisons.

Saez and Stantcheva (2016) do not formalize the observation that GSMWW is not consistent with a global objective; in this paper, I do. I show that for some specifications of generalized social welfare weights, any binary relation that attempts to rationalize them will contain a cycle of the form: T_0 is better than T_1, which is better than T_2, which in turn is better than T_0 (where the T_j's are tax policies). Indeed, my main result–Theorem 1–shows that such cycles occur precisely when GSMWW are not of a utilitarian kind–that is, the GSMWW approach is inconsistent precisely in those cases when the GSMWW approach is supposed to go beyond the standard approach.

While the GSMWW approach is intended to only make local comparisons, it is not surprising these local rankings imply global comparisons; it is analogous to the fact that we can recover the function $f : \mathbb{R}^n \to \mathbb{R}$ up to a constant from its gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, which informs us about its local behavior.

The problem with the GSMWW approach is that, rather than working from foundations, it attempts to reverse engineer a solution that resembles the traditional approach. This would be analogous to directly modifying the optimality conditions derived from an optimization problem rather than modifying the original underlying problem. It is the optimization problem that gives significance to the optimality conditions, not the other way around. Modifying the optimality conditions directly might not lead to a coherent solution to any problem; in fact that is what I claim happens in this case.

If we really want to incorporate broader values into optimal tax policy, then it is unlikely that the solution will take the form (1) analogous to the traditional approach. We should expect an approach that incorporates broader values to depart more radically from the traditional approach. In particular, it may be that we are willing to trade off welfare against other moral considerations, meaning, for example, that everyone may be worse off, and only slightly worse off on average, while, say, fairness is much better satisfied, and we would deem that an improvement. In other words, taking diverse moral principles seriously may lead to a conflict with Pareto efficiency. This point is discussed further in Sections 3 and 8.

I should note that the argument of this paper does not preclude moral values that cannot be represented by some social objective. It may be for example that certain systems of obligations and permissions cannot be so represented. It may be that the *right*–what one ought to do–is really separate from the *good*–what is better or worse–in such a way that moral behavior is not a matter of maximizing some objective. What the argument of this paper precludes is a system that puts the good before the right, directing one to act so as to maximize a *local* objective that cannot be coherently extended to a global objective.

The outline of this paper is as follows. Section 2 presents the GSMWW framework. Section 3 explains informally why we might expect the GSMWW approach to run into problems, elaborating

on some of the points that I have made in this introduction. Section 4 defines what it means for a global relation to rationalize social marginal welfare weights. Section 5 presents a simple example in which welfare weights are rationalizable, and Section 6 presents a simple example in which welfare weights are not rationalizable. Section 7 presents my main result, which shows that generalized social welfare weights are rationalizable only when they take an essentially utilitarian form, so that they are not rationalizable precisely in those cases when the theory of GSMWW purports to go beyond the standard theory. Section 8 concludes with a discussion of the significance of the results.

2 Model

There is a continuum of agents in the interval $I = [0,1]$ distributed with Lebesgue measure λ. Each agent $i \in I$ has **observable characteristics** $x_i \in X$ and **unobservable characteristics** $y_i \in Y$, where X and Y are compact subsets of a Euclidean space. I assume that $i \mapsto x_i$ and $i \mapsto y_i$ have at most finitely many discontinuities. If it is impermissible to condition taxes on a certain characteristic, then we treat that characteristic as unobservable.

Let c_i be agent i's consumption and z_i be agent i's income. I assume that consumption can take values in $\mathbb{R}$ and income can take values in the set $Z = [0, \bar{z}]$ for some $\bar{z} > 0$.

Agent $i \in [0,1]$ has a utility function

$$u_i(c_i, z_i) = u(c_i - v_i(z_i)),$$

where

$$v_i(z_i) = v(z_i; x_i, y_i)$$

and $u : \mathbb{R} \to \mathbb{R}$ is a strictly increasing twice continuously differentiable function that is common across agents, and $v : Z \times X \times Y \to$ is continuous, and $\frac{\partial}{\partial z} v(z; x, y)$ and $\frac{\partial^2}{\partial z^2} v(z; x, y)$ exist and are continuous at all $(z; x, y) \in Z \times X \times Y$. $v(z_i; x_i, y_i)$ is interpreted as the cost in terms of consumption of earning income z_i given characteristics (x_i, y_i), and u transforms the utility representation $c_i - v_i(z_i)$ into a representation that is adequate for making utilitarian interpersonal comparisons. I assume that $\forall i, v_i'(z_i) > 1$, which implies that the maximum "possible" income $\bar{z}$ in Z is selected so that it is so large that no agent would actually choose it in the absence of taxes.

A **tax policy** is a function $T\colon Z \times X \to \mathbb{R}$, where $T(z; x)$ is the tax paid by citizens with income z and x is the agent's observable characteristics. $\mathscr{T}$ is the set of all tax policies. Formally, $\mathscr{T}$ is the set of all continuous functions T on $Z \times X$ such that $\frac{\partial}{\partial z} T(z; x)$ and $\frac{\partial^2}{\partial z^2} T(z; x)$ exist and are continuous at all $(z; x) \in Z \times X$.

Define the norm $\rho(T) = \max\left\{|T(z; x)| + \left|\frac{\partial}{\partial z} T(z; x)\right| + \left|\frac{\partial^2}{\partial z^2} T(z; x)\right| : (z; x) \in Z \times X\right\}$, and let $\mathscr{T}$ be endowed with the metric topology induced by ρ. Thus the notion of closeness for tax policies depends on the first two derivatives of T and not just its values.

Let $\mathscr{T}^\circ = \left\{T \in \mathscr{T} : \frac{\partial^2}{\partial z^2} T(z; x) > 0, \forall (z; x) \in Z \times X\right\}$. Thus $\mathscr{T}^\circ$ is the set of functions with

a positive second z derivative everywhere. It is easy to see that $\mathscr{T}^{\circ}$ is an open set in the metric topology induced by ρ.

We write

$$T_i(z_i) = T(z_i; x_i).$$

Given a tax policy T, we have $c_i = z_i - T_i(z_i)$. Define $z_i(T)$ to be i's optimal income when facing tax system T; that is,

$$z_i(T) \in \arg\max_{z_i} z_i - T_i(z_i) - v_i(z_i).$$

Observe that since u is strictly monotone, $z_i(T) \in \arg\max_{z_i} u(z_i - T_i(z_i) - v_i(z_i))$. Now define

$$U_i(T) = u(z_i(T) - T_i(z_i(T)) - v(z_i(T))), \tag{2}$$
$$\tilde{U}_i(T) = z_i(T) - T_i(z_i(T)) - v(z_i(T)). \tag{3}$$

Thus $U_i(T)$ is the utility induced by tax system T expressed in terms adequate for utilitarian interpersonal comparisons; and $\tilde{U}_i(T)$ is an alternative utility representation in dollar terms giving the consumption such that i would be indifferent between optimizing against tax system T and having consumption $\tilde{U}_i(T)$ without incurring the costs of earning income.

For any tax system T, define $R : \mathscr{T} \to \mathbb{R}$ by[1]

$$R(T) = \int T_i(z_i(T))\,\mathrm{d}i.$$

The novelty in the GSMWW approach is the way that tax systems are evaluated. Let $g(c_i, z_i; x_i, y_i)$ be the **generalized social welfare weight**. Thus, we assign a certain weight to each agent depending on their consumption c_i, their income z_i, and their characteristics x_i, y_i. Formally, a **system of generalized social welfare weights** is an integrable function $g : \mathbb{R} \times Z \times X \times Y \to \mathbb{R}$ such that

$$g(c_i, z_i; x_i, y_i) > 0, \forall c_i, z_i, x_i, y_i. \tag{4}$$

Let $\mathscr{G}$ be the set of all systems of generalized social welfare weights. 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 value of giving a dollar to each person i.

Under the utilitarian approach, the goal of a tax system T–or any other policy for that matter–is

[1]For $R(T)$ to be uniquely defined, we must assume that the optimal income $z_i(T)$ is unique for almost all i.

to maximize the sum of agent utilities

$$\int U_i(T)\,\mathrm{d}i. \tag{5}$$

The generalized social welfare weight approach attempts to bring into play more general normative considerations. Rather than having a global objective like (5), the social welfare weights approach is local: it looks at a tax policy and in considering small changes to the tax policy, it weighs the incremental dollars to different individuals according not just to their marginal utility, but according to other considerations (possibly including marginal utility for consumption), such as those involving fairness.

A key aspect of the approach is that there is *no global objective*. Given the tax system T to be evaluated, the local marginal welfare weight

$$g_i(T) = g_i\left(z_i(T) - T_i\left(z_i(T)\right), z_i(T)\right)$$

(local at T) is endogenously determined.

Some examples from Saez and Stantcheva (2016) to illustrate generalized social welfare weights are as follows:

- *Utilitarian weights*: $g_i(c_i, z_i) = u'\left(c_i - v_i(z_i)\right)$. These are the weights that arise out of the standard utilitarian framework. That is, the priority put on giving a dollar to any individual is proportional to its marginal utility.

- *Libertarian weights*: $g_i(c_i, z_i) = \tilde{g}(z_i - c_i) = \tilde{g}(t_i)$ where $\tilde{g}'(t_i) > 0$, where $t_i = z_i - c_i$ is the tax paid. That is, the more tax a person has already paid, the greater the weight placed on that person.

- *Libertarian-utilitarian mix*: $g_i(c_i, z_i) = \tilde{g}\left(c_i - v_i(z_i), z_i - c_i\right) = \tilde{g}(\tilde{u}_i, t_i)$ where $\tilde{u}_i = c_i - v_i(z_i)$ with $\frac{\partial \tilde{g}}{\partial \tilde{u}_i} < 0$ and $\frac{\partial \tilde{g}}{\partial t_i} > 0$; the first inequality can be interpreted as saying that weights are increasing in marginal utility for consumption (since $u_i'\left(c_i - v_i(z_i)\right)$ is decreasing in $c_i - v_i(z_i)$) and the second says that they are also increasing in taxes paid.

- *Poverty elimination*: $g(c_i, z_i) = 1$ 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.[2]

- *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 (but such weights go beyond the formal

[2]Such weights violate the positivity condition (4).

framework in that they depend on the entire income distribution and not just on c_i, z_i, x_i, and y_i.)

A **tax reform** is a function $\Delta T \in \mathscr{T}$ whose interpretation is that it represents some change to the status quo tax policy. Define $\Delta T_i(z_i) = \Delta T(z_i, x_i)$. Let $\hat{\mathscr{T}}$ be on open subset of $\mathscr{T}$ that contains $\mathscr{T}^\circ$ as a subset of $\mathscr{T}$ (in the metric topology induced by ρ) and such that for all $T \in \hat{\mathscr{T}}$, and all $\Delta T \in \mathscr{T}$ and almost all $i \in [0,1]$, $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\tilde{U}_i(T + \varepsilon\Delta T)$ exists at $\varepsilon = 0$ and is continuous at $\varepsilon = 0$.[3,4] By the envelope theorem, when $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\tilde{U}_i(T + \varepsilon\Delta T)$ exists at $\varepsilon = 0$ and is continuous at $\varepsilon = 0$, $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\big|_{\varepsilon=0}\tilde{U}_i(T + \varepsilon\Delta T) = -\Delta T_i(z(T))$.

Say tax reform ΔT is **locally budget neutral at** T if $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\big|_{\varepsilon=0} R(T + \varepsilon\Delta T) = 0$. For any tax policy $T \in \hat{\mathscr{T}}$, tax reform $\Delta T \in \mathscr{T}$, and system of social welfare weights $g \in \mathscr{G}$, say that a locally budget neutral tax reform ΔT is **locally desirable** if

$$\int g_i(T)\,\Delta T_i(z_i(T))\,\mathrm{d}i < 0.$$

In other words, ΔT is desirable if the cost of the tax change to different individuals, weighted by the local welfare weights, is negative. Say that tax system $T \in \hat{\mathscr{T}}$ satisfies the **local optimal tax criterion** if

$$\forall \Delta T \in \mathscr{T}, \left[\left.\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0} R(T + \varepsilon\Delta T) = 0 \Rightarrow \int g_i(T)\,\Delta T_i(z_i(T))\,\mathrm{d}i = 0\right]. \tag{6}$$

Saez and Stantcheva (2016) say that (6) 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.

There is however a question of why the local optimal tax criterion is the right thing to look at. Why should we think that it is good if a tax system satisfies (6)? What normative assumptions justify the criterion (6)?

In a more traditional framework, the above questions are answered on the basis of the properties of a *global* optimization problem from which a condition such as (6) is *derived*. To see this, consider the case of utilitarianism: Suppose the goal is to maximize the utilitarian objective $\int U_i(T)\,\mathrm{d}i$ subject to a revenue constraint $R(T) \geq E$, where E represents required government expenditures. Then, using the envelope theorem, a necessary condition for T to be an optimum is:[5]

$$\forall \Delta T \in \mathscr{T}, \left[\left.\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0} R(T + \varepsilon\Delta T) = 0 \Rightarrow \int u'(z_i(T) - T_i(z_i(T)) - v_i(z_i(T)))\,\Delta T_i(z_i(T))\,\mathrm{d}i = 0\right]. \tag{7}$$

[3]Note that if $\hat{\mathscr{T}} = \mathscr{T}^\circ$, then these requirements are satisfied.

[4]Note that because u is continuously differentiable, for any $T, \Delta T \in \mathscr{T}$, $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\tilde{U}_i(T + \varepsilon\Delta T)$ exists at $\varepsilon = 0$ and is continuous at $\varepsilon = 0$ if and only if $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}U_i(T + \varepsilon\Delta T)$ exists at $\varepsilon = 0$ and is continuous at $\varepsilon = 0$.

[5]This requires that for almost all i, and all $\Delta T \in \mathscr{T}$, $h_i(\varepsilon) = U_i(T + \varepsilon\Delta T)$ is continuously differentiable in a neighborhood around $\varepsilon = 0$.

If the weights are utilitarian (that is, $g_i(c_i, z_i) = u'(c_i - v_i(z_i))$), then (7) coincides with (6). But in the utilitarian case, the justification for condition (7) is that it is a necessary condition for an optimum if one's aim is to maximize the objective $\int U_i(T)\,\mathrm{d}i$ subject to a revenue requirement, and one has independent reasons for thinking that $\int U_i(T)\,\mathrm{d}i$ is a reasonable objective.

However a justification analogous to (7) is not available to Saez and Stantcheva (2016) because, under their approach, the local optimality conditions are not derived from a global optimization problem; it is a purely local optimality condition. The question is then why such a local optimality condition is justified. Usually such conditions are justified by a broader global optimization problem from which they are derived.

In their appendix, Saez and Stantcheva (2016) do however describe a foundation for their generalized social welfare weights. In fact, their results depend on this foundation; in particular the proof that (6) is indeed a necessary condition for a local optimum–their Proposition 1–depends on the definitions and arguments presented in the appendix. Proposition 1, in turn, serves as a foundation for other results, such as the characterization of optimal marginal tax rates (Proposition 2). Let us then explore this foundation.

Any system of social welfare weights g and tax policy $\tilde{T}$ together define a social welfare function $W^g_{\tilde{T}}$ by which tax policies T in general can be evaluated. In particular, this social welfare function takes the form:

$$W^g_{\tilde{T}}(T) = \int g_i\left(\tilde{T}\right)\tilde{U}_i(T)\,\mathrm{d}i. \tag{8}$$

$W^g_{\tilde{T}}$ is the social welfare function that evaluates all tax policies T using welfare weights defined locally by the tax policy $\tilde{T}$. Recall that $\tilde{U}_i(T)$ is the version of agent's indirect utility function in dollar terms (3).

Proposition 1 *For all systems of social welfare weights g, tax policies $T \in \hat{\mathscr{T}}$ and tax reforms ΔT,*

$$\left.\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0} W^g_T(T + \varepsilon\Delta T) = -\int g_i(T)\,\Delta T_i(z_i(T))\,\mathrm{d}i.$$

This result follows immediately from the envelope theorem.

Saez and Stantcheva (2016) provide the following definition (in their appendix):

Definition 1 *A tax system $T \in \mathscr{T}$ as **locally optimal** if there exists a neighborhood N of T such that for all $T' \in N, R(T) = R(T') \Rightarrow W^g_T(T) \geq W^g_T(T')$.*

Saez and Stantcheva (2016) establish the following result (Proposition 1 in Saez and Stantcheva (2016)).

Proposition 2 *If $T \in \hat{\mathscr{T}}$ is locally optimal, then T satisfies the local optimal tax criterion (6).*[6]

[6]This proposition also requires the assumption that for almost all i, and all $\Delta T \in \mathscr{T}$, $h_i(\varepsilon) = U_i(T + \varepsilon\Delta T)$ is continuously differentiable in a neighborhood of $\varepsilon = 0$.

One might think that Proposition 2 is a justification for the local optimal tax criterion (6) as a necessary condition for optimality in the same way that the fact that (7) is a necessary condition for an optimum if one's aim is to maximize the objective (5) subject to a revenue requirement. But this is not so. The question is: Why should the local optimality criterion in Definition 1 be a compelling one? In contrast to maximizing aggregate utility, it is not clear why the criterion in Definition 1 should be normatively compelling. The issue here is not about a utilitarian vs a non-utilitarian objective. The issue is that Definition 1 does not give us a global objective at all, and it does not explain why $W_T^g(T) \geq W_T^g\left(\tilde{T}\right)$ is a good criterion for making local comparisons between T and $\tilde{T}$.

One seeming virtue of the GSMWW approach is that it leads to tax policies that are locally Pareto optimal. Formally, say that a tax system $T \in \mathscr{T}$ is **locally Pareto optimal** if there exists a neighborhood N such that for all $\tilde{T} \in N$,

$$R(T) = R\left(\tilde{T}\right) \Rightarrow \left[\lambda\left(i : U_i\left(\tilde{T}\right) > U_i(T)\right) > 0 \Rightarrow \lambda\left(i : U_i(T) > U_i\left(\tilde{T}\right)\right) > 0\right],$$

where recall λ is Lebesgue measure. In other words, if T is locally Pareto optimal, then within some neighborhood N of T, any revenue-equivalent tax system $\tilde{T}$ that is better for some positive measure set of agents and raises the same revenue as T is worse for some other positive measure set of agents; there is no tax policy $\tilde{T}$ in N that is superior for a positive measure set while being worse only for a zero measure set.

Proposition 3 *If $T \in \mathscr{T}$ is locally optimal, then T is locally Pareto optimal.*[7]

Proposition 3 is significant, but it does not by itself provide a justification for the optimality criterion in Definition 1. Moreover, as I discuss in Section 3, there is a tension between the Pareto principle and the representation of broad ethical values.

3 A reason to expect problems for the GSMWW approach

In this section I explore a prima facie reason to expect that the GSMWW approach may encounter problems.

Several authors, including Sen (1970, 1979) and Kaplow and Shavell (2001), 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.[8] As they say, one philosopher's modus ponens is another philosopher's modus tollens.

The question is then how the GSMWW approach avoids the problem. The GSMWW approach incorporates broader social values into normative evaluations and at the same time it appears to

[7] This is closely related to Proposition 4 of Saez and Stantcheva (2016).

[8] See also Weymark (2017).

satisfy a local version of the Pareto principle, as stated in Proposition 3. Is this consistent with the results of Sen and Kaplow and Shavell?

What I will argue below is that the compatibility of Pareto and the incorporation broader normative values is purchased at a steep price: In particular, it is possible to create the appearance of a compatibility between the Pareto principle and broader normative principles because the GSMWW approach is just not consistent with *any* global ranking over tax systems.

The basic idea behind the Sen (1970, 1979) and Kaplow and Shavell (2001) is as follows. Consider two states s_1 and s_2 and a set of agents, such that each agent i has a utility $u_i(s_j)$ for each of the two states $j = 1, 2$. Suppose that $u_i(s_1) = u_i(s_2)$ for all i, so that all agents are indifferent between the two states. Suppose there is another moral difference between s_1 and s_2, having to do with fairness or freedom or desert or rights or any other moral consideration that is not captured in the agents' utility functions. Pareto leaves essentially no room for such considerations to enter. In particular, let W be a social welfare function on the states that satisfies a Pareto indifference condition:

$$[u_i(s_1) = u_i(s_2), \forall i] \Rightarrow W(s_1) = W(s_2).$$

Pareto indifference implies that we cannot view any moral criterion other than preference satisfaction as being *independently good*; that is we cannot say that there is a trade-off between how fair a situation is and how well preferences are satisfied: If preference satisfaction is held fixed (i.e., $u_i(s_1) = u_i(s_2), \forall i$), then nothing else matters to social evaluation (i.e., $W(s_1) = W(s_2)$). So making a situation more fair for example or reducing rights violations will only matter if it has an effect on preferences, but will never matter in itself. Similar considerations apply if we consider other versions of the Pareto principle, such as weak or strong Pareto.

Fleurbaey, Tungodden and Chang (2003) criticize Kaplow and Shavell (2001) for claiming that their results show that Pareto implies *welfarism*; this criticism is based on the technical definition of welfarism.[9] Fleurbaey and Maniquet (2011) put forward a theory of social welfare that incorporates fairness considerations while respecting Pareto. Fleurbaey and Maniquet (2018) review the literature incorporating fairness into optimal tax. Fleurbaey and Maniquet (2018) write, "One of our main points ... is that the classical social welfare function framework is more flexible than commonly thought, and can accommodate a very large set of non-utilitarian values." However Fleurbaey and Maniquet (2018) do not dispute the above reasoning: namely, that Pareto indifference restricts the incorporation of broader values in precisely the way described in the preceding paragraph.

When we consider Proposition 3 and the claim that welfare weights incorporate broader ethical values we may wonder how this is consistent with the the considerations raised by Sen (1970, 1979) and Kaplow and Shavell (2001), and whether somehow the considerations raised by Fleurbaey and

[9]See Theorem 2 of Weymark (2016). In addition to a Pareto principle, an additional independence condition is required for welfarism. See also d'Aspremont and Gevers (1977), Sen (1977, 1986), Hammond (1979), Roberts (1980), Bossert and Weymark (2004).

Maniquet (2018)–albeit that approach advocates a social welfare function approach–can redeem the GSMWW approach. Let us suppose that we want the following combination features:

1. Social evaluations incorporate moral considerations not captured in preferences, and moreover considerations that can vary independently of preferences across policies that we would like to compare.

2. Some version of a Pareto-like principle.

3. Social evaluations are consistent with the existence of a global objective that ranks alternatives as better or worse.

Another way of thinking of the results of the results of Sen (1970, 1979) and Kaplow and Shavell (2001) is as saying that the three desiderata are not mutually satisfiable if the Pareto-like principle is Pareto indifference, or, under mild conditions, another version of Pareto. In the case of Saez and Stantcheva (2016), the third desideratum is not explicitly required. Moreover the second desideratum is not explicitly spelled out. First, there is Proposition 3, which falls short of Pareto indifference; Proposition 3 says only that optima are Pareto efficient. However, the fact that GSMWW make evaluations via weighted averages of utilities suggests something closer to Pareto indifference. This is formalized in Section 4. Usually, we are not focused on the third desideratum, because it is taken for granted and left implicit; so the conflict is phrased as holding between the first two. however, in this case, the third desideratum is explicitly at issue. Below I will show that the Saez and Stantcheva (2016) approach conflicts with the third.

In a way, Saez and Stantcheva (2016) acknowledge the above, as they say that there is no global welfare objective; however, they do not discuss whether this feature of their approach is problematic. In a more critical vein, in discussing the Saez and Stantcheva, approach Fleurbaey and Maniquet (2018) 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 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)

In a footnote, Fleurbaey and Maniquet (2018) clarify their first sentence, writing, "This claim is formally true only if transitivity is logically strengthened into being representable (by a function). Transitive Paretian social preferences may or may not be representable by a function. The social welfare function approach is here meant to include all such social preferences." (p. 1059)

Fleurbaey and Maniquet (2018) only discuss this point 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. I do both of these things below. In what follows, I will show precisely when the generalized social welfare weights approach is consistent–I will characterize the property of welfare weights that make them consistent, and I will show how to explicitly construct a cycle whenever welfare weights are not consistent (such a cycle is constructed in the proof of Theorem 1). I do not require a complete or representable rankings; I provide conditions under which there is no preorder that is consistent with the implicit welfare weights. Moreover, all of my global comparisons for tax policies hold revenue constant, so that I show that for welfare weights that are not of a utilitarian kind, one can construct cycles even while holding the revenue of the tax policy constant.

4 Rationalization of generalized social welfare weights

The GSMWW approach makes local comparisons among tax policies. The question that I address here is whether these local comparisons are *consistent*. But how do we assess the consistency of local welfare weights? We do so by asking whether there is a global ranking of tax policies that agrees locally with the GSMWW approach. If there is no such global ranking, then we say that the weights are inconsistent. Why is this a reasonable way of assessing consistency? The reason is that if it is not possible to have a global ranking of tax policies whose judgements agree with those of local welfare weights, then *anyone who arrives at a global ranking will disagree with the judgements of the welfare weights on some local comparisons*. That is, to follow the GSMWW approach requires that you *cannot* have a global ranking of tax policies. Another way of expressing the idea is to say that if you pooled all the local judgements in order to create a global ranking, that global ranking would be intransitive. That is the sense in which welfare weights deliver inconsistent judgements if they cannot be extended to a global ranking.

The basic idea is analogous to the following. Imagine that I know the gradient of a function $\nabla f : \mathbb{R}^n \to \mathbb{R}$. This allows me to make local comparisons: For any direction d, if $\nabla f(x) \cdot d > 0$, then at x, f is increasing in direction d. And if I know the gradient of the function everywhere, then for any pair of points, x and y, I can recover the difference $f(y) - f(x) = \int_0^1 \nabla f(\rho(\theta)) \cdot \rho'(\theta) \, \mathrm{d}\theta$, where ρ is a differentiable path from x to y. So local comparisons via ∇f imply global comparisons of the form $f(y) > f(x)$ inherent in f. In an analogous way, we can derive global comparisons from the local comparisons of the GSMWW approach.

Define a **path** to be a continuous function $\rho : [a, b] \to \hat{\mathscr{T}}; \theta \mapsto T^{\rho,\theta}$, where $a, b \in \mathbb{R}, a < 0, 1 < b$. Now for paths ρ and all $i \in [0, 1]$, define the function $T_i^\rho : \mathbb{R} \times [a, b] \to \mathbb{R}$ by $T_i^\rho(z_i, \theta) := T_i^{\rho,\theta}(z_i)$. For any path ρ and $\theta \in [a, b]$, define $z_i^\rho(\theta) = z_i\left(T^{\rho,\theta}\right)$.

Let P be the set of paths ρ such that

(i) for all i, $(z_i, \theta) \mapsto T_i^\rho(z_i, \theta)$ is continuously differentiable,

(ii) for all $\theta \in [a, b]$, there are at most finitely many $i_0 \in [0, 1]$ such that for some $z \in Z$, $i \mapsto \frac{\partial}{\partial \theta} T_i^\rho(z, \theta)$ is discontinuous at $i = i_0$, and

(iii) $\forall \theta \in [a,b], R\left(T^{\rho,\theta}\right) = R\left(T^{\rho,0}\right)$.

In particular observe that by (iii), revenue is constant on any path ρ in P.

Consider a binary relation $\precsim$ on $\hat{\mathscr{T}}$. Say that $\precsim$ is a **preorder** if $\precsim$ is transitive and reflexive. For any binary relation $\precsim$ on $\hat{\mathscr{T}}$, let $\sim$ and $\prec$ be the symmetric and asymmetric parts of $\precsim$, respectively.

Definition 2 *Let g be a system of welfare weights. Say that a binary relation $\precsim$ **rationalizes** g if for all $\rho \in P$,*

local improvement principle:

$$\begin{aligned} &\int g_i\left(T^{\rho,0}\right) \left.\frac{\partial}{\partial \theta}\right|_{\theta=0} T_i^{\rho}\left(z_i^{\rho}(0), \theta\right) \mathrm{d}i < 0 \Rightarrow \left[\exists \bar{\theta} \in (0,1], \forall \theta \in \left(0, \bar{\theta}\right), T^{\rho,0} \prec T^{\rho,\theta}\right], \\ &\int g_i\left(T^{\rho,0}\right) \left.\frac{\partial}{\partial \theta}\right|_{\theta=0} T_i^{\rho}\left(z_i^{\rho}(0), \theta\right) \mathrm{d}i > 0 \Rightarrow \left[\exists \bar{\theta} \in (0,1], \forall \theta \in \left(0, \bar{\theta}\right), T^{\rho,0} \succ T^{\rho,\theta}\right], \end{aligned} \tag{9}$$

indifference principle:

$$\text{and } \left(\forall \theta' \in [0,1], \int g_i\left(T^{\rho,\theta'}\right) \left.\frac{\partial}{\partial \theta}\right|_{\theta=\theta'} T_i^{\rho}\left(z_i^{\rho}\left(\theta'\right), \theta\right) \mathrm{d}i = 0\right) \Rightarrow T^{\rho,0} \sim T^{\rho,1}. \tag{10}$$

*Say that the system of welfare weights g is **rationalizable** if there exists a preorder that rationalizes g.*

Remark 1 *Note that we define the what it means for $\precsim$ to rationalize g for an arbitrary binary relation $\precsim$, not just a preorder. But we say that g is rationalizable only if it is rationalized by a preorder.*

The GSMWW approach implies a series of local comparisons. The relation $\precsim$ attempts to *simultaneously* capture all of the local comparisons implied by the social welfare weights. In this sense, it is a global relation.

I will discuss the two conditions (9) and (10). Consider first the *local improvement principle* (9). Consider a path ρ such that for some $T \in \hat{\mathscr{T}}$ and $\Delta T \in \mathscr{T}$,[10]

$$T^{\rho,\theta} = T + \theta \Delta T. \tag{11}$$

In this case, the condition

$$\int g_i\left(T^{\rho,0}\right) \left.\frac{\partial}{\partial \theta}\right|_{\theta=0} T_i^{\rho}\left(z_i^{\rho}(0), \theta\right) \mathrm{d}i < 0, \tag{12}$$

from the antecedent of the first part of (9) reduces to:

$$\int g_i(T) \Delta T_i\left(z_i(T)\right) \mathrm{d}i < 0. \tag{13}$$

[10] We may assume that for all $\theta \in [a,b], T + \theta \Delta T \in \hat{\mathscr{T}}$; observe that since $T \in \hat{\mathscr{T}}$, for all $\Delta T \in \mathscr{T}$, there exists $\xi > 0$ such that $T + \theta \xi \Delta T \in \hat{\mathscr{T}}$ for all $\theta \in [a,b]$. So we could consider $\xi \Delta T$ rather than ΔT to begin with.

But if ΔT is a locally budget neutral tax reform so that

$$\left.\frac{\mathrm{d}}{\mathrm{d}\theta}\right|_{\theta=0} R\left(T+\theta\Delta T\right)=0, \tag{14}$$

then (13) just amounts to Saez and Stantcheva's definition of a *locally desirable* tax reform. So, in this case, the first line in the local improvement principle (9) reduces to the claim that if ΔT is a locally budget neutral and locally desirable tax reform (locally desirable in the sense of (13)), then a sufficiently small version of the reform $T+\theta\Delta T$–that is, for sufficiently small $\theta>0$–is strictly socially preferred to T according to the global relation $\prec$; that is $T\prec T+\theta\Delta T$.

Now if we are to be strict with our definitions, we will notice that the definition of P is such that it requires that for all $\rho\in P$,

$$\forall\theta\in[a,b],R\left(T^{\rho,\theta}\right)=R\left(T^{\rho,0}\right), \tag{15}$$

but if ρ is chosen so that (11) holds, this constant revenue condition may not hold for ρ. However, suppose that instead we choose ρ so that

$$T^{\rho,\theta}=T+\theta\Delta T+\left(R\left(T\right)-R\left(T+\theta\Delta T\right)\right), \tag{16}$$

where the last term $\left(R\left(T\right)-R\left(T+\theta\Delta T\right)\right)$ is a lumpsum transfer, independent of income and observable characteristics. Then (15) does hold, and if we continue to assume (14), then (12) still reduces to (13).[11]

Above we have examined a special case of the local improvement principle which basically says that a locally budget neutral and locally desirable tax reform in Saez and Stantcheva's sense is deemed to be good according to the global relation $\precsim$. It seems that Saez and Stancheva ought to be committed to the local improvement principle in the special case. However it is difficult to see why the local improvement should be compelling in this special case, but not more generally when $T^{\rho,\theta}$ does not necessarily take the form given in the preceding paragraph. The local improvement condition maintains the same spirit in the more general case: It says that if we have a parameterized family of *revenue equivalent* tax policies $\left(T^{\rho,\theta}\right)_{\theta\in[a,b]}$ and start at $\theta=0$ and an increase in θ is locally desirable, then for a sufficiently small $\theta>0$, $T^{\rho,\theta}$ should be socially preferred to $T^{\rho,0}$. This seems like a weak and reasonable way of inferring properties of global social preference from local social preference.

Note that the second line of the local improvement is analogous to the first, but it applies to the case where the local change is undesirable rather than desirable.

Whereas the local improvement principle allows us to draw inferences about global social *strict preference*, the indifference principle allows us to draw inferences about global social *indifference*.

The indifference principle says that if there is a smooth path ρ of revenue-equivalent tax polices such that at every point θ' on the path, welfare weights cannot detect either an improvement or

[11] In particular, observe that if (16) and (14) hold, then $\left.\frac{\partial}{\partial\theta}\right|_{\theta=0} T_i^{\rho}\left(z_i^{\rho}\left(0\right),\theta\right)=\Delta T_i\left(z_i\left(T\right)\right)$.

deterioration as we move through θ':

$$\int g_i\left(T^{\rho,\theta'}\right)\left.\frac{\partial}{\partial\theta}\right|_{\theta=\theta'} T_i^{\rho}\left(z_i^{\rho}\left(\theta'\right),\theta\right)\mathrm{d}i=0,$$

then the endpoints of the path are indifferent according to the global relation. It is useful to make an analogy. Suppose that $\rho:[0,1]\to\mathbb{R}^n$ is a smooth path in $\mathbb{R}^n$ and $f:\mathbb{R}^n\to\mathbb{R}$ is a smooth function. Define $f^{\rho}:\mathbb{R}\to\mathbb{R}$ by $f^{\rho}(\theta)=f(\rho(\theta))$. Then $\frac{\mathrm{d}}{\mathrm{d}\theta}f^{\rho}(\theta)=\nabla f(\rho(\theta))\cdot\rho'(\theta)$, where $\rho'(\theta)=(\rho_1'(\theta),\cdots,\rho_n'(\theta))$ and $\rho_i'(\theta)$ is the derivative of the i-component $\rho_i(\theta)$ of $\rho(\theta)$. Then the indifference condition is analogous to the condition that if $\frac{\mathrm{d}}{\mathrm{d}\theta}f^{\rho}(\theta)=0$ for all $\theta\in[0,1]$, then $f^{\rho}(0)=f^{\rho}(1)$, which in this case is a consequence of the fundamental theorem of calculus.

For the utilitarian case we have the following proposition.

Proposition 4 *Consider the utilitarain social welfare function $W:\hat{\mathscr{T}}\to\mathbb{R}$ defined by*

$$W(T)=\int U_i(T)\,\mathrm{d}i,$$

and define the relation $\precsim$ on $\hat{\mathscr{T}}$ by

$$\left(T_0\precsim T_1\Leftrightarrow\left[W(T_0)\leq W(T_1)\text{ and }R(T_0)\leq R(T_1)\right]\right),\forall T_0,T_1\in\hat{\mathscr{T}}. \tag{17}$$

That is T_1 is ranked T_0 weakly higher by $\precsim$ if only if T_1 is weakly better in terms of both utilitarian welfare and revenue raised. Let welfare weights be utilitarian: $g_i(c_i,z_i)=u'(c_i-v_i(z_i))$. Then $\precsim$ rationalizes g.[12]

The proof is instructive. First, I establish the local improvement principle in the utilitarian case. Choose a path $\rho\in P$ and suppose that (12) holds.

Observe that

$$\begin{aligned}\left.\frac{\mathrm{d}}{\mathrm{d}\theta}\right|_{\theta=0}W\left(T^{\rho,\theta}\right)&=-\int u'\left(z_i^{\rho}(0)-T_i\left(z_i^{\rho}(0)\right)-v_i\left(z_i^{\rho}(0)\right)\right)\left.\frac{\partial}{\partial\theta}\right|_{\theta=0}T_i^{\rho}\left(z_i^{\rho}(0),\theta\right)\mathrm{d}i\\&=-\int g_i\left(T^{\rho,0}\right)\left.\frac{\partial}{\partial\theta}\right|_{\theta=0}T_i^{\rho}\left(z_i^{\rho}(0),\theta\right)\mathrm{d}i,\end{aligned}$$

where the first line follows from the envelope theorem and the second follows from the fact that welfare weights are utilitarian. It now follows from (12) that $\left.\frac{\mathrm{d}}{\mathrm{d}\theta}\right|_{\theta=0}W\left(T^{\rho,0}\right)>0$. So there exists $\bar{\theta}>0$ such that for all $\theta\in\left(0,\bar{\theta}\right)$, $W\left(T^{\rho,0}\right)<W\left(T^{\rho,\theta}\right)$. It now follows from (17) and the fact that revenue is constant along any path in P that $T^{\rho,0}\prec T^{\rho,\theta}$. This, along with a similar argument for

[12] Observe that the theorem would remain true if (17) were replaced by either

$$\left(T_0\precsim T_1\Leftrightarrow W(T_0)\leq W(T_1)\right),\forall T_0,T_1\in\hat{\mathscr{T}}$$

or

$$\left(T_0\precsim T_1\Leftrightarrow\left[W(T_0)\leq W(T_1)\text{ and }R(T_0)=R(T_1)\right]\right),\forall T_0,T_1\in\hat{\mathscr{T}}. \tag{18}$$

the case where the inequality in (12) is reversed, establishes the local improvement principle in the utilitarian case.

Now I establish the indifference principle in the utilitarian case. Suppose that for $\rho \in P$,

$$\forall \theta' \in [0,1], \int g_i\left(T^{\rho,\theta'}\right) \frac{\partial}{\partial \theta} T_i^{\rho}\left(z_i^{\rho}\left(\theta'\right);\theta'\right) \mathrm{d}i = 0. \tag{19}$$

We have

$$\begin{aligned} W\left(T^{\rho,1}\right) - W\left(T^{\rho,0}\right) &= \int U_i\left(T^{\rho,1}\right) \mathrm{d}i - \int U_i\left(T^{\rho,0}\right) \mathrm{d}i = \int_0^1 \left[\frac{\mathrm{d}}{\mathrm{d}\theta} \int U_i\left(T^{\theta}\right) \mathrm{d}i\right] \mathrm{d}\theta \\ &= \int_0^1 \int \frac{\mathrm{d}}{\mathrm{d}\theta} U_i\left(T^{\rho,\theta}\right) \mathrm{d}i \mathrm{d}\theta \\ &= \int_0^1 \int u'\left(z_i^{\rho}(\theta) - T_i^{\rho,\theta}\left(z_i^{\rho}\left(T^{\rho,\theta}\right)\right) - v_i\left(z_i^{\rho}\left(T^{\rho,\theta}\right)\right)\right) \frac{\partial}{\partial \theta} T_i\left(z_i\left(T^{\theta}\right),\theta\right) \mathrm{d}i \mathrm{d}\theta \\ &= \int_0^1 \int g_i\left(T^{\rho,\theta}\right) \frac{\partial}{\partial \theta} T_i^{\rho}\left(z_i^{\rho}(\theta),\theta\right) \mathrm{d}i \mathrm{d}\theta = \int_0^1 0 \mathrm{d}\theta = 0, \end{aligned} \tag{20}$$

where the second equality follows from the fundamental theorem of calculus, the fourth from the envelope theorem, the fifth from the assumption that welfare weights are utilitarian, and the sixth from (19). It follows that $W\left(T^{\rho,0}\right) = W\left(T^{\rho,1}\right)$ and by the definition of P, $R\left(T^{\rho,0}\right) = R\left(T^{\rho,1}\right)$. So by (17), $T^{\rho,0} \sim T^{\rho,1}$. This establishes the indifference principle in the utilitarian case, and completes the proof of the proposition. □

Proposition 4 provides support for the notion of rationalization in Definition 2. If welfare weights are utilitarian and the global ranking of revenue equivalent tax policies is determined by a utilitarian social welfare function, then the local improvement and indifference principles are *theorems*. This provides further plausibility to these conditions.

In this paper, I do not assume that welfare weights are utilitarian and, in Definition 2, the local improvement and indifference principles are treated as *axioms* that characterize rationalization. My claim is that these principles are reasonable axioms for capturing the implicit local comparisons made by social welfare weights and integrating them into an overall ranking. Note that for my purposes it is not so important that the rationalizing relation captures *all* such comparisons, just that the comparisons it does capture are genuinely implied.

Before concluding this section, I would like to include a couple of definitions of rankings $\precsim$ on tax policies, as well as an alternative definition of rationalization. Recall that λ is the Lebesgue measure on the interval $[0,1]$, interpreted as the set of agents. Say that a ranking $\precsim$ on tax policies satisfies **Pareto indifference** if:

$$\forall T_0, T_1 \in \hat{\mathscr{T}}, R\left(T_0\right) = R\left(T_1\right) \Rightarrow \left[\lambda\left(i : \tilde{U}_i\left(T_0\right) = \tilde{U}_i\left(T_1\right)\right) = 1 \Rightarrow T_0 \sim T_1\right]. \tag{21}$$

That is $\precsim$ satisfies Pareto indifference if, whenever almost all agents are indifferent between two

tax policies, T_0 and T_1, then the ranking is indifferent among these policies as well.

Definition 3 *Let g be a system of welfare weights. Say that a binary relation $\precsim$* ***rationalizes*** *g* ***with Pareto indifference*** *if for all $\rho \in P$, g and $\precsim$ jointly satisfy the local improvement principle (9) and Pareto indifference (21). Say that g is* ***rationalizable with Pareto indifference*** *if there exists a preorder that rationalizes g with Pareto indifference.*

Definition 3 provides an alternative definition of rationalization that substitutes the Pareto indifference condition for the indifference principle. It will help to establish the robustness of the phenomenon studied here because it will help to show that generalized social welfare weights fail to be rationalizable under alternative definitions of rationalization: See Section 6.

One might think that Saez and Stantcheva (2016) need not be committed to the Pareto indifference condition for a rationalizing relation: After all, this the Pareto indifference condition restricts social evaluation to be measurable with respect to welfare and the point of GSMWW is to bring broader normative values into the analysis. However notice how GSMWW brings broader values into the analysis: by placing *weights* on the resources going to each agent. If all agents are indifferent among two tax policies, then it might seem that if does not matter how those weights are assigned, the tax policies should be regarded as equally good. An alternative framework that did not represent the fairness in terms of the weights attached to different individuals might more clearly articulate why indifference among all agents to those tax policies might not imply that they should be regarded as socially indifferent to each other when one is in some sense more fair than the other.

5 A special case in which welfare weights are rationalizable: libertarian weights with fixed incomes

In the previous section we saw a special case in which welfare weights are rationalizable: namely when the welfare weights are utilitarian (Proposition 4). In this section, we present another, perhaps less obvious case in which welfare weights can be rationalized: libertarian welfare weights with fixed incomes; as we shall see below, the fixed incomes are crucial. This special case of libertarian weights with fixed incomes is discussed in Section II.A of Saez and Stantcheva (2016).

Recall that libertarian welfare weights are weights of the form

$$g_i(c_i, z_i) = \tilde{g}_i(z_i - c_i) = \tilde{g}_i(t_i)$$

where $t_i = z_i - c_i$ is the tax paid and $\tilde{g}_i(t_i)$ is increasing in t_i. Assume moreover that for all i there

exists income z_i^* such that[13]

$$v_i(z_i) = \begin{cases} 0, & \text{if } z_i \leq z_i^* \\ \infty, & \text{if } z_i > z_i^* \end{cases} \tag{22}$$

Strictly speaking this cost function does not fit into the general assumptions laid out in Section 2 because the cost function in (22) is discontinuous at z_i^* and hence not differentiable.

In this section I restrict attention to differentiable tax policies T_i is differentiable and $T_i'(z) < 1$ for all z. Let $\hat{\mathscr{T}}' = \left\{T \in \hat{\mathscr{T}} : T'(z) < 1, \forall z \in Z\right\}$. In this section we say that a binary relation $\precsim$ **rationalizes** g on $\mathscr{T}'$ if the the conditions of Definition 9 are satisfied when we restrict attention to tax policies in $\mathscr{T}'$. For all such tax policies $T \in \mathscr{T}'$, $z_i(T) = z_i^*$. So the assumptions in this section essentially amount to the assumption that incomes are fixed and there are no behavioral responses to taxes.

Now define the function

$$H_i(t_i) = -\int_0^{t_i} g_i\left(t_i'\right) dt_i'. \tag{23}$$

Note that

$$H_i'(t_i) = -g_i(t_i). \tag{24}$$

Now consider the social welfare function

$$W(T) = \int H_i\left(T\left(z_i^*\right)\right) \mathrm{d}i \tag{25}$$

Define $\precsim$ by

$$\left(T_0 \precsim T_1 \Leftrightarrow \left[W(T_0) \leq W(T_1) \text{ and } R(T_0) \leq R(T_1)\right]\right), \quad \forall T_0, T_1 \in \hat{\mathscr{T}}'. \tag{26}$$

Proposition 5 *$\precsim$ rationalizes g on $\hat{\mathscr{T}}'$.*

Proof. Choose $\rho \in P$.[14] Observe that for any $\theta' \in [a,b]$,

$$\left.\frac{\mathrm{d}}{\mathrm{d}\theta}\right|_{\theta=\theta'} W\left(T^{\rho,\theta}\right) = \int H_i'\left(T^{\rho,\theta'}\left(z_i^*\right)\right) \left.\frac{\partial}{\partial\theta}\right|_{\theta=\theta'} T^{\rho,\theta}\left(z_i^*\right) \mathrm{d}i = -\int g_i\left(T^{\rho,\theta'}\right) \left.\frac{\partial}{\partial\theta}\right|_{\theta=\theta'} T^{\rho,\theta}\left(z_i^\rho\left(\theta'\right)\right) \mathrm{d}i, \tag{27}$$

where the second equality follows from (24) and the fact that $z_i^\rho(\theta') = z_i^*$. First consider the local improvement principle, and assume that (12) holds. It follows from (12) and (27) that $\left.\frac{\mathrm{d}}{\mathrm{d}\theta}\right|_{\theta=0} W\left(T^{\rho,0}\right) > 0$. It follows that there exists $\bar{\theta} \in (0,1]$ such that for all $\theta \in \left(0,\bar{\theta}\right)$ $W\left(T^{\rho,0}\right) <$

[13] It is not necessary assume that the cost of earning income is literally infinite above z_i^*; it sufficient to assume that this cost is very large.

[14] Recall that in this section we assume that for all $\rho \in P$ and $\theta \in [a,b]$, $T^{\rho,\theta} \in \hat{\mathscr{T}}'$.

$W\left(T^{\rho,\theta}\right)$. It follows from (26) and the fact that revenue is constant along any path $\rho \in P$ that for all $\theta \in \left(0, \bar{\theta}\right)$, $T^{\rho,0} \prec T^{\rho,\theta}$. This establishes the local improvement principle.

Next consider the indifference principle. Suppose that (19) holds. We have

$$\begin{aligned} W\left(T^{\rho,1}\right) - W\left(T^{\rho,0}\right) &= \int H_i\left(T^{\rho,1}\left(z_i^*\right)\right) \mathrm{d}i - \int H_i\left(T^{\rho,0}\left(z_i^*\right)\right) \mathrm{d}i \\ &= \int_0^1 \left[\left. \frac{\mathrm{d}}{\mathrm{d}\theta} \right|_{\theta=\theta'} \int H_i\left(T^{\theta}\left(z_i^*\right)\right) \mathrm{d}i \right] \mathrm{d}\theta' = \int_0^1 \int \left. \frac{\mathrm{d}}{\mathrm{d}\theta} \right|_{\theta=\theta'} H_i\left(T^{\rho,\theta}\left(z_i^*\right)\right) \mathrm{d}i \mathrm{d}\theta' \\ &= \int_0^1 \int g_i\left(T^{\rho,\theta}\right) \left. \frac{\partial}{\partial\theta} \right|_{\theta=\theta'} T_i^{\rho}\left(z_i^{\rho}\left(\theta'\right);\theta\right) \mathrm{d}i \mathrm{d}\theta = \int_0^1 0 \mathrm{d}\theta' = 0, \end{aligned} \tag{28}$$

where the derivation is justified in a way similar to (20) and (27). It now follows from (28) and (26) that and the fact that revenue is constant along any path ρ in P that $T^{\rho,0} \sim T^{\rho,1}$. This establishes the indifference principle, and hence also that $\precsim$ rationalizes g. $\square$

An attentive reader will have noticed that the proof of Proposition 5 on rationalization for libertarian weights and fixed incomes resembles the proof of Proposition 4 for utilitarian weights. This is not a coincidence. Despite appearances to the contrary, the fixed incomes–or in other words, the lack of behavioral responses to taxes–make the case of libertarian case very similar to the utilitarian case. Essentially, in this case (25) amounts to a quasi-utilitarian social welfare function that essentially rationalizes the weights. To see that (25) is "quasi"-utilitarian, observe that if i's income is fixed at z_i^*, then i's utility (in consumption terms) if i pays tax t_i is z_i^* is $z_i^* - t_i - v_i\left(z_i^*\right)$ and we have

$$z_i^* - t_i - v_i\left(z_i^*\right) \geq z_i^* - t_i' - v_i\left(z_i^*\right) \Leftrightarrow -t_i \geq -t_i'.$$

So with incomes fixed the negative of i's tax, $-t_i$, actually represents i's utility because i does not alter her income in response to the tax; in other words, i's utility only varies insofar as i pays more or less tax, so the negative of i's tax represents i's utility. If we define the function u_i via the condition $u_i\left(z_i^* - t_i - v_i\left(z_i^*\right)\right) = H_i\left(t_i\right)$, where H_i is defined by (23), and $u_i\left(z_i^* - t_i - v_i\left(z_i^*\right)\right)$ is decreasing in t_i, so that $u_i\left(z_i^* - t_i - v_i\left(z_i^*\right)\right)$ represents i's utility. Moreover, $\frac{\mathrm{d}}{\mathrm{d}t} u_i\left(z_i^* - t_i - v_i\left(z_i^*\right)\right) = g_i\left(t_i\right)$, so that welfare weights can be interpreted in terms of marginal utilities; we can now see that these are the welfare weights that result from the "utilitarian" social welfare function (25).

It is somewhat surprising that libertarian welfare weights can be captured by a welfarist objective. Intuitively, *reasons* justifying the comparisons encoded in the welfare weights are not welfarist. According to a (rights-based) libertarian view, it is bad to tax people more *because* it is unfair, rather than because it lowers their utility. However, it turns out that with *fixed incomes*, these fairness considerations can be fitted into a welfarist framework. This is because raising a person's taxes reduces their utility and, and when incomes are fixed, the framework cannot distinguish between a situation in which a person is given additional priority *because* their utility was reduced or *because* it was unfair to tax them. That we can fit this example into a welfarist framework and

derive the welfare weights from a utilitarian social welfare function also means that the GSMWW approach was not really necessary and we could have derived our evaluation of tax policies from a traditional social welfare function approach. The GSMWW approach really comes into its own when there is no such social welfare function.

It turns out that our ability to shoehorn the the libertarian weights into a utilitarian social welfare function and to rationalize them in the sense of Definition 2 is an artifact of the assumption that incomes are fixed. When incomes respond to taxes, we can vary taxes in such a way as to hold fixed a person's utility while varying the total tax paid. However, in this more general case, as we shall see in the next section, libertarian welfare weights are not rationalizable at all.

6 A simple example showing that welfare weights may not be rationalizable

In this section I present a simple example showing that social welfare weights may not in general be rationalizable. The example presented in this section presents the essence of the more general result, presented in Section 7, that welfare weights are in general not rationalizable when they are not of a utilitarian kind.

Suppose there are three types of agents, each with measure $\frac{1}{3}$. The types are distinguished by an observable characteristic x_i, which belongs to the set

$$\{\text{West Coast}, \text{Interior}, \text{East Coast}\}.$$

Ordinarily, geographic location is a characteristic that an individual can choose and so might be affected by taxes; however, I only choose these categories because they are easy to remember, so let us assume that they are immutable, and that an agent cannot change these characteristics in response to taxes. To fit this into our model, we may assume that

$$x_i = \begin{cases} \text{East Coast}, & \text{if } i \in \left[0, \frac{1}{3}\right), \\ \text{Interior}, & \text{if } i \in \left[\frac{1}{3}, \frac{2}{3}\right], \\ \text{West Coast}, & \text{if } i \in \left(\frac{2}{3}, 1\right], \end{cases}$$

where recall there is a uniform measure on $[0, 1]$. I abbreviate East Coast, Interior, and West Coast respectively by E, I, and W. There are no unobservable characteristics.

It is possible to condition taxes on the observable characteristic, but it is not relevant to payoffs. In particular, all types share the same cost of of earning income $v(z_i; E) = v(z_i; I) = v(z_i; W) =: v(z_i) := \frac{1}{2}z_i^2$.

I assume that welfare weights are libertarian and identical across types, so that for all $i \in [0, 1]$, welfare weights are of the form $g_i(t_i) = g(t_i)$, where g is increasing in the tax $t_i = z_i - c_i$ paid by each agent. These are welfare weights of the same form as was assumed in Section 5, except that here the precise weights are assumed the same for all agents i.

We now prove the following proposition

Proposition 6 *1. Under the above assumptions, there does not exist a preorder $\precsim$ that rationalizes g. Any binary relation that rationalizes g contains a cycle of the form*

$$T_0 \prec T_1 \sim T_2 \prec T_3 \sim T_0. \tag{29}$$

2. Moreover, there does not exist a preorder $\precsim$ that rationalizes g with Pareto indifference. Any binary relation that rationalizes g with Pareto indifference creates a cycle of the form (29).

So under these assumptions it is impossible to rationalize the welfare weights and the impossibility is robust to the precise definition of rationalization–rationalization in either the sense of Definition 2 or Definition 3 works in the result. Observe that in this example, we can condition taxes on observable characteristics and not just income. Allowing taxes to be conditioned on observable characteristics makes it easier to generate an example. In the next section, I will present a general result that shows that it is generally possible to construct such examples when taxes depend only on income. The general result will assume that agents are heterogenous in the senses that (i) they may face different costs of earning income and (ii) their welfare weight may depend on their characteristics as well as their consumption and income.

Let us now see why the result is true. Let us consider linear taxes of the form $T(z) = \alpha + \beta z$. If an agent with cost function $v(z) = \frac{1}{2}z^2$ faces such a tax schedule, the agent will solve the following problem

$$\max_z \left[z - (\alpha + \beta z) - \frac{1}{2}z^2 \right]$$

The first order condition is

$$(1 - \beta) - z = 0.$$

So if we define $z(\beta)$ as the optimal solution to the agent's problem when the agent faces a marginal tax rate of β, we have:

$$z(\beta) = 1 - \beta.$$

For any marginal tax rate β and utility level u, we can set select a transfer $\alpha(\beta, u)$ so that the agent's utility when facing taxes $\alpha(\beta, u) + \beta z$ is u. Formally, $\alpha(\beta, u)$ solves

$$z(\beta) - (\alpha(\beta, u) + \beta z(\beta)) - \frac{1}{2}[z(\beta)]^2 = u.$$

In particular[15]

$$\alpha(\beta, u) = \frac{1}{2}(1-\beta)^2 - u. \tag{30}$$

Let

$$T^u_\beta(z) = \alpha(\beta, u) + \beta z.$$

That is T^u_β is the unique tax policy that gives the agent a utility level of u at marginal tax rate β, given that the agent responds optimally to the policy. Note that for any pair of marginal tax rates, β_0 and β_1, agents are indifferent between facing the tax policies $T^u_{\beta_0}$ and $T^u_{\beta_1}$ because both generate utility u. However, these different tax policies do not generate the same revenue. While we can adjust the lumpsum tax $\alpha(\beta, u)$, so that he agent's utility is held fixed as we raise β and correspondingly lower α, the total taxes raised from the agent fall. This is because when marginal tax rates are lower, the agent works more and must be compensated for their efforts to keep their utility constant. For example, to achieve a utility level of zero, it is possible to (i) set the marginal tax rate equal to 1 and collect no transfer, raising no taxes (tax policy T^0_1), or (ii) set the marginal tax rate to 0 and collect a lump-sum tax of $\frac{1}{2}$ (tax policy T^0_0). While agents are indifferent between T^0_0 and T^0_1, there is a sense in which T^0_0 is clearly better because it raises more revenue, which can be used for beneficial purposes (which are not modeled formally).

The proof of Proposition 6 does require that we hold revenue fixed because the definition of rationalization concerns comparisons among tax policies raising equal revenue. Now consider the following family of tax policies $\tilde{T}_\beta$, where $\beta \in [0, 1]$:

$$\tilde{T}_\beta(z_i; x_i) = \begin{cases} \alpha(\beta, 0) + \beta z, & \text{if } x_i = \text{East Coast} \\ \alpha\left(\sqrt{1-\beta^2}, 0\right) + \left(\sqrt{1-\beta^2}\right) z_i, & \text{if } x_i = \text{Interior} \\ \frac{1}{4}, & \text{if } x_i = \text{West Coast} \end{cases}$$

This family of tax policies is as follows. Agents on the West Coast pay a tax of $\frac{1}{4}$ independently of their income. Agents on the East Coast pay a marginal tax rate β and those in the interior pay a marginal tax rate of $\sqrt{1-\beta^2}$ and in both cases the lumpsum transfer is adjusted to keep agents' utility at 0. So all agents are indifferent among all tax policies of the form $\tilde{T}_\beta$, as β varies between

[15] To see that (30) holds, observe that if we assume (30), we get

$$\begin{aligned} z(\beta) - (\alpha(\beta, u) + \beta z(\beta)) - \frac{1}{2}[z(\beta)]^2 &= (1-\beta) - \left(\left[\frac{1}{2}(1-\beta)^2 - u\right] + \beta(1-\beta)\right) - \frac{1}{2}(1-\beta)^2 \\ &= \left[1 - \frac{1}{2}(1-\beta) + \beta - \frac{1}{2}(1-\beta)\right](1-\beta) + u = u. \end{aligned}$$

0 and 1. Observe that

$$\begin{aligned} R\left(\tilde{T}_\beta\right) &= \frac{1}{3}\left[\frac{1}{2}(1-\beta)^2 + \beta(1-\beta)\right] + \frac{1}{3}\left[\frac{1}{2}\left(1-\sqrt{1-\beta^2}\right)^2 + \sqrt{1-\beta^2}\left(1-\sqrt{1-\beta^2}\right)\right] + \frac{1}{3}\frac{1}{4} \\ &= \frac{1}{3}\left[\frac{1}{2}(1+\beta)(1-\beta)\right] + \frac{1}{3}\left[\frac{1}{2}\left(1+\sqrt{1-\beta^2}\right)\left(1-\sqrt{1-\beta^2}\right)\right] + \frac{1}{3}\frac{1}{4} \\ &= \frac{2}{12}\left[\left(1-\beta^2\right) + \left(1-\left(1-\beta^2\right)\right)\right] + \frac{1}{12} = \frac{3}{12} \\ &= \frac{1}{4}. \end{aligned}$$

Observe that the revenue of $\tilde{T}_\beta$ is $\$\frac{1}{4}$ independently of the value of β. This is because as we raise β, the revenue from taxpayers on the east coast falls, but the revenue in the interior rises exactly so as to offset it.

Let $\precsim$ be a binary relation that represents g. Consider the parameterized collection of tax policies $\left(\tilde{T}_\beta\right)_{\beta\in[0,1]}$. It follows from the construction of taxes–the fact that for all agents utility is unchanged as β varies–and the envelope theorem that for all $\beta_0 \in [0,1]$, $0 = \frac{\mathrm{d}}{\mathrm{d}\beta}\Big|_{\beta=\beta_0} \tilde{U}_i\left(\tilde{T}_\beta\right) = -\frac{\mathrm{d}}{\mathrm{d}\beta}\Big|_{\beta=\beta_0} \tilde{T}_\beta\left(z\left(\beta_0\right)\right)$. This can also be confirmed by direct calculation. It follows that for all $\beta_0 \in [0,1]$,

$$\int g_i\left(\tilde{T}_{\beta_0}\right) \frac{\mathrm{d}}{\mathrm{d}\beta}\Big|_{\beta=\beta_0} \tilde{T}_\beta\left(z\left(\beta_0\right)\right) \mathrm{d}i = 0.$$

It follows from the indifference principle (10) that

$$\tilde{T}_0 \sim \tilde{T}_1. \tag{31}$$

Alternatively we can derive (31) from Pareto indifference if we assume instead that $\precsim$ rationalizes g with Pareto indifference.

Now define the tax reform

$$\Delta T\left(z\right), = \begin{cases} 1 & \text{if } i \in \text{East Coast} \\ 0, & \text{if } i \in \text{Interior} \\ -1, & \text{if } i \in \text{West Coast} \end{cases}$$

This tax reform amounts to a transfer of \$1 from the East Coast to the West Coast. Then for any $\varepsilon > 0$, the tax policy $\tilde{T}_\beta + \varepsilon\Delta T$ is the tax policy that results from $\tilde{T}_\beta$ by transferring $\$\varepsilon$ lumpsum from agents in the East Coast to agents in the West Coast. It is easy to see that for fixed ε, all agents are indifferent among all tax policies of the form $\tilde{T}_\beta + \varepsilon\Delta T$ as β varies; the reasoning is essentially the same as for $\tilde{T}_\beta$. Moreover since ΔT is just a lumpsum transfer that does not affect

behavior, we have

$$R\left(\tilde{T}_\beta + \varepsilon\Delta T\right) = R\left(\tilde{T}_\beta\right) = \frac{1}{4}, \ \ \forall\beta \in [0,1], \forall\varepsilon \in \mathbb{R}.$$

It then follows again from the indifference principle or, alternatively, from Pareto indifference, that

$$\tilde{T}_0 + \varepsilon\Delta T \sim \tilde{T}_1 + \varepsilon\Delta T. \tag{32}$$

Observe that an agent i with $x_i =$ East Coast pays a tax of $\$\frac{1}{2}$ under $\tilde{T}_0$ and a tax of \$0 under $\tilde{T}_1$. The respective calculations are:

$$\begin{aligned}
\tilde{T}_0\left(z_i\left(\tilde{T}_0\right); E\right) &= \alpha(0,0) + 0 \times z(0) = \frac{1}{2}(1-0)^2 = \frac{1}{2}, \\
\tilde{T}_1\left(z_i\left(\tilde{T}_1\right); E\right) &= \alpha(1,0) + 1 \times z(1) = \frac{1}{2}(1-1)^2 + 1 \times (1-1) = 0.
\end{aligned}$$

Observe that the tax paid by i with $x_i =$ West Coast is $\$\frac{1}{4}$ under both $\tilde{T}_0$ and $\tilde{T}_1$. Since weights $g(t)$ are increasing in taxes paid, we have

$$g(0) < g\left(\frac{1}{4}\right) < g\left(\frac{1}{2}\right).$$

So consider the parameterized the two collections of tax policies parameterized by ε, $\left(\tilde{T}_0 + \varepsilon\Delta T\right)_{\varepsilon\in[0,1]}$ and $\left(\tilde{T}_1 + \varepsilon\Delta T\right)_{\varepsilon\in[0,1]}$. We have

$$\begin{aligned}
\int g\left(\tilde{T}_0\right)\Delta T\left(z_i\left(\tilde{T}_0\right)\right)\mathrm{d}i &= \frac{1}{3}g\left(\frac{1}{2}\right) - \frac{1}{3}g\left(\frac{1}{4}\right) > 0, \\
\int g\left(\tilde{T}_0\right)\Delta T\left(z_i\left(\tilde{T}_1\right)\right)\mathrm{d}i &= \frac{1}{3}g(0) - \frac{1}{3}g\left(\frac{1}{4}\right) < 0.
\end{aligned}$$

It follows from the local improvement principle that (9) that for sufficiently small $\varepsilon > 0$, respectively,

$$\begin{aligned}
\tilde{T}_0 &\succ \tilde{T}_0 + \varepsilon\Delta T \\
\tilde{T}_1 &\prec \tilde{T}_0 + \varepsilon\Delta T.
\end{aligned} \tag{33}$$

In other words, transferring a dollar from the East Coast to the West Coast is

- *bad* at $\tilde{T}_0$ because at $\tilde{T}_0$ agents on the East Coast pay *more* in tax than those on the East Coast; and
- *good* at $\tilde{T}_1$ because at $\tilde{T}_1$ agents on the East Coast pay *less* in tax than those on the East Coast.

Putting together (31), (32), and (33), we get

$$\tilde{T}_1 \prec \left[\tilde{T}_1 + \varepsilon\Delta T\right] \sim \left[\tilde{T}_0 + \varepsilon\Delta T\right] \prec \tilde{T}_0 \sim \tilde{T}_1.$$

This establishes the desired cycle and completes the proof of Proposition 6.

This shows that in general GSMWW do not provide a coherent way of ranking tax policies. Perhaps the proponent of such weights would reply that it is not the aim to provide a full ranking but just to find an optimum. However, I think that if the implied ranking is incoherent, then there is no basis for trusting GSMWW's verdict on an optimum. After all an optimum amounts to a series of pairwise comparisons of rank with other tax policies.

In the next section, I present a much more general version of the result.

7 The main result: welfare weights of a non-utilitarian kind are inconsistent

This section presents my main result, which is a much more general version of the example presented in the presented in the previous section. What I show is that whenever welfare weights do not take an essentially utilitarian form, they are not rationalizable and it is possible to construct a cycle. Thus it is essentially precisely in those cases in which the GSMWW approach is more general than the standard utilitarian approach that it delivers inconsistent rankings. In order to prove this result I will make some additional assumptions about the form of the agents utility functions, but the model will remain quite general.

Suppose that we can parameterize the cost of effort in terms of a one dimensional parameter $\phi \in \left[\underline{\phi}, \overline{\phi}\right] \subseteq \mathbb{R}$ with $\underline{\phi} < \overline{\phi}$. A higher value of ϕ corresponds to a lower cost of effort. Rather than expressing the agent's cost of earning income as $v(z; x, y)$, we assume that we can express it as $v(z; \phi)$, where $v : \mathbb{R} \times \left[\underline{\phi}, \overline{\phi}\right] \to \mathbb{R}$ is a three times continuously differentiable function. I make the following assumptions on v:

$$\begin{aligned} \forall z \in Z, \forall \phi \in \left[\underline{\phi}, \overline{\phi}\right], \quad & \underbrace{\frac{\partial}{\partial z} v(z; \phi) > 0}_{(i)}, \quad \underbrace{\frac{\partial^2}{\partial \phi \partial z} v(z; \phi) < 0}_{(ii)}, \quad \underbrace{\frac{\partial^2}{\partial z^2} v(z, \phi) > 0}_{(iii)}, \\ & \underbrace{\frac{\partial}{\partial z} v\left(0, \underline{\phi}\right) < 1}_{(iv)}, \quad \underbrace{\frac{\partial}{\partial z} v\left(\bar{z}, \overline{\phi}\right) > 1}_{(v)}, \end{aligned} \tag{34}$$

Condition (i) says that earning income is costly; (ii) says that the cost function is convex for each type; (iii) says that earning income is less costly for higher types; (iv) says that the even the highest cost type will find it worthwhile to earn positive income if there are no taxes; and similarly (v) says that the even the lowest cost type will only find it worthwhile to earn less income than the technically highest possible level $\bar{z}$ if there are not taxes. In what follows, I sometimes write $v_\phi(z)$ instead of $v(z; \phi)$. I assume a probability density f on types $\left[\underline{\phi}, \overline{\phi}\right]$ such that $f(\phi) > 0, \forall \phi \in \left[\underline{\phi}, \overline{\phi}\right]$.

Similarly welfare weights are now assumed to depends on ϕ rather than (x, y), so that I write

welfare weights as $g(c,z,\phi)$ rather than $g(c,z;x,y)$. I also write $g_\phi(c,z) = g(c,z,\phi)$. I assume that $g(c,z,\phi)$ is three times continuously differentiable.

In this section, I assume that taxes depend *only on income* z and not on personal characteristics, so that we can simply write $T(z)$ rather than $T(z,x)$ or $T(z,\phi)$. This makes it *harder* to prove my main result; it would be easier to construct a cycle if I could condition taxes on characteristics. In this section, I also find it convenient to restrict the class of tax policies a bit further. For each $\phi \in \left[\underline{\phi}, \overline{\phi}\right]$, let z_ϕ solve

$$v'_\phi(z_\phi) = 1 \tag{35}$$

This z_ϕ is the income that type ϕ would earn in the absence of taxes. In particular $z_{\underline{\phi}}$ is the income that the lowest type $\underline{\phi}$ would choose to earn in the absence of taxes. I restrict attention to tax policies for which the marginal tax rate below $z_{\underline{\phi}}$ is 0. I also restrict attention to tax policies for which the marginal tax rate is never more than 1, and I assume that the set of tax policies contains all convex tax policies with these properties. Formally, if we let $\mathscr{T}$ be the set of three times continuously differentiable functions from Z to $\mathbb{R}$, then in this section, I assume that the set of admissible tax policies $\hat{\mathscr{T}}$ has the following properties:

$$\begin{aligned}&\mathscr{S} \cap \left\{T \in \mathscr{T} : \frac{\mathrm{d}^2}{\mathrm{d}z^2} T\left(z'\right) \leq 0\right\} \subseteq \hat{\mathscr{T}} \subseteq \mathscr{S} \\ &\text{where } \mathscr{S} = \left\{T \in \mathscr{T} : \frac{\mathrm{d}}{\mathrm{d}z} T\left(z'\right) = 0, \forall z' \in \left[0, z_{\underline{\phi}}\right]; 0 \leq \frac{\mathrm{d}}{\mathrm{d}z} T\left(z'\right) \leq 1, \forall z' \in \left[z_{\underline{\phi}}, \bar{z}\right]\right\}.\end{aligned} \tag{36}$$

Moreover, I assume that for all $T \in \hat{\mathscr{T}}$, the agent's problem has a unique maximizer $z_\phi(T)$ for all types T; this condition holds in $\mathscr{S}$. Subject to these adjustments, the assumptions on the set $\hat{\mathscr{T}}$ of tax policies are the same as in Section 2, and likewise the definition of the set of smooth paths P holding revenue fixed is the same as in Section 4. For each $\phi \in \left[\underline{\phi}, \overline{\phi}\right]$, I define $Z_\phi = \left[z_{\underline{\phi}}, z_\phi\right]$. Thus Z_ϕ is the set of income levels that type ϕ might earn given some tax policy in $\hat{\mathscr{T}}$.

Now, analogously to the definitions in Section 2, define

$$\begin{aligned}z_\phi(T) &\in \arg\max_{z \in Z} z - T(z) - v_\phi(z) \\ U_\phi(T) &= u\left(z_\phi(T)\right) - T\left(z_\phi(T)\right) - v_\phi\left(z_\phi(T)\right) \\ g_\phi(T) &= g_\phi\left(z_\phi(T) - T\left(z_\phi(T)\right), z_\phi(T)\right)\end{aligned}$$

Similarly for any $\rho \in P$ and $\theta \in [a,b]$, define

$$z^\rho_\phi(\theta) = z_\phi\left(T^{\rho,\theta}\right)$$

I now repeat the definition of rationalizability in this new setting, which is essentially identical to Definition 2, except that it incorporates types ϕ and the density f over types.

Definition 4 *Let g be a system of welfare weights. Say that a binary relation $\precsim$* ***rationalizes*** *g if for all $\rho \in P$,*

local improvement principle:

$$\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(T^{\rho,0}\right) \left.\frac{\partial}{\partial\theta}\right|_{\theta=0} T^\rho\left(z_\phi^\rho(0),\theta\right) f(\phi)\,\mathrm{d}\phi < 0 \Rightarrow \left[\exists\bar{\theta}\in(0,1],\forall\theta\in\left(0,\bar{\theta}\right), T^{\rho,0} \prec T^{\rho,\theta}\right], \tag{37}$$

$$\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(T^{\rho,0}\right) \left.\frac{\partial}{\partial\theta}\right|_{\theta=0} T^\rho\left(z_\phi^\rho(0),\theta\right) f(\phi)\,\mathrm{d}\phi > 0 \Rightarrow \left[\exists\bar{\theta}\in(0,1],\forall\theta\in\left(0,\bar{\theta}\right), T^{\rho,0} \succ T^{\rho,\theta}\right],$$

indifference principle:

$$\text{and } \left(\forall\theta'\in[0,1], \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(T^{\rho,\theta'}\right) \left.\frac{\partial}{\partial\theta}\right|_{\theta=\theta'} T^\rho\left(z_\phi^\rho\left(\theta'\right),\theta\right) f(\phi)\,\mathrm{d}\phi = 0\right) \Rightarrow T^{\rho,0} \sim T^{\rho,1}. \tag{38}$$

Say that the system of welfare weights g is ***rationalizable*** *if there exists a preorder that rationalizes g.*

For each $\phi \in \left[\underline{\phi},\overline{\phi}\right]$, let $z_\phi \in Z$ solve $v'_\phi(z_\phi) = 1$. Thus z_ϕ is the income level that type ϕ would choose in the absence of taxes. Let

Definition 5 *Say that g* ***depends only on utility*** *if and only if $\forall\phi \in \left[\underline{\phi},\overline{\phi}\right], \forall z,z' \in Z_\phi, \forall c,c' \in C$,*

$$c - v(z;\phi) = c' - v\left(z',\phi\right) \Rightarrow g_\phi(c,z) = g_\phi\left(c',z'\right) \tag{39}$$

To understand this definition, observe that $c - v(z;\phi)$ is a representation of the agent's utility. So (39) says that the welfare weight of agent of type ϕ depends only on the utility of type ϕ; it means that in two allocations (z,c) and (z',c') in which type ϕ has the same utility, type ϕ gets the same welfare weight. It also means that in two allocations (z,c) and (z',c') in which type ϕ has the same marginal utility for consumption *for every possible choice of the function* u,[16] type ϕ has the same welfare weight. Observe that (39) is satisfied in the standard utilitarian case, where

$$g_\phi(c,z) = u'(c - v(z,\phi)).$$

It might be justified to call welfare weights g that depend only on utility *utilitarian*, because whenever weights depend only on utility, there exists type-dependent utility functions $u_\phi : \mathbb{R} \to \mathbb{R}$ and $w_\phi : \mathbb{R} \times Z \to \mathbb{R}$ such that

$$w_\phi(c,z) = u_\phi(c - v_\phi(z))$$

[16] If we restrict attention to strictly convex u, then we can change the word "every" to "any" in the above statement.

and

$$g_\phi = \frac{\partial w_\phi}{\partial c};$$

and, moreover, that weights g_ϕ are those that are induced by the social welfare function $W = \int w_\phi f(\phi)\,\mathrm{d}\phi$.[17] So such weights can be thought of as coming from a utilitarian social welfare function. However, one might instead prefer to call weights that depend only on utility in the sense of Definition 5 *quasi-utilitarian*, rather than utilitarian, in that the Definition 5 does not preclude the possibility the implicit choice of the function u_ϕ can depend on non-utilitarian ethical considerations.

This terminological and interpretive choice resonates with two comments from Fleurbaey and Maniquet (2018), which concern their social welfare function approach (rather than the local social welfare weight approach); Fleurbaey and Maniquet write, "Our defense of the social welfare function could even be understood as a defense of the utilitarian approach, for an ecumenical notion of utilitarianism that is flexible about the degree of inequality aversion and the definition of individual utility." (p. 1031) This ecumenical sense of utilitarianism supports the terminology "utilitarian" for welfare weights that satisfy (39). However Fleurbaey and Maniquet also write "... we highlight another way in which at least some fairness principles can remain compatible with the Pareto principle. Fairness principles can indeed guide the selection of the individual utility functions that serve to measure well-being and perform interpersonal comparisons." (p. 1040) That is, non-utilitarian "fairness" considerations can guide the choice of weights g_ϕ compatible with (39); Fleurbaey and Maniquet provide a number of examples; other examples of this principle include Piacquadio (2017) and Berg and Piacquadio (2020). The "utilitarian" social welfare function constructed in Section 5 also has this flavor. From this perspective, the term "quasi-utilitarian" may be better (although note that Piacquadio (2017) refers to his approach as "a fairness justification *of utilitarianism*").

Whatever the appropriate way of interpreting condition (39), it is the critical condition for the rationalizability of welfare weights.

Theorem 1 *Under the assumptions of this section, g depends only on utility if and only if g is rationalizable.*

So the theorem shows that, outside of the essentially utilitarian case, welfare weights are not rationalizable. Outside of the utilitarian case, the proof of Theorem 1 shows that if the binary relation $\precsim$ rationalizes g, it is possible to construct a sequence of tax policies T_0, T_1, T_2, T_3, which generate the same revenue, such that they generate a cycle:

$$T_0 \sim T_1 \prec T_2 \sim T_3 \prec T_0.$$

[17] One might wonder how we can always attain a separable social welfare function whenever welfare weights are rationalizable; the answer is that type θ's welfare weight is assumed to depend only on type θ's consumption, income, and characteristics, and not on the distribution of these in society.

The proof essentially shows that whenever welfare weights don't depend only on utility, it is possible to construct an example similar to the one found in Section 6. Of course, the argument is much more general than that found in Section 6, and the construction more intricate. However, the example of Section 6 captures of the spirit of the construction in the proof of the main theorem, Theorem 1. The key point is that this theorem shows that the generalized welfare weight approach is problematic precisely in those cases in which welfare weights come into their own and purport to go beyond the standard approach.

Remark 2 *The specific condition that welfare weights depend only on utility depends on the fact that we have assumed that utility takes the quasilinear form $u(c - v(z))$, so that there are no income effects. In a more in which utility took the more general form $u(c, z)$, then the statement of the theorem would have to be modified, and the necessary and sufficient condition for rationalizability would no longer be (39). I believe that in that case an analogous result would hold: that outside of an essentially "utilitarain case", welfare weights would not be rationalizable. However, this is not something I have yet proven.*

8 Discussion

I conclude this paper with some thoughts about how broader values ought to be represented. Consider libertarianism as an example.[18] Suppose that one thinks that people are entitled to their pre-tax incomes and in some way taxation is like theft. This view is not faithfully rendered as saying that additional income to people 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 $c_i(t_i)$ for each agent i, that measures how bad it is to violate i's entitlements. Then the social objective is given by

$$W(T) = -\int_i c_i(T_i(z_i(T)))\,\mathrm{d}i. \tag{40}$$

Note that $c_i(t_i)$ is not derived from agents' preferences but rather measures a social judgement about how bad it is to violate the person's entitlements. One could also consider a more pluralist value function, which incorporates both utilitarian and libertarian considerations.

$$W(T) = \alpha \int_i U_i(T)\,\mathrm{d}i - (1-\alpha)\int_i c_i(T_i(z_i(T)))\,\mathrm{d}i, \tag{41}$$

where α measures the weight put on utilitarianism rather than libertarianism. $\alpha = 0$ corresponds to pure libertarianism (as in (40)), $\alpha = 1$ to pure utilitarianism. If Ann and Bob respectively make

[18] For different approaches to libertarian taxation, see Nozick (1974), Feldstein (1976), Young (1987), Weinzierl (2014), and Vallentyne (2018).

social judgements with parameters α_0 and α_1 with $\alpha_0 < \alpha_1$, that means that Ann places more weight on libertarian considerations than Bob (relative to utilitarian considerations). One could then maximize (41) subject to a revenue requirement. The analog to the local desirability of a tax reform ΔT (13) with objective (41) is

$$\int_i \left\{ \left[\alpha g_i \left(T \left(z_i \left(T \right) \right) \right) + \left(1 - \alpha \right) c_i' \left(T_i \left(z_i \left(T \right) \right) \right) \right] \Delta T_i \left(z_i \left(T \right) \right) \right.$$
$$\left. + \left(1 - \alpha \right) c_i' \left(T_i \left(z_i \left(T \right) \right) \right) T_i' \left(z_i \left(T \right) \right) \left. \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \right|_{\varepsilon = 0} z_i \left(T + \varepsilon \Delta T \right) \right\} \mathrm{d}i < 0,$$

where

$$g_i \left(T \right) = u_i' \left(z_i \left(T \right) - T \left(z_i \left(T \right) \right) - v_i \left(z_i \left(T \right) \right) \right)$$

is the utilitarian welfare weight. It follows that under this system, changes in tax policies are not evaluated by just putting a weight on changes in taxes ΔT. There is no system of welfare weights that is equivalent to the social welfare function (41). At the same time, judgments rendered by (41) must be consistent and rationalizable because they come from a global objective.

One potential criticism of the objective (41) is that it will lead to Pareto inefficient optima. However, the reply to this is that that is what is involved in taking rights and other values seriously: These other values will compete with preference satisfaction, and one may want to sacrifice some measure of preference satisfaction to realize these other values. This point was made by Sen (1979). Now some economists may view such violations of Pareto as unacceptable, and argue that any change that improves everyone's preferences must be deemed as a social improvement. That may be true (or not), but in making such a claim, innocuous as it may sound, economists become moral philosophers and it is then incumbent on them to provide a philosophical defense. Papers that criticize the Pareto principle include Mongin (1997/2016) and Sher (2020).

There is another potential alternative way that broader values may enter into economic analysis suggested by the above analysis. Theorem 1 shows that welfare weights are rationalizable provided they take the form $g_\phi \left(c, z \right) = \tilde{g}_\phi \left(c - v_\phi \left(z \right) \right)$. g_ϕ may be marginal utility for consumption but it need not be. Perhaps the dependence of welfare weights of this form on characteristics ϕ can allow us to incorporate some broader ethical values, while preserving Pareto. This possibility was discussed above in Sections 3, 5, and 7, and it is a possible way of incorporating broader values, but I do not think that we should expect it to succeed in capturing *all* values. Again thinking of the libertarian case, if we are thinking about violations of people's entitlements, it does not intuitively seem that what is at stake is just how much weight we should put on different people's income; likewise, if we think about merit, fairness, or procedural values, it is not clear why their representation should be of this form. To fully address these issues would take us beyond the scope of the current paper and would entail considering other literatures. However, the consideration of what can and cannot be done with welfare weights puts these issues in sharper focus.

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.

Bossert, W. and Weymark, J. A. (2004), Utility in social choice, *in* 'Handbook of utility theory', Kluwer Academic Publishers, pp. 1099–1177.

d'Aspremont, C. and Gevers, L. (1977), 'Equity and the informational basis of collective choice', *The Review of Economic Studies* **44**(2), 199–209.

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.

Hammond, P. J. (1979), 'Equity in two person situations: some consequences', *Econometrica* pp. 1127–1135.

Hartman, P. (1987), *Ordinary Differential Equations*, SIAM.

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.

Lang, S. (2012), *Real and functional analysis*, Springer.

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.

Roberts, K. W. (1980), 'Interpersonal comparability and social choice theory', *The Review of Economic Studies* pp. 421–439.

Saez, E. (2001), 'Using elasticities to derive optimal income tax rates', *The review of economic studies* **68**(1), 205–229.

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'.

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

Sen, A. (1977), 'On weights and measures: informational constraints in social welfare analysis', *Econometrica* pp. 1539–1572.

Sen, A. (1979), 'Utilitarianism and welfarism', *The Journal of Philosophy* **76**(9), 463–489.

Sen, A. (1986), 'Social choice theory', *Handbook of mathematical economics* **3**, 1073–1181.

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.

Weymark, J. A. (2016), Social welfare functions, *in* M. D. Adler and M. Fleurbaey, eds, 'The Oxford Handbook of Well-Being and Public Policy', Oxford University Press, pp. 126–159.

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.

A Appendix

A.1 Preliminaries

Consider welfare weights of the form $g_\phi(c,z)$. Define the variable u by $u = c - v_\phi(z)$. So $c = u + v_\phi(z)$. Then we can re-express welfare weights in terms of u and z, rather than in terms of c and z as follows.

$$\hat{g}_\phi(u,z) = g_\phi(u + v_\phi(z), z), \quad \forall u \in \mathbb{R}, \forall z \in Z. \tag{A.1}$$

Formally, welfare weights can then be represented as a function $\hat{g} : \mathbb{R} \times Z \times \left[\underline{\phi}, \overline{\phi}\right] \to \mathbb{R}_+$ where I often write $\hat{g}_\phi(u,z)$ instead of $\hat{g}(u,z,\phi)$.

Notice that the two formulations–in terms of u and z and in terms of c and z–are equivalent because u is recoverable from c, z, and ϕ, and c is recoverable from u, z and ϕ.

Proposition A.1 *Let g and $\hat{g}$ be related as in (A.1). Then the following conditions are equivalent:*[19]

1. *g depends only on utility.*
2. *$\forall \phi \in \left[\underline{\phi}, \overline{\phi}\right], \forall u \in \mathbb{R}, \ \forall z, z' \in Z_\phi, \ \ \hat{g}_\phi(u,z) = \hat{g}_\phi(u,z')$.*
3. *$\forall \phi \in \left[\underline{\phi}, \overline{\phi}\right], \forall u \in \mathbb{R}, \forall z \in Z_\phi, \frac{\partial}{\partial z}\hat{g}_\phi(u,z) = 0$.*

Proof. First I argue that condition 1 of the proposition implies condition 2. First that g depends only on utility. Now choose $\phi \in \left[\underline{\phi}, \overline{\phi}\right], z, z' \in Z_\phi$, and $u \in \mathbb{R}$. Define $c = u + v_i(z)$ and $c' = u + v_i(z')$. Then observe that $c - v_i(z) = u = c' - v_i(z')$. So by (A.1), $\hat{g}_\phi(u,z) = g_\phi(c,z) = g_\phi(c',z') = \hat{g}_\phi(u,z')$, where the middle equality follows from the assumption that g depends only on utility. It follows that condition 2 of the proposition holds.

Next I argue that condition 2 implies condition 1. So assume condition 2. Choose $\phi \in \left[\underline{\phi}, \overline{\phi}\right], c, c' \in \mathbb{R}$, and $z, z' \in Z_\phi$ and assume that $u = c - v(z;\phi) = c' - v(z';\phi) = u'$. It follows that $g_\phi(c,z) = \hat{g}_\phi(u,z) = \hat{g}_\phi(u,z') = \hat{g}_\phi(u',z') = g_\phi(c',z')$, where the second equality follows from condition 2. This establishes condition 1.

Finally, consider the equivalence of conditions 2 and 3. First observe that by continuous differentiability condition 2 implies: $\forall \phi \in \left[\underline{\phi}, \overline{\phi}\right], \forall u \in \mathbb{R}, \forall z \in Z_\phi, \frac{\partial}{\partial z}\hat{g}_\phi(u,z) = 0$. The equivalence now follows from the fundamental theorem of calculus. □

Given the proposition, observe that

$$g_\phi(T) = g_\phi(U_\phi(T), z_\phi(T)) \tag{A.2}$$

The Proposition has the following corollary:

Corollary A.1 *If g does not depend only on utility, then there exists three times continuously differentiable tax policy $T(z)$ in $\hat{\mathscr{T}}$, which is strictly increasing and strictly convex on $\left[\underline{z_\phi}, \bar{z}\right]$, and is such that either*

$$\forall \phi \in (\phi_1, \phi_2), \frac{\partial}{\partial z}\hat{g}_\phi(U_\phi(T), z_\phi(T)) < 0 \tag{A.3}$$

or

$$\forall \phi \in (\phi_1, \phi_2), \frac{\partial}{\partial z}\hat{g}_\phi(U_\phi(T), z_\phi(T)) > 0 \tag{A.4}$$

[19] Recall that we have assumed that $g(c,z;\phi)$ is continuously differentiable in $(c,z;\phi)$.

Proof. Assume that g does not depends on utility. It follows from Proposition A.1 that there exists $\phi^\circ \in \left(\underline{\phi}, \overline{\phi}\right)$, $z^\circ \in \left(z_{\underline{\phi}}, z_{\phi^\circ}\right), u^\circ \in \mathbb{R}$ and such that

$$\frac{\partial}{\partial z}\hat{g}_{\phi^\circ}\left(u^\circ, z^\circ\right) \neq 0. \tag{A.5}$$

It follows from (34) that $1 - v'_{\phi^\circ}\left(z^\circ\right) > 0$. Let $r^\circ = 1 - v'_{\phi^\circ}\left(z^\circ\right)$. Then observe that z° is the unique maximizer of $z\left(1 - r^\circ\right) - v_{\phi^\circ}\left(z\right)$ over Z, where this follows form the fact that the objective in this optimization is strictly concave. It follows that if we choose any three times continuously differentiable tax policy T in $\hat{\mathscr{T}}$ that is strictly convex on $\left[z_{\underline{\phi}}, \bar{z}\right]$ and such that $T'\left(z^\circ\right) = r^\circ$, then $z_{\phi^\circ}\left(T\right) = z^\circ$. Observe that strict convexity of T on $\left[z_{\underline{\phi}}, \bar{z}\right]$ together with the fact that since $T \in \hat{\mathscr{T}}$, $\frac{\partial}{\partial z}T\left(z_{\underline{\phi}}\right) = 0$ (see (36)) imply that T is strictly increasing on $\left[z_{\underline{\phi}}, \bar{z}\right]$. By adding a lumpsum transfer, we can insure that $U_\phi\left(T\right) = u^\circ$. (A.5) now implies that either (A.3) or (A.4) holds. □

Let P_0 be the set of all paths $\rho : [a, b] \to \hat{\mathscr{T}}$ (where recall $a < 0, 1 < b$) that satisfy condition i at the beginning of Section 4 but not necessarily condition iii; note that because, in Section 7, we assume that there are no observable characteristics, condition ii is vacuously satisfied.[20] Observe that that the difference between paths P and P_0 is that paths ρ in P hold revenue fixed, whereas paths in P_0 need not satisfy this requirement. Thus $P \subset P_0$.

Definition 6 *Say that a binary relation $\precsim$ on $\hat{\mathscr{T}}$ **strongly rationalizes** g if the local improvement principle (37) and the indifference principle (38) are satisfied for all paths ρ in P_0, and g is **strongly rationalizable** if there exists a preorder $\precsim$ that strongly rationalizes g.*

Strong rationalizability differs from rationalizability (as in Definition 4) only in that in the former P_0 is substituted for P; or in other words, that the conditions (37) and (38) are required to hold on all smooth paths, not just those on which revenue is held constant. It follows that if g is strongly rationalizable, then g is rationalizable, but the converse does not hold in general. By contraposition, if g is not rationalizable.

A.2 The core of the proof of the main theorem

I first prove a weaker version of one direction of the theorem.

Proposition A.2 *Under the assumptions of section 7, if g does not depend only on utility then g is not strongly rationalizable.*

This proposition is weaker than the main theorem in two ways: first, it is only one direction–the harder direction–and also since it employs strong rationalizability rather than rationalizability, it does not require that revenue is held constant along a path. Below, in Proposition A.3, I will strengthen the result to hold revenue constant. However, I separate that from the rest of the argument because it introduces complexity but does not really engage with the main ideas in the proof–in the example of Section 6, holding revenue constant was accomplished simply by varying the revenue in one part of the tax schedule to offset changes in another part; the same sort of idea works here.

So to prove Proposition A.5, I assume that g does not depend only on utility, and I argue that g is not strongly rationalizable. Appealing to Corollary A.1, I consider the case in which there exists $\phi_1 < \phi_2$ and three times continuously differentiable tax policy T in $\hat{\mathscr{T}}$, which is strictly increasing and strictly convex on $\left[z_{\underline{\phi}}, \bar{z}\right]$, and such that (A.3) holds; the case in which (A.4) instead of (A.3) holds is similar.

[20] Also, in condition (i), $T_i^\rho\left(z_i, \theta\right) = T^\rho\left(z_i, \theta\right) \forall i$.

Let $z_1 = z_{\phi_1}(T), z_2 = z_{\phi_2}(T)$, so that $z_1 < z_2$. Now consider a three times continuously differentiable real-valued function $\nu(z)$ with with domain $[0, \bar{z}]$ and support $[z_1, \bar{z}]$ and such that its derivative ν' has support $[z_1, z_2]$ and $\nu'(z) > 0, \forall z \in (z_1, z_2)$. For each $\gamma \in \mathbb{R}$, define the tax policy

$$T_\gamma = T + \gamma\nu.$$

Observe that T_γ is three times continuously differentiable and, if $|\gamma|$ is sufficiently small, then T_γ is strictly increasing and strictly convex on $\left[z_{\underline{\phi}}, \bar{z}\right]$.

Next choose $\hat{\phi}_1, \hat{\phi}_2 \in \left(\underline{\phi}, \overline{\phi}\right)$ with $\hat{\phi}_1 < \hat{\phi}_2 < \phi_1$ and let $\hat{z}_1 = z_{\hat{\phi}_1}(T), \hat{z}_2 = z_{\hat{\phi}_2}(T)$. It follows that $\hat{z}_1 < \hat{z}_2 < z_1$. Let $\mu(z)$ be a three times continuously differentiable function with domain $[0, \bar{z}]$ and support $[\hat{z}_1, \bar{z}]$ and such that μ' has support $[\hat{z}_1, \hat{z}_2]$ with $\mu'(z) > 0$ on $(\hat{z}_1, \hat{z}_2)$ and $\mu(z) = 1$ on $[\hat{z}_2, \bar{z}]$.

It follows from the Picard-Lindelöf theorem[21] that there exists an interval $[-\gamma^*, \gamma^*]$ (with $\gamma^* > 0$) and a unique function $\zeta(\gamma) : [-\gamma^*, \gamma^*] \to \mathbb{R}$ such that $\zeta(0) = 0$ and

$$\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T_\gamma - \zeta(\gamma)\mu)\left[-\frac{\mathrm{d}}{\mathrm{d}\gamma}\zeta(\gamma)\mu(z_\phi(T_\gamma - \zeta(\gamma)\mu)) + \nu(z_\phi(T_\gamma - \zeta(\gamma)\mu))\right] f(\phi)\,\mathrm{d}\phi = 0. \tag{A.6}$$

To interpret this expression, note that $T_\gamma - \zeta(\gamma)\mu$ is the tax policy defined by $[T_\gamma - \zeta(\gamma)\mu](z) = T_\gamma(z) - \zeta(\gamma)\mu(z), \forall z \in [0, \infty)$. Note that for sufficiently small γ, $T_\gamma - \zeta(\gamma)\mu$ is strictly increasing and strictly convex on $\left[z_{\underline{\phi}}, \bar{z}\right]$. We may assume that γ^* is sufficiently small that for all $\gamma \in [-\gamma^*, \gamma^*]$, T_γ and $T_\gamma - \zeta(\gamma)\mu$ are strictly increasing and strictly convex on $\left[z_{\underline{\phi}}, \bar{z}\right]$.

We can rewrite (A.6) as

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}\zeta(\gamma) = \frac{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T_\gamma - \zeta(\gamma)\mu)\,\nu(z_\phi(T_\gamma - \zeta(\gamma)\mu))\,f(\phi)\,\mathrm{d}\phi}{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T_\gamma - \zeta(\gamma)\mu)\,\mu(z_\phi(T_\gamma - \zeta(\gamma)\mu))\,f(\phi)\,\mathrm{d}\phi}.$$

[21] In particular, observe that if we define $F(\gamma, \zeta) : \mathbb{R}^2 \to \mathbb{R}$ by

$$F(\gamma, \zeta) = \frac{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T_\gamma - \zeta\mu)\,\nu(z_\phi(T_\gamma - \zeta\mu))\,f(\phi)\,\mathrm{d}\phi}{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T_\gamma - \zeta\mu)\,\mu(z_\phi(T_\gamma - \zeta\mu))\,f(\phi)\,\mathrm{d}\phi},$$

then F is continuously differentiable in (γ, ζ) in a neighborhood N of $(\gamma, \zeta) = (0, 0)$. It follows that on any closed set within N, F is Lipschitz continuous.

Using the fact that when $\gamma = 0$, $T_\gamma - \zeta(\gamma)\mu = T$, we have

$$\begin{aligned}
\left.\frac{\mathrm{d}}{\mathrm{d}\gamma}\right|_{\gamma=0}\zeta(\gamma) =& \frac{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T)\,\nu(z_\phi(T))\,f(\phi)\,\mathrm{d}\phi}{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T)\,\mu(z_\phi(T))\,f(\phi)\,\mathrm{d}\phi} \\
=& \left[\int_{\underline{\phi}}^{\phi_1} g_\phi(T)\,\nu(z_1)\,f(\phi)\,\mathrm{d}\phi + \int_{\phi_1}^{\phi_2} g_\phi(T)\,\nu(z_\phi(T))\,f(\phi)\,\mathrm{d}\phi\right. \\
&\left. + \int_{\phi_2}^{\overline{\phi}} g_\phi(T)\,\nu(z_2)\,f(\phi)\,\mathrm{d}\phi\right] \Big/ \int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T)\,\mu(z_\phi(T))\,f(\phi)\,\mathrm{d}\phi \\
=& \left[\int_{\underline{\phi}}^{\phi_1} g_\phi(T)\,\nu(z_1)\,\mu(z_\phi(T))\,f(\phi)\,\mathrm{d}\phi + \int_{\phi_1}^{\phi_2} g_\phi(T)\,\nu(z_\phi(T))\,\mu(z_\phi(T))\,f(\phi)\,\mathrm{d}\phi\right. \\
&\left. + \int_{\phi_2}^{\overline{\phi}} g_\phi(T)\,\nu(z_2)\,\mu(z_\phi(T))\,f(\phi)\,\mathrm{d}\phi\right] \Big/ \int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T)\,\mu(z_\phi(T))\,f(\phi)\,\mathrm{d}\phi
\end{aligned} \tag{A.7}$$

To understand the second equality, observe that by our assumptions on ν, $\nu(z) = 0, \forall z \in [0, z_1]$. This explains the first integral in the numerator of the right-hand side. Similarly, our assumptions imply that for all $z \in [z_2, \bar{z}]$, $\nu(z) = \nu(z_2)$, from which it follows that for all $\phi \in \left[\phi_2, \overline{\phi}\right]$, $\nu(z_\phi(T_\gamma)) = \nu(z_2)$; this explains the third integral of the numerator of the right-hand side of the second equality. To understand the third equality, observe that $\nu(z_1) = 0$, so adding $\mu(z_\phi(T))$ to the first integral in the numerator has no effect; also when $\phi > \phi_1$, $z_\phi(T) > z_1 > \hat{z}_2$, so $\mu(z_\phi(T)) = 1$, which justifies adding $\mu(z_\phi(T))$ to the second and third integrals in the numerator in the third equality.

Observe moreover that it follows from the facts that (i)

$$\{\nu(z_\phi(T)) : \phi \in (\phi_1, \phi_2)\} = (\nu(z_1), \nu(z_2)),$$

(ii) ν is strictly increasing on $[z_1, z_2]$, and (iii) $g_\phi(T)\,\mu(z_\phi(T))\,f(\phi) \geq 0$, for all $\phi \in \left(\underline{\phi}, \overline{\phi}\right)$ and $\exists \phi' < \phi_1$ such that for all $\phi \in \left(\phi', \overline{\phi}\right)$, $g_\phi(T)\,\mu(z_\phi(T))\,f(\phi) > 0$, and (A.7) that there exists $\hat{\phi} \in (\phi_1, \phi_2)$ such that

$$\left.\frac{\mathrm{d}}{\mathrm{d}\gamma}\right|_{\gamma=0}\zeta(\gamma) = \nu\left(z_{\hat{\phi}}(T)\right). \tag{A.8}$$

Let $\hat{z} = z_{\hat{\phi}}(T)$.

Now consider two three times continuously differentiable functions η_1 and η_2 such that η_1 has support $[z_1', z_2']$, η_2 has support $[z_1'', z_2'']$, where $z_1', z_2', z_1'', z_2'' \in \mathbb{R}_+$ and $z_1' < z_2' < \hat{z}_1 < z_1 < z_1'' < \hat{z} < z_2'' < z_2$ and there exist $\phi_1', \phi_2', \phi_1'', \phi_2'' \in \left(\underline{\phi}, \overline{\phi}\right)$ such that $z_1' = z_{\phi_1'}(T_0), z_2' = z_{\phi_2'}(T_0), z_1'' = z_{\phi_1''}(T_0)$, and $z_2'' = z_{\phi_2''}(T_0)$; it follows that $z_{\underline{\phi}} < z_1'$. I also assume that $\eta_2(z) < 0, \forall z \in (z_1'', z_2''), \eta_2'(z) < 0, \forall z \in (z_1'', \hat{z})$, and $\eta'(z) > 0, \forall z \in (\hat{z}, z_2'')$, and

$$\int_{\phi_1'}^{\phi_2'} g_\phi(T_0)\,\eta_1(z_\phi(T_0))\,f(\phi)\,\mathrm{d}\phi = -\int_{\phi_1''}^{\phi_2''} g_\phi(T_0)\,\eta_2(z_\phi(T_0))\,f(\phi)\,\mathrm{d}\phi. \tag{A.9}$$

Now consider the parameterized collection of tax policies

$$T_{\gamma,\varepsilon,\lambda} = T + \gamma\nu + \varepsilon(\lambda\eta_1 + \eta_2) \tag{A.10}$$

Next observe that by the Picard-Lindelöf theorem,[22] there exist $\delta^*, \varepsilon^* > 0$ for which there is a unique function $\zeta(\gamma, \varepsilon, \lambda) : [-\gamma^*, \gamma^*] \times [-\varepsilon^*, \varepsilon^*] \times [1-\delta^*, 1+\delta^*] \to \mathbb{R}$ to solve the equations

$$\zeta(0, \varepsilon, \lambda) = 0, \quad \forall \varepsilon, \forall \lambda \tag{A.11}$$

$$\begin{aligned} &\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T_{\gamma,\varepsilon,\lambda} - \zeta(\gamma, \varepsilon, \lambda)\mu) \\ &\times \left[-\frac{\partial}{\partial \gamma} \zeta(\gamma, \varepsilon, \lambda)\mu \left(z_\phi(T_{\gamma,\varepsilon,\lambda} - \zeta(\gamma, \varepsilon, \lambda)\mu) \right) + \nu\left(z_\phi(T_{\gamma,\varepsilon,\lambda} - \zeta(\gamma, \varepsilon, \lambda)\mu) \right) \right] f(\phi)\, \mathrm{d}\phi = 0. \end{aligned} \tag{A.12}$$

Note that because (A.12) reduces to (A.6) when $\varepsilon = 0$, the uniqueness of solutions implied by the Picard-Lindelöf theorem imply that

$$\zeta(\gamma) = \zeta(\gamma, 0, \lambda), \quad \forall \lambda \in [1-\delta^*, 1+\delta^*].$$

Moreover the smoothness of the primitive functions $T, \nu, \mu, \eta_1, \eta_2$ imply that $\zeta(\varepsilon, \gamma, \lambda)$ is twice continuously differentiable.[23]

Define

$$\hat{T}_{\gamma,\varepsilon,\lambda} = T_{\gamma,\varepsilon,\lambda} - \zeta(\gamma, \varepsilon, \lambda)\mu. \tag{A.13}$$

Now consider the optimization problem

$$\max_z z - \hat{T}_{\gamma,\varepsilon,\lambda} - v_\phi(z).$$

[22] Extending the construction in footnote 21, we can write

$$F_{\varepsilon,\lambda}(\gamma, \zeta) = \frac{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T_{\gamma,\varepsilon,\lambda} - \zeta\mu)\, \nu\left(z_\phi(T_{\gamma,\varepsilon,\lambda} - \zeta\mu)\right) f(\phi)\, \mathrm{d}\phi}{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T_{\gamma,\varepsilon,\lambda} - \zeta\mu)\, \mu\left(z_\phi(T_{\gamma,\varepsilon,\lambda} - \zeta\mu)\right) f(\phi)\, \mathrm{d}\phi}$$

If follows from the Picard-Lindelöf theorem that for each (ε, λ) in a neighborhood N of $(0,0)$, there exists $\gamma_{\varepsilon,\lambda} > 0$ such that the function $\zeta_{\varepsilon,\lambda}(\gamma) : [-\gamma_{\varepsilon,\lambda}, \gamma_{\varepsilon,\lambda}] \to \mathbb{R}$ satisfies $\frac{\mathrm{d}}{\mathrm{d}\gamma}\zeta_{\varepsilon,\lambda}(\gamma) = F(\gamma, \zeta_{\varepsilon,\lambda}(\gamma))$ for all $\gamma \in [-\gamma_{\varepsilon,\lambda}, \gamma_{\varepsilon,\lambda}]$. How large we can choose $\gamma_{\varepsilon,\lambda}$ depends on the range over which $F_{\varepsilon,\lambda}$ is Lipschitz continuous in γ for each value of ζ and the supremum of $F_{\varepsilon,\lambda}$ over the relevant range. Because $F_{\varepsilon,\lambda}$ varies smoothly in (ε, λ) we can choose $\gamma_{\varepsilon,\lambda}$ that is independent of ε and λ within the relevant range; that is, we can choose γ^* such that $\gamma^* = \gamma_{\varepsilon,\lambda}, \forall \varepsilon, \lambda \in [-\varepsilon^*, \varepsilon^*] \times [1-\delta^*, 1+\delta^*]$.

[23] In particular, the function ζ is implicitly defined by a differential equation of the form

$$\zeta' = H(\zeta; \varepsilon, \gamma, \lambda).$$

It follows from Corollary 4.1 on p. 101 of Hartman (1987) that to show that $\zeta(\varepsilon, \gamma, \lambda)$ is twice continuously differentiable, it is sufficient to show that H is twice continuously differentiable. Moreover, in our case H can be written in the form

$$H(\zeta; \varepsilon, \gamma, \lambda) = \frac{\int_{\underline{\phi}}^{\overline{\phi}} h_\phi(z_\phi(\varepsilon, \gamma, \lambda, \zeta); \varepsilon, \gamma, \lambda, \zeta)\, \mathrm{d}\phi}{\int_{\underline{\phi}}^{\overline{\phi}} k_\phi(z_\phi(\varepsilon, \gamma, \lambda, \zeta); \varepsilon, \gamma, \lambda, \zeta)\, \mathrm{d}\phi},$$

where $z_\phi(\varepsilon, \gamma, \lambda, \zeta)$ is the optimal solution to $\max_z z - T_{\gamma,\varepsilon,\lambda}(z) + \zeta\mu(z) - v_\phi(z)$, and h_ϕ and k_ϕ are twice continuously differentiable functions–where this follows from the smoothness of T, ν, μ, η_1, and η_2. Note next that $z_\phi(\varepsilon, \gamma, \lambda, \zeta)$ is defined as an implicit function by the first order condition to the above-mentioned optimization problem, which is of the form

$$0 = L(z; \varepsilon, \gamma, \lambda, \zeta),$$

where the smoothness of L, which again follows from the smoothness of T, ν, μ, η_1, and η_2 implies that $z_\phi(\varepsilon, \gamma, \lambda, \zeta)$ is twice continuously differentiable (see Theorem 2.1 on p. 364 of Lang (2012)) These facts together imply that H is twice continuously differentiable, which, in turn, implies that $\zeta(\varepsilon, \gamma, \lambda)$ is twice continuously differentiable as explained above.

The optimal solution $z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)$ satisfies the first order condition:

$$\begin{aligned}0 =& 1 - T'\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) - v'_\phi\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) - \gamma\nu'\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) \\ & - \varepsilon\left(\lambda\eta'_1\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + \eta'_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right) + \zeta(\gamma,\varepsilon,\lambda)\,\mu'\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\end{aligned} \tag{A.14}$$

Applying the implicit function theorem to (A.14) when $\gamma = \varepsilon = 0$, we derive

$$\left.\frac{\mathrm{d}}{\mathrm{d}\gamma}\right|_{\gamma=\varepsilon=0} z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) = \left.-\frac{\nu'\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + \left[\frac{\mathrm{d}}{\mathrm{d}\gamma}\zeta(\gamma,\varepsilon,\lambda)\right]\mu'\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}{T''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + v''_\phi\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}\right|_{\gamma=\varepsilon=0}, \tag{A.15}$$

$$\left.\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\right|_{\gamma=\varepsilon=0} z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) = \left.-\frac{\lambda\eta'_1\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + \eta'_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}{T''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + v''_\phi\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}\right|_{\gamma=\varepsilon=0}, \tag{A.16}$$

where (A.16) uses the fact that $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\zeta(\gamma,\varepsilon,\lambda)\big|_{\gamma=0} = 0$, which follows from (A.11). Observe that when $\phi \in (\phi''_1, \phi''_2)$, and $\gamma = \varepsilon = 0$, $\mu'\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) = 0$ and $\eta'_1\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) = 0$. So, in that case, (A.15)-(A.16) simplify to

$$\left.\frac{\mathrm{d}}{\mathrm{d}\gamma}\right|_{\gamma=\varepsilon=0} z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) = \left.-\frac{\nu'\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}{T''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + v''_\phi\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}\right|_{\gamma=\varepsilon=0}, \quad \forall\phi \in (\phi''_1, \phi''_2). \tag{A.17}$$

$$\left.\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\right|_{\gamma=\varepsilon=0} z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) = \left.-\frac{\eta'_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}{T''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + v''_\phi\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}\right|_{\gamma=\varepsilon=0}, \quad \forall\phi \in (\phi''_1, \phi''_2). \tag{A.18}$$

It follows from (A.17) and (A.18) that

$$\left.\frac{\mathrm{d}}{\mathrm{d}z}\eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\frac{\mathrm{d}}{\mathrm{d}\gamma}z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right|_{\gamma=\varepsilon=0} = \left.\frac{\mathrm{d}}{\mathrm{d}z}\nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right|_{\gamma=\varepsilon=0}, \quad \forall\phi \in (\phi''_1, \phi''_2). \tag{A.19}$$

Differentiating (A.12) with respect to ε, it follows that

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left[-\frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda)\,\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + \nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right] f(\phi)\,\mathrm{d}\phi = 0.$$

Applying Leibniz's integral rule and the product rule, we have

$$\begin{aligned}&\int_{\underline{\phi}}^{\overline{\phi}} \frac{\mathrm{d}}{\mathrm{d}\varepsilon} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left[-\frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda)\,\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + \nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right] f(\phi)\,\mathrm{d}\phi \\ &+ \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left[-\frac{\partial^2}{\partial\varepsilon\partial\gamma}\zeta(\gamma,\varepsilon,\lambda)\,\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right. \\ &\left. + \left[-\frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda)\,\frac{\mathrm{d}}{\mathrm{d}z}\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + \frac{\mathrm{d}}{\mathrm{d}z}\nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right]\frac{\mathrm{d}}{\mathrm{d}\varepsilon}z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right] f(\phi)\,\mathrm{d}\phi \\ &= 0.\end{aligned}$$

Rearranging terms,

$$\begin{aligned}
&\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \frac{\partial^2}{\partial\varepsilon\partial\gamma}\zeta\left(\gamma,\varepsilon,\lambda\right)\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) f\left(\phi\right)\mathrm{d}\phi \\
&= \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left[-\frac{\partial}{\partial\gamma}\zeta\left(\gamma,\varepsilon,\lambda\right)\frac{\mathrm{d}}{\mathrm{d}z}\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)+\frac{\mathrm{d}}{\mathrm{d}z}\nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right]\frac{\mathrm{d}}{\mathrm{d}\varepsilon}z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) f\left(\phi\right)\mathrm{d}\phi \qquad (A.20)\\
&\quad + \int_{\underline{\phi}}^{\overline{\phi}} \frac{\mathrm{d}}{\mathrm{d}\varepsilon}g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left[-\frac{\partial}{\partial\gamma}\zeta\left(\gamma,\varepsilon,\lambda\right)\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)+\nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right] f\left(\phi\right)\mathrm{d}\phi
\end{aligned}$$

Write $\hat{T}_{\gamma,\varepsilon,\lambda}\left(z\right)=\hat{T}\left(z;\gamma,\varepsilon,\lambda\right)$ and observe that (explanations for the derivation at the bottom):

$$\begin{aligned}
&\frac{\partial}{\partial\gamma}\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left[\frac{\partial}{\partial\varepsilon}\hat{T}\left(\tilde{z}_\phi;\gamma,\varepsilon,\lambda\right)\right]_{\tilde{z}_\phi=z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)} f\left(\phi\right)\mathrm{d}\phi\Bigg|_{\gamma=\varepsilon=0} \\
&= \frac{\partial}{\partial\gamma}\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left(\lambda\eta_1\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)+\eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)-\frac{\partial}{\partial\varepsilon}\zeta\left(\gamma,\varepsilon,\lambda\right)\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right) f\left(\phi\right)\mathrm{d}\phi\Bigg|_{\gamma=\varepsilon=0} \qquad (A.21)\\
&= \int_{\underline{\phi}}^{\overline{\phi}} \frac{\partial}{\partial\gamma}g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left(\lambda\eta_1\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)+\eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)-\frac{\partial}{\partial\varepsilon}\zeta\left(\gamma,\varepsilon,\lambda\right)\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right) f\left(\phi\right)\mathrm{d}\phi\Bigg|_{\gamma=\varepsilon=0} \\
&\quad + \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left(\left[\lambda\frac{\mathrm{d}}{\mathrm{d}z}\eta_1\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)+\frac{\mathrm{d}}{\mathrm{d}z}\eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)-\frac{\partial}{\partial\varepsilon}\zeta\left(\gamma,\varepsilon,\lambda\right)\frac{\mathrm{d}}{\mathrm{d}z}\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right]\right. \\
&\quad \left.\times\frac{\mathrm{d}}{\mathrm{d}\gamma}z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)-\frac{\partial^2}{\partial\varepsilon\partial\gamma}\zeta\left(\gamma,\varepsilon,\lambda\right)\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right) f\left(\phi\right)\mathrm{d}\phi\Bigg|_{\gamma=\varepsilon=0} \qquad (A.22)\\
&= \int_{\underline{\phi}}^{\overline{\phi}} \frac{\partial}{\partial\gamma}g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left(\eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right) f\left(\phi\right)\mathrm{d}\phi\Bigg|_{\gamma=\varepsilon=0} \qquad (A.23)\\
&\quad + \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left(\frac{\mathrm{d}}{\mathrm{d}z}\eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\frac{\mathrm{d}}{\mathrm{d}\gamma}z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)-\frac{\partial^2}{\partial\varepsilon\partial\gamma}\zeta\left(\gamma,\varepsilon,\lambda\right)\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right) f\left(\phi\right)d\phi\Bigg|_{\gamma=\varepsilon=0} \\
&= \int_{\underline{\phi}}^{\overline{\phi}} \frac{\partial}{\partial\gamma}g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) f\left(\phi\right)\mathrm{d}\phi\Bigg|_{\gamma=\varepsilon=0} \\
&\quad + \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left(\frac{\mathrm{d}}{\mathrm{d}z}\eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\frac{\mathrm{d}}{\mathrm{d}\gamma}z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) f\left(\phi\right)\mathrm{d}\phi\Bigg|_{\gamma=\varepsilon=0} \\
&\quad - \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\frac{\mathrm{d}}{\mathrm{d}z}\nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) f\left(\phi\right)\mathrm{d}\phi\Bigg|_{\gamma=\varepsilon=0} \qquad (A.24)\\
&\quad - \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\frac{\partial}{\partial\gamma}\zeta\left(\gamma,\varepsilon,\lambda\right)\frac{\mathrm{d}}{\mathrm{d}z}\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) f\left(\phi\right)\mathrm{d}\phi\Bigg|_{\gamma=\varepsilon=0} \\
&\quad - \int_{\underline{\phi}}^{\overline{\phi}} \frac{\mathrm{d}}{\mathrm{d}\varepsilon}g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left[-\frac{\partial}{\partial\gamma}\zeta\left(\gamma,\varepsilon,\lambda\right)\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)+\nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right] f\left(\phi\right)\mathrm{d}\phi\Bigg|_{\gamma=\varepsilon=0}
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\underline{\phi}}^{\overline{\phi}} \frac{\partial}{\partial\gamma} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \\
&\quad + \int_{\phi_1''}^{\phi_2''} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \left(\frac{\mathrm{d}}{\mathrm{d}z}\eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) \frac{\mathrm{d}}{\mathrm{d}\gamma} z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \\
&\quad - \int_{\phi_1''}^{\phi_2''} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \frac{\mathrm{d}}{\mathrm{d}z}\nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) \frac{\mathrm{d}}{\mathrm{d}\varepsilon} z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \qquad \text{(A.25)} \\
&\quad - \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda) \frac{\mathrm{d}}{\mathrm{d}z}\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) \frac{\mathrm{d}}{\mathrm{d}\varepsilon} z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \\
&\quad - \int_{\underline{\phi}}^{\overline{\phi}} \frac{\mathrm{d}}{\mathrm{d}\varepsilon} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \left[-\frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda)\,\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + \nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right] f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \\
&= \int_{\underline{\phi}}^{\overline{\phi}} \frac{\partial}{\partial\gamma} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \\
&\quad - \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda) \frac{\mathrm{d}}{\mathrm{d}z}\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) \frac{\mathrm{d}}{\mathrm{d}\varepsilon} z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \qquad \text{(A.26)} \\
&\quad - \int_{\underline{\phi}}^{\overline{\phi}} \frac{\mathrm{d}}{\mathrm{d}\varepsilon} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \left[-\frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda)\,\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + \nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right] f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \\
&= \int_{\underline{\phi}}^{\overline{\phi}} \frac{\partial}{\partial\gamma} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \qquad \text{(A.27)} \\
&\quad - \int_{\underline{\phi}}^{\overline{\phi}} \frac{\mathrm{d}}{\mathrm{d}\varepsilon} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \left[-\frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda) + \nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right] f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \\
&= \int_{\phi_1''}^{\phi_2''} \frac{\partial}{\partial\gamma} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \qquad \text{(A.28)} \\
&\quad - \int_{\phi_1''}^{\phi_2''} \frac{\mathrm{d}}{\mathrm{d}\varepsilon} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \left[-\frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda) + \nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right] f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \\
&= \int_{\phi_1''}^{\phi_2''} \left\{\frac{\partial}{\partial u}\hat{g}_\phi\left(U_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right), z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) \left[\frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda) - \nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right]\right. \\
&\quad \left. + \frac{\partial}{\partial z}\hat{g}_\phi\left(U_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right), z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) \frac{\mathrm{d}}{\mathrm{d}\gamma} z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right\} \times \eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \qquad \text{(A.29)} \\
&\quad - \int_{\phi_1''}^{\phi_2''} \left\{\frac{\partial}{\partial u}\hat{g}_\phi\left(U_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right), z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) \left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\zeta(\gamma,\varepsilon,\lambda) - \eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right]\right. \\
&\quad \left. + \frac{\partial}{\partial z}\hat{g}_\phi\left(U_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right), z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) \frac{\mathrm{d}}{\mathrm{d}\varepsilon} z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right\} \times \left[-\frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda) + \nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right] f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \\
&= \int_{\phi_1''}^{\phi_2''} + \frac{\partial}{\partial z}\hat{g}_\phi\left(U_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right), z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) \frac{\mathrm{d}}{\mathrm{d}\gamma} z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \times \eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \\
&\quad - \int_{\phi_1''}^{\phi_2''} \frac{\partial}{\partial z}\hat{g}_\phi\left(U_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right), z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) \frac{\mathrm{d}}{\mathrm{d}\varepsilon} z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right) \left[-\frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda) + \nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right] f(\phi)\,\mathrm{d}\phi \Big|_{\gamma=\varepsilon=0} \qquad \text{(A.30)}
\end{aligned}
$$

$$= -\int_{\phi_1''}^{\phi_2''} \frac{\partial}{\partial z}\hat{g}_\phi\left(U_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right), z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) \frac{\nu'\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}{T''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + v_\phi''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)} \eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) f(\phi)\,\mathrm{d}\phi\Big|_{\gamma=\varepsilon=0}$$
$$+ \int_{\phi_1''}^{\phi_2''} \frac{\partial}{\partial z}\hat{g}_\phi\left(U_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right), z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) \frac{\eta_2'\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}{T''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + v_\phi''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}$$
$$\times \left[-\frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda) + \nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right] f(\phi)\,\mathrm{d}\phi\Big|_{\gamma=\varepsilon=0} \tag{A.31}$$

$$= -\int_{\phi_1''}^{\phi_2''} \underbrace{\frac{\partial}{\partial z}\hat{g}_\phi\left(U_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right), z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}_{-} \underbrace{\frac{\nu'\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}{T''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + v_\phi''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}}_{+} \underbrace{\eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}_{-} \underbrace{f(\phi)}_{+}\mathrm{d}\phi\Big|_{\gamma=\varepsilon=0}$$
$$+ \int_{\phi_1''}^{\hat{\phi}} \underbrace{\frac{\partial}{\partial z}\hat{g}_\phi\left(U_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right), z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}_{-} \underbrace{\frac{\eta_2'\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}{T''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + v_\phi''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}}_{-}$$
$$\times \underbrace{\left[-\frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda) + \nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right]}_{-} \underbrace{f(\phi)}_{+}\mathrm{d}\phi\Big|_{\gamma=\varepsilon=0}$$
$$+ \int_{\hat{\phi}}^{\phi_2''} \underbrace{\frac{\partial}{\partial z}\hat{g}_\phi\left(U_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right), z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}_{-} \underbrace{\frac{\eta_2'\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}{T''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) + v_\phi''\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)}}_{+}$$
$$\times \underbrace{\left[-\frac{\partial}{\partial\gamma}\zeta(\gamma,\varepsilon,\lambda) + \nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\right]}_{+} \underbrace{f(\phi)}_{+}\mathrm{d}\phi\Big|_{\gamma=\varepsilon=0} \tag{A.32}$$

$$< 0 \tag{A.33}$$

where (A.21) follows from (A.10) and (A.13); (A.22) follows from the Leibniz integral rule, the product rule, and the symmetry of partial derivatives of ζ, which follows from the twice continuous differentiability of ζ, which was established above; (A.23) follows from the fact that by (A.11), $\frac{\partial}{\partial\varepsilon}\zeta(\gamma,\varepsilon,\lambda) = 0$ when $\gamma = 0$, the fact that $\frac{\partial}{\partial\gamma}g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\Big|_{\gamma=\varepsilon=0} = 0$ outside of $\left[\hat{\phi}_1, \overline{\phi}\right]$ because neither marginal taxes nor total taxes are changing locally outside of that interval as γ varies, and the fact that $\frac{\mathrm{d}}{\mathrm{d}\gamma}z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\Big|_{\gamma=\varepsilon=0} \neq 0$ only when $\phi \in [\phi_1, \phi_2]$, $\eta_1\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\Big|_{\gamma=\varepsilon=0} \neq 0$ only when $\phi \in [\phi_1', \phi_2']$, and $[\phi_1, \phi_2] \cap [\phi_1', \phi_2'] = \emptyset$; (A.24) follows from (A.20); (A.26) follows from the fact that $\frac{\mathrm{d}}{\mathrm{d}z}\eta_2\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\Big|_{\gamma=\varepsilon=0}$ and $\frac{\mathrm{d}}{\mathrm{d}z}\nu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\Big|_{\gamma=\varepsilon=0}$ equal zero outside of $[\phi_1'', \phi_2'']$; (A.26) follows from (A.19); (A.27) from the fact that $\frac{\mathrm{d}}{\mathrm{d}z}\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right)\Big|_{\gamma=\varepsilon=0} \neq 0$ only when $\phi \in \left(\hat{\phi}_1, \hat{\phi}_2\right)$, by (A.16), $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\Big|_{\gamma=\varepsilon=0} \neq 0$ only when $\phi \in (\phi_1', \phi_2') \cup (\phi_1'', \phi_2'')$ and $\left(\hat{\phi}_1, \hat{\phi}_2\right)$ and $(\phi_1', \phi_2') \cup (\phi_1'', \phi_2'')$ are disjoint; (A.28) follows from the fact that $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\Big|_{\gamma=\varepsilon=0} = 0$ outside of $(\phi_1', \phi_2') \cup (\phi_1'', \phi_2'')$, $\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) = 0$

on (ϕ_1', ϕ_2') and $\mu\left(z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\right) = 1$ on (ϕ_1'', ϕ_2''). (A.29) follows from expanding the terms $\frac{\partial}{\partial\gamma} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)$ and $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)$ using (A.2), and appealing to the envelope theorem; (A.30) follows from the fact that $\frac{\partial}{\partial\varepsilon}\zeta(\gamma,\varepsilon,\lambda) = 0$ when $\gamma = 0$ by (A.11), and the fact that once this term is eliminated, other terms cancel out; (A.31) follows from (A.17)-(A.18); where in (A.32), I have signed terms on the basis of assumptions that I have made above, including (A.3), assumptions above on ν and η_2, the convexity of T and v_ϕ, the fact that, by (34), $z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)$ is increasing in ϕ, and (A.8); and finally the final inequality (A.33) follows from keeping track of the preceding signs.

To summarize, above we have established that for all $\lambda \in [1-\delta^*, 1+\delta^*]$

$$\frac{\partial}{\partial\gamma}\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left[\frac{\partial}{\partial\varepsilon}\hat{T}\left(\tilde{z}_\phi;\gamma,\varepsilon,\lambda\right)\right]_{\tilde{z}_\phi = z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)} f(\phi)\,\mathrm{d}\phi\Bigg|_{\gamma=\varepsilon=0} < 0. \tag{A.34}$$

Next observe that it follows from (A.9) and our assumptions on η_1 and η_2 that

$$\left[\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left[\frac{\partial}{\partial\varepsilon}\hat{T}\left(\tilde{z}_\phi;\gamma,\varepsilon,\lambda\right)\right]_{\tilde{z}_\phi = z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)} f(\phi)\,\mathrm{d}\phi\Bigg|_{\gamma=\varepsilon=0,\lambda=1} \left\{\begin{array}{c} > \\ = \\ < \end{array}\right\} 0\right] \Leftrightarrow \lambda \left\{\begin{array}{c} > \\ = \\ < \end{array}\right\} 1 \tag{A.35}$$

Let us now assume for contradiction that there exists a preorder $\precsim$ that strongly rationalizes g. It follows from the local improvement principle (37) that there exists $\bar{\varepsilon} > 0$ such that for all $\varepsilon \in [0, \bar{\varepsilon}]$,

$$\lambda > 1 \Rightarrow T \succ T + \varepsilon\left(\lambda\eta_1 + \eta_2\right), \tag{A.36}$$

$$\lambda < 1 \Rightarrow T \prec T + \varepsilon\left(\lambda\eta_1 + \eta_2\right). \tag{A.37}$$

It follows from (A.34) and (A.35) that if $\gamma > 0$ is sufficiently small,

$$\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left[\frac{\partial}{\partial\varepsilon}\hat{T}\left(\tilde{z}_\phi;\gamma,\varepsilon,\lambda\right)\right]_{\tilde{z}_\phi = z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)} f(\phi)\,\mathrm{d}\phi\Bigg|_{\varepsilon=0,\lambda=1} < 0.$$

By continuity of the above integral in λ, for fixed small $\gamma > 0$, there exists δ_γ such that for all $\lambda \in [1-\delta_\gamma, 1+\delta_\gamma]$,

$$\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)\left[\frac{\partial}{\partial\varepsilon}\hat{T}\left(\tilde{z}_\phi;\gamma,\varepsilon,\lambda\right)\right]_{\tilde{z}_\phi = z_\phi\left(\hat{T}_{\gamma,\varepsilon,\lambda}\right)} f(\phi)\,\mathrm{d}\phi\Bigg|_{\varepsilon=0} < 0. \tag{A.38}$$

It follows from the local improvement principle (37) that for sufficiently small $\gamma > 0$ and $\lambda \in [1-\delta_\gamma, 1+\delta_\gamma]$, there exists $\varepsilon_{\gamma,\lambda} \in (0, \bar{\varepsilon})$, such that

$$T + \gamma\nu - \zeta(\gamma, 0, \lambda)\,\mu \prec T + \gamma\nu + \varepsilon_{\gamma,\lambda}\left(\lambda\eta_1 + \eta_2\right) - \zeta\left(\gamma, \varepsilon_{\gamma,\lambda}, \lambda\right)\mu, \tag{A.39}$$

Next observe that by (A.11)-(A.12) and the indifference principle (38),

$$T \sim T + \gamma\nu - \zeta(\gamma, 0, \lambda)\,\mu \tag{A.40}$$

$$T + \varepsilon_{\gamma,\lambda}\left(\lambda\eta_1 + \eta_2\right) \sim T + \gamma\nu + \varepsilon_{\gamma,\lambda}\left(\lambda\eta_1 + \eta_2\right) - \zeta\left(\gamma, \varepsilon_{\gamma,\lambda}, \lambda\right)\mu, \tag{A.41}$$

If we choose $\lambda \in (1, 1+\delta_\gamma)$, then putting together (A.60), (A.39), (A.64), and (A.36), we have

$$T \sim T + \gamma\nu - \zeta(\gamma, 0, \lambda)\,\mu \prec T + \gamma\nu + \varepsilon_{\gamma,\lambda}\left(\lambda\eta_1 + \eta_2\right) - \zeta\left(\gamma, \varepsilon_{\gamma,\lambda}, \lambda\right)\mu \sim T + \varepsilon_{\gamma,\lambda}\left(\lambda\eta_1 + \eta_2\right) \prec T. \tag{A.42}$$

Since $\precsim$ is assumed to be transitive, it follows that $T \prec T$, a contradiction. So g is not strongly rationalizable. This completes the proof of Proposition A.5 in the case that T satisfies (A.3). The proof is similar in the case that T instead satisfied (A.4). $\square$

A.3 Holding revenue constant

In this section I strengthen Proposition A.5; in particular, I prove the following.

Proposition A.3 *Under the assumptions of section 7, if g does not depend only on utility then g is not rationalizable.*

The difference between Propositions A.5 and A.3 is that the latter appeals to rationalizability rather than strong rationalizability. It is harder to show that welfare weights are not rationalizable than that they are not strongly rationalizable. In particular we have to show that the local improvement and indifference principles, even when restricted to paths along which *revenue is held constant*, imply the existence of a cycle. The key difference then is that our constructions must hold revenue constant, a feature that we did not worry about in Section A.2.

Now assume that g does not only depend on utility. So in this proof we are entitled to all of the constructions and conditions that were derived on the basis of this assumption in the course of the proof of Proposition A.5. In particular, note that I consider the case in which (A.3) is assumed to hold, but the proof for the case in which (A.4) is assumed to hold is similar. In this spirit, let T be the same as the tax policy T used in the proof of Proposition A.5. Let ϕ_1' and z_1' also be as in the proof or Proposition A.5, and observe by our assumptions on $\hat{\mathscr{T}}$, $z_{\underline{\phi}} = z_{\underline{\phi}}(T)$, where $z_{\underline{\phi}}$ is defined by (35). Then $z_{\underline{\phi}} < z_1'$. Let Ψ be the set of all three times continuously differentiable real-valued functions functions ψ on Z whose support is a proper subset of $\left(z_{\underline{\phi}}, z_1'\right)$ such that $\psi \not\equiv 0$, and $\psi(z) \geq 0, \forall z \in \left(z_{\underline{\phi}}, z_1'\right)$.

Choose $\omega, \psi \in \Psi$ with disjoint support. It follows from the Picard-Lindelöf theorem that there exists $\bar{\sigma} > 0$ such that there exists a function $\chi : [-\bar{\sigma}, \bar{\sigma}] \to \mathbb{R}$ satisfying:

$$\chi(0) = 0 \tag{A.43}$$

$$\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(T + \sigma\omega - \chi(\sigma)\psi\right)\left[-\frac{\mathrm{d}}{\mathrm{d}\sigma}\chi(\sigma)\psi\left(z_\phi\left(T + \sigma\omega - \chi(\sigma)\psi\right)\right) + \omega\left(z_\phi\left(T + \sigma\omega - \chi(\sigma)\psi\right)\right)\right] f(\phi)\,\mathrm{d}\phi = 0. \tag{A.44}$$

For $\sigma \in [-\bar{\sigma}, \bar{\sigma}]$, define

$$\tilde{T}_\sigma = T + \sigma\omega - \chi(\sigma)\psi. \tag{A.45}$$

Observe that $\tilde{T}_0 = T$. Observe that $\tilde{T}_\sigma$ (as well as χ) depend on the choice of ω and ψ. When I want to express this dependence, I write $\tilde{T}_\sigma^{\omega,\psi}$ instead of . There are two possibilities

1. There exist $\omega, \psi \in \Psi$ with disjoint support such that $\frac{\mathrm{d}}{\mathrm{d}\sigma}\big|_{\sigma=0} R\left(\tilde{T}_\sigma^{\omega,\psi}\right) \neq 0$.
2. For all $\omega, \psi \in \Psi$ with disjoint support, $\frac{\mathrm{d}}{\mathrm{d}\sigma}\big|_{\sigma=0} R\left(\tilde{T}_\sigma^{\omega,\psi}\right) = 0$.

In this Appendix, I will assume that we are in the first case. However, in Appendix A.4, I will consider the second case and show that in that case, it is always possible to make a modification to the original tax policy T such that steps in the proof of Proposition A.5 still hold and such that for some $\omega, \psi \in \Psi$, $\frac{\mathrm{d}}{\mathrm{d}\sigma}\big|_{\sigma=0} R\left(\tilde{T}_\sigma^{\omega,\psi}\right) \neq 0$. Thus I will

now assume that $\frac{\mathrm{d}}{\mathrm{d}\sigma}\big|_{\sigma=0} R\left(\tilde{T}_\sigma\right) \neq 0$, but I will show in Appendix A.4 that I could have always chosen the original tax policy T so that this assumption (or a similar assumption) holds. I separate this part of the argument so as to make clearer the structure of the proof, and focus, in this section, on what I take to be the more important details.

Since $\sigma \mapsto R\left(\tilde{T}_\sigma\right)$ is continuously differentiable, it follows that there is some interval $[\underline{\sigma}, \overline{\sigma}]$ containing 0 on which $R\left(\tilde{T}_\sigma\right)$ is either increasing or decreasing in σ. We may assume that the interval is chosen so that $\left|R\left(\tilde{T}_0\right) - R\left(\tilde{T}_{\overline{\sigma}}\right)\right| = \left|R\left(\tilde{T}_0\right) - R\left(\tilde{T}_{\underline{\sigma}}\right)\right| =: \overline{r}$. It follows that for each value $r \in [-\overline{r}, \overline{r}]$, there exists a unique value $\sigma(r) \in [\underline{\sigma}, \overline{\sigma}]$ such that

$$R\left(\tilde{T}_0\right) - R\left(\tilde{T}_{\sigma(r)}\right) = r. \tag{A.46}$$

Then let $\tilde{T}_r^* := \tilde{T}_{\sigma(r)}$. Moreover because T is strictly convex on $\left[z_{\underline{\phi}}, \bar{z}\right]$–and moreover $\frac{\mathrm{d}^2}{\mathrm{d}z^2} T(z) < 0$ on the supports of ω and ψ–if $\overline{r}$ is chosen to be small enough, then $\tilde{T}_r^*$ is strictly convex on the same interval for all $r \in [-\overline{r}, \overline{r}]$. Let us assume that $\overline{r}$ is so chosen.

I now prove the following lemma

Lemma A.1 *For any $\rho_0 \in P_0$, define $r^{\rho_0} : [a,b] \to \mathbb{R}$ by*

$$r^{\rho_0}(\theta) = R\left(T^{\rho_0,\theta}\right) - R\left(T^{\rho_0,0}\right), \ \ \forall \theta \in [a,b]. \tag{A.47}$$

Let $r_0 \in (-\overline{r}, \overline{r})$, and consider $\rho_0 \in P_0$ such that such that $\forall \theta \in [a,b], T^{\rho_0,\theta} \in \hat{\mathscr{T}}$ is strictly convex on $\left[z_{\underline{\phi}}, \bar{z}\right]$ and

$$|r^{\rho_0}(\theta)| < \overline{r} - |r_0|, \ \ \forall \theta \in [a,b]. \tag{A.48}$$

and

$$\forall \theta \in [a,b], T^{\rho_0,\theta} = \tilde{T}^*_{r_0} \text{ on } [0, z_1'].$$

Then there exists $\rho \in P$ such that

$$T^{\rho,\theta}(z) = \begin{cases} \tilde{T}^*_{r_0 + r^{\rho_0}(\theta)}(z) & \text{if } z < z_1', \\ T^{\rho_0,\theta}(z) & \text{if } z \geq z_1'. \end{cases}$$

and

$$\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(T^{\rho,\theta}\right) \frac{\partial}{\partial \theta} T\left(z_\phi^\rho(\theta), \theta\right) f(\phi)\, \mathrm{d}\phi = \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(T^{\rho_0,\theta}\right) \frac{\partial}{\partial \theta} T\left(z_\phi^{\rho_0}(\theta), \theta\right) f(\phi)\, \mathrm{d}\phi, \forall \theta \in [a,b]. \tag{A.49}$$

I now prove the lemma. So assume that there is $\rho_0 \in P_0$ with the properties given in the assumptions of the lemma. Now consider a family of tax policies $\left(S^\theta\right)_{\theta \in [a,b]}$ such that

$$S^\theta(z) = \begin{cases} \tilde{T}^*_{r_0 + r^{\rho_0}(\theta)}(z) & \text{if } z < z_1', \\ T^{\rho_0,\theta}(z) & \text{if } z \geq z_1'. \end{cases} \tag{A.50}$$

Thus, we would like to establish that there exists $\rho \in P$ such that $T^{\rho,\theta} = S^\theta$, for all $\theta \in [a,b]$. Observe that because the supports of ω and ψ were assumed contained in $(0, z_1')$, there is some $z_0' < z_1'$ such that for all

$\theta \in [a,b]$, $T^{\rho_0,\theta} = \tilde{T}^*_{r(\theta)}$ on $[z_0', z_1']$.[24] It now follows from the twice continuous differentiability and strict convexity of $T^{\rho_0,\theta}$ and $\tilde{T}^*_{r(\theta)}$ that $z \mapsto S^\theta(z)$ is twice continuously differentiable and strictly convex on $\left[z_{\underline{\phi}}, \bar{z}\right]$ for all $\theta \in [a,b]$;[25] moreover, $(z,\theta) \mapsto S^\theta(z)$ is twice continuously differentiable. It follows that there exists $\rho \in P_0$ such that $T^{\rho,\theta} = S^\theta$ for all $\theta \in [a,b]$. We want to show that $\rho \in P$.

Note also that by construction and convexity of the relevant tax policies,

$$\forall \phi \in \left[\underline{\phi}, \overline{\phi}\right], \forall \theta \in [a,b], \quad z_\phi\left(S^\theta\right) = \begin{cases} z_\phi\left(\tilde{T}^*_{r(\theta)}\right) & \text{if } \phi < \phi_1' \\ z_\phi^{\rho_0}(\theta) & \text{if } \phi \geq \phi_1' \end{cases}. \tag{A.51}$$

Observe that

$$\begin{aligned}
R\left(T^{\rho_0,0}\right) - R\left(S^\theta\right) &= \int_{\underline{\phi}}^{\overline{\phi}} T^{\rho_0,0}\left(z_\phi^{\rho_0}(0)\right) f(\phi)\,\mathrm{d}\phi - \int_{\underline{\phi}}^{\overline{\phi}} S^\theta\left(z_\phi\left(S^\theta\right)\right) f(\phi)\,\mathrm{d}\phi \\
&= \left[\int_{\underline{\phi}}^{\phi_1'} T^{\rho_0,0}\left(z_\phi^{\rho_0}(0)\right) f(\phi)\,\mathrm{d}\phi - \int_{\underline{\phi}}^{\phi_1'} S^\theta\left(z_\phi\left(S^\theta\right)\right) f(\phi)\,\mathrm{d}\phi\right] \\
&\quad + \left[\int_{\phi_1'}^{\overline{\phi}} T^{\rho_0,0}\left(z_\phi^{\rho_0}(0)\right) f(\phi)\,\mathrm{d}\phi - \int_{\phi_1'}^{\overline{\phi}} S^\theta\left(z_\phi\left(S^\theta\right)\right) f(\phi)\,\mathrm{d}\phi\right] \\
&= \left[\int_{\underline{\phi}}^{\phi_1'} T^{\rho_0,0}\left(z_\phi^{\rho_0}(0)\right) f(\phi)\,\mathrm{d}\phi - \int_{\underline{\phi}}^{\phi_1'} \tilde{T}^*_{r_0+r^{\rho_0}(\theta)}\left(z_\phi\left(\tilde{T}^*_{r_0+r^{\rho_0}(\theta)}\right)\right) f(\phi)\,\mathrm{d}\phi\right] \\
&\quad + \left[\int_{\phi_1'}^{\overline{\phi}} T^{\rho_0,0}\left(z_\phi^{\rho_0}(0)\right) f(\phi)\,\mathrm{d}\phi - \int_{\phi_1'}^{\overline{\phi}} T^{\rho_0,\theta}\left(z_\phi^{\rho_0}(\theta)\right) f(\phi)\,\mathrm{d}\phi\right] \\
&= \left[\int_{\underline{\phi}}^{\overline{\phi}} \tilde{T}^*_{r_0}\left(z_\phi\left(\tilde{T}^*_{r_0}\right)\right) f(\phi)\,\mathrm{d}\phi - \int_{\underline{\phi}}^{\overline{\phi}} \tilde{T}^*_{r_0+r^{\rho_0}(\theta)}\left(z_\phi\left(\tilde{T}^*_{r_0+r^{\rho_0}(\theta)}\right)\right) f(\phi)\,\mathrm{d}\phi\right] \\
&\quad + \left[\int_{\underline{\phi}}^{\overline{\phi}} T^{\rho_0,0}\left(z_\phi^{\rho_0}(0)\right) f(\phi)\,\mathrm{d}\phi - \int_{\underline{\phi}}^{\overline{\phi}} T^{\rho_0,\theta}\left(z_\phi^{\rho_0}(\theta)\right) f(\phi)\,\mathrm{d}\phi\right] \\
&= \left[R\left(\tilde{T}^*_{r_0}\right) - R\left(\tilde{T}^*_{r_0+r^{\rho_0}(\theta)}\right)\right] + \left[R\left(T^{\rho_0,0}\right) - R\left(T^{\rho_0,\theta}\right)\right] \\
&= \left(\left[R\left(\tilde{T}_0\right) - R\left(\tilde{T}^*_{r_0+r^{\rho_0}(\theta)}\right)\right] - \left[R\left(\tilde{T}_0\right) - R\left(\tilde{T}^*_{r_0}\right)\right]\right) + \left[R\left(T^{\rho_0,0}\right) - R\left(T^{\rho_0,\theta}\right)\right] \\
&= \left(r_0 + r^{\rho_0(\theta)}\right) - r_0 + \left[R\left(T^{\rho_0,0}\right) - R\left(T^{\rho_0,\theta}\right)\right] \\
&= r^{\rho_0}(\theta) + \left[R\left(T^{\rho_0,0}\right) - R\left(T^{\rho_0,\theta}\right)\right] \\
&= \left[R\left(T^{\rho_0,\theta}\right) - R\left(T^{\rho_0,0}\right)\right] + \left[R\left(T^{\rho_0,0}\right) - R\left(T^{\rho_0,\theta}\right)\right] \\
&= 0,
\end{aligned}$$

where the third equality follows from (A.50) and (A.51), the fourth from the fact that $T^{\rho_0,0}\left(z_\phi^{\rho_0}(0)\right) = T^{\rho_0,\theta}\left(z_\phi^{\rho_0}(\theta)\right) = \tilde{T}^*_{r_0}\left(z_\phi\left(\tilde{T}^*_{r_0}\right)\right)$ when $\phi \in \left[\phi_1', \overline{\phi}\right]$ and $\tilde{T}^*_{r_0}\left(z_\phi\left(\tilde{T}^*_{r_0}\right)\right) = \tilde{T}^*_{r_0+r^{\rho_0}(\theta)}\left(z_\phi\left(\tilde{T}^*_{r_0+r^{\rho_0}(\theta)}\right)\right) = T\left(z_\phi(T)\right)$ when $\phi \in \left[\underline{\phi}, \phi_1'\right)$, the seventh from (A.46), and the ninth from (A.47). So I have now established that

$$R\left(T^{\rho_0,0}\right) = R\left(S^\theta\right), \forall \theta \in [a,b].$$

It follows from this and the other properties of S^θ established above that $\rho : \theta \mapsto S^\theta$ belongs to P. Thus we can

[24] The existence of such a z_0' follows from the fact that the support of a function is a closed set.

[25] This follows from the fact that we have assumed $\overline{r}$ is sufficiently small that $\tilde{T}^*_r$ is strictly convex on $\left[z_{\underline{\phi}}, \bar{z}\right]$ when $r \in [-\overline{r}, \overline{r}]$, and that, by (A.48), for all $\theta \in [a,b]$, $|r_0 + r^{\rho_0}(\theta)| \leq |r_0| + |r^{\rho_0}(\theta)| \leq \overline{r}$.

take $T^{\rho,\theta} = S^{\theta}$.

Let me write $S(z,\theta) = S^{\theta}(z)$ and $T(z,\theta) = T^{\theta}(z)$. Then we have:

$$
\begin{aligned}
&\int_{\underline{\phi}}^{\overline{\phi}} g_{\phi}\left(S^{\theta}\right) \frac{\partial}{\partial\theta} S\left(z_{\phi}\left(S^{\theta}\right),\theta\right) f(\phi)\,\mathrm{d}\phi \\
&= \int_{\underline{\phi}}^{\phi_1'} g_{\phi}\left(S^{\theta}\right) \frac{\partial}{\partial\theta} S\left(z_{\phi}\left(S^{\theta}\right),\theta\right) f(\phi)\,\mathrm{d}\phi + \int_{\phi_1'}^{\overline{\phi}} g_{\phi}\left(S^{\theta}\right) \frac{\partial}{\partial\theta} S\left(z_{\phi}\left(S^{\theta}\right),\theta\right) f(\phi)\,\mathrm{d}\phi \\
&= \sigma'\left(r_0 + r^{\rho_0}(\theta)\right) \frac{\mathrm{d}}{\mathrm{d}\theta} r^{\rho_0}(\theta) \\
&\quad \times \int_{\underline{\phi}}^{\phi_1'} g_{\phi}\left(S^{\theta}\right) \left[\omega\left(z_{\phi}\left(\tilde{T}_{\sigma(r_0+r^{\rho_0}(\theta))}\right)\right) - \frac{\mathrm{d}}{\mathrm{d}\sigma}\chi\left(\sigma\left(r_0 + r^{\rho_0}(\theta)\right)\right) \psi\left(z_{\phi}\left(\tilde{T}_{\sigma(r_0+r^{\rho_0}(\theta))}\right)\right)\right] f(\phi)\,\mathrm{d}\phi \\
&\quad + \int_{\phi_1'}^{\overline{\phi}} g_{\phi}\left(T^{\rho_0,\theta}\right) \frac{\partial}{\partial\theta} T^{\rho_0}\left(z_{\phi}^{\rho_0}(\theta),\theta\right) f(\phi)\,\mathrm{d}\phi \qquad \text{(A.52)} \\
&= \sigma'\left(r_0 + r^{\rho_0}(\theta)\right) \frac{\mathrm{d}}{\mathrm{d}\theta} r^{\rho_0}(\theta) \\
&\quad \times \int_{\underline{\phi}}^{\overline{\phi}} g_{\phi}\left(S^{\theta}\right) \left[\omega\left(z_{\phi}\left(\tilde{T}_{\sigma(r_0+r^{\rho_0}(\theta))}\right)\right) - \frac{\mathrm{d}}{\mathrm{d}\sigma}\chi\left(\sigma\left(r_0 + r^{\rho_0}(\theta)\right)\right) \psi\left(z_{\phi}\left(\tilde{T}_{\sigma(r_0+r^{\rho_0}(\theta))}\right)\right)\right] f(\phi)\,\mathrm{d}\phi \\
&\quad + \int_{\underline{\phi}}^{\overline{\phi}} g_{\phi}\left(T^{\rho_0,\theta}\right) \frac{\partial}{\partial\theta} T^{\rho_0}\left(z_{\phi}^{\rho_0}(\theta),\theta\right) f(\phi)\,\mathrm{d}\phi \\
&\quad = \int_{\underline{\phi}}^{\overline{\phi}} g_{\phi}\left(T^{\rho_0,\theta}\right) \frac{\partial}{\partial\theta} T^{\rho_0}\left(z_{\phi}^{\rho_0}(\theta),\theta\right) f(\phi)\,\mathrm{d}\phi,
\end{aligned}
$$

where the second equality follows from (A.50), (A.51), and (A.45), the third from the fact that when $\phi \in \left[\phi_1', \overline{\phi}\right]$, $0 = \omega\left(z_{\phi}(T)\right) = \omega\left(z_{\phi}\left(\tilde{T}_{\sigma(r_0+r^{\rho_0}(\theta))}\right)\right)$ and $0 = \psi\left(z_{\phi}(T)\right) = \psi\left(z_{\phi}\left(\tilde{T}_{\sigma(r_0+r^{\rho_0}(\theta))}\right)\right)$ and when $\phi \in \left[\underline{\phi}, \phi_1'\right)$, $T\left(z_{\phi}^{\rho_0}(\theta),\theta\right) = \tilde{T}_r^*\left(z_{\phi}\left(\tilde{T}_r^*\right)\right)$, and the last equality by (A.44). (A.52) establishes (A.49) and completes the proof of Lemma A.1. □

Choose γ and λ for which the cycle (A.42) holds, and let $\gamma^* = \gamma, \lambda^* = \lambda, \varepsilon^* = \varepsilon_{\gamma,\lambda}$. Now consider $\rho_1, \rho_2, \rho_3, \rho_4 \in P_0$ such that

$$
\begin{aligned}
T^{\rho_1,\theta} &= \tilde{T}_{r_1}^* + \gamma^*\theta\nu - \zeta\left(\gamma^*\theta, 0, \lambda^*\right)\mu && \text{(A.53)} \\
T^{\rho_2,\theta} &= \tilde{T}_{r_2}^* + \varepsilon^*\theta\left(\lambda^*\eta_1 + \eta_2\right) && \text{(A.54)} \\
T^{\rho_3,\theta} &= \tilde{T}_{r_3}^* + \gamma^*\theta\nu + \varepsilon^*\left(\lambda^*\eta_1 + \eta_2\right) - \zeta\left(\gamma^*\theta, \varepsilon^*, \lambda^*\right)\mu && \text{(A.55)} \\
T^{\rho_4,\theta} &= \tilde{T}_{r_4}^* + \gamma^*\nu + \varepsilon^*\theta\left(\lambda^*\eta_1 + \eta_2\right) - \zeta\left(\gamma^*, \varepsilon^*\theta, \lambda^*\right)\mu, && \text{(A.56)}
\end{aligned}
$$

where

$$
\begin{aligned}
r_1 &= 0, \\
r_2 &= 0, \\
r_3 &= R\left(T + \varepsilon^*\left(\lambda^*\eta_1 + \eta_2\right)\right) - R(T), \\
r_4 &= R\left(T + \gamma^*\nu - \zeta\left(\gamma^*, 0, \lambda^*\right)\mu\right) - R(T),
\end{aligned} \qquad \text{(A.57)}
$$

If ε^* and γ^* are chosen small enough–which is consistent with (A.42)–the ρ_1, ρ_2, ρ_3, and ρ_4 indeed belong to P_0. It

follows from Lemma A.1 that for $i = 1, 2, 3, 4$, there exists $\hat{\rho}_i \in P$ such that

$$T^{\hat{\rho}_i,\theta}(z) = \begin{cases} \tilde{T}^*_{r_i + r^{\rho_i}(\theta)}(z) & \text{if } z < z_1', \\ T^{\hat{\rho}_i,\theta}(z) & \text{if } z \geq z_1'. \end{cases} \tag{A.58}$$

and

$$\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(T^{\hat{\rho}_i,\theta}\right) \frac{\partial}{\partial\theta} T\left(z_\phi^{\hat{\rho}_i}(\theta), \theta\right) f(\phi)\,\mathrm{d}\phi = \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(T^{\rho_i,\theta}\right) \frac{\partial}{\partial\theta} T\left(z_\phi^{\rho_i}(\theta), \theta\right) f(\phi)\,\mathrm{d}\phi, \forall\theta \in [a,b]. \tag{A.59}$$

I now assume that $\precsim$ is a relation that rationalizes g. It follows from (A.53), (A.12), the indifference principle (38), and (A.59) that

$$T^{\hat{\rho}_1,0} \sim T^{\hat{\rho}_1,1}. \tag{A.60}$$

Similarly, it follows from (A.54), (A.35), the fact that $\lambda^* > 1$, the local improvement principle (37), and (A.59) that

$$T^{\hat{\rho}_2,0} \succ T^{\hat{\rho}_2,1}. \tag{A.61}$$

Next observe that for all $\theta_0 \in [0,1]$,

$$\begin{aligned} &\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma^*\theta_0,\varepsilon^*,\lambda^*}\right)\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\bigg|_{\theta=\theta_0} \hat{T}_{\gamma^*\theta,\varepsilon^*,\lambda^*}\left(z_\phi\left(\hat{T}_{\gamma^*\theta_0,\varepsilon^*,\lambda^*}\right)\right)\right] f(\phi)\,\mathrm{d}\phi \\ &= \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(T^{\rho_3,\theta_0}\right)\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\bigg|_{\theta=\theta_0} T^{\rho_3,\theta}\left(z_\phi\left(T^{\rho_3,\theta_0}\right)\right)\right] f(\phi)\,\mathrm{d}\phi, \end{aligned} \tag{A.62}$$

$$\begin{aligned} &\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(\hat{T}_{\gamma^*,\varepsilon^*\theta_0,\lambda^*}\right)\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\bigg|_{\theta=\theta_0} \hat{T}_{\gamma^*,\varepsilon^*\theta,\lambda^*}\left(z_\phi\left(\hat{T}_{\gamma^*,\varepsilon^*\theta_0,\lambda^*}\right)\right)\right] f(\phi)\,\mathrm{d}\phi \\ &= \int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(T^{\rho_4,\theta_0}\right)\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\bigg|_{\theta=\theta_0} T^{\rho_4,\theta}\left(z_\phi\left(T^{\rho_4,\theta_0}\right)\right)\right] f(\phi)\,\mathrm{d}\phi. \end{aligned} \tag{A.63}$$

(A.62) follows because, in both integrals, the integrand is equal to zero when $\phi < \phi_1'$ and the integrands in the two integrals are equal elsewhere.[26] The reasoning justifying (A.63) is similar.

It follows from (A.55), (A.12), the indifference principle (38), (A.59), and (A.62) that

$$T^{\hat{\rho}_3,0} \sim T^{\hat{\rho}_3,1}. \tag{A.64}$$

Similarly, it follows from (A.56), (A.38), the local improvement principle (37), (A.59), and (A.63) that

$$T^{\hat{\rho}_4,0} \prec T^{\hat{\rho}_4,1}. \tag{A.65}$$

[26] So, in fact, the two integrands are equal everywhere.

It follows from (A.53)-(A.56) and (A.57) that

$$\begin{aligned}
r_1 + r^{\rho_1}(1) =& 0 + \left[R\left(T^{\rho_1,1}\right) - R\left(T^{\rho_1,0}\right)\right] = r_4 + r^{\rho_4(0)} \\
r_4 + r^{\rho_4}(1) =& \left[R\left(T + \gamma^*\nu - \zeta\left(\gamma^*, 0, \lambda^*\right)\mu\right) - R\left(T\right)\right] + \left[R\left(T^{\rho_4,1}\right) - R\left(T^{\rho_4,0}\right)\right] \\
=& \left[R\left(T + \gamma^*\nu - \zeta\left(\gamma^*, 0, \lambda^*\right)\mu\right) - R\left(T\right)\right] \\
&+ \left[R\left(T + \gamma^*\nu + \varepsilon^*\left(\lambda^*\eta_1 + \eta_2\right) - \zeta\left(\gamma^*, \varepsilon^*, \lambda^*\right)\mu\right) - R\left(T + \gamma^*\nu - \zeta\left(\gamma^*, 0, \lambda^*\right)\mu\right)\right] \\
=& \left[R\left(T + \varepsilon^*\left(\lambda^*\eta_1 + \eta_2\right) - \zeta\left(\gamma^*, \varepsilon^*, \lambda^*\right)\mu\right) - R\left(T\right)\right] \\
&+ \left[R\left(T + \gamma^*\nu + \varepsilon^*\left(\lambda^*\eta_1 + \eta_2\right) - \zeta\left(\gamma^*, \varepsilon^*, \lambda^*\right)\mu\right) - R\left(T + \varepsilon^*\left(\lambda^*\eta_1 + \eta_2\right) - \zeta\left(\gamma^*, \varepsilon^*, \lambda^*\right)\mu\right)\right] \\
=& \left[R\left(T + \varepsilon^*\left(\lambda^*\eta_1 + \eta_2\right) - \zeta\left(\gamma^*, \varepsilon^*, \lambda^*\right)\mu\right) - R\left(T\right)\right] + \left[R\left(T^{\rho_3,1}\right) - R\left(T^{\rho_3,0}\right)\right] \\
=& r_3 + r^{\rho_3}(1) \\
r_3 + r^{\rho_3}(0) =& \left[R\left(T + \varepsilon^*\left(\lambda^*\eta_1 + \eta_2\right)\right) - R\left(T\right)\right] + 0 = r_2 + r^{\rho_2}(1) \\
r_2 + r^{\rho_2}(0) =& 0 = r_1 + r^{\rho_2}(0)
\end{aligned} \tag{A.66}$$

One can use (A.66) to establish that

$$T^{\rho_1,1} = T^{\rho_4,0},\ \ T^{\rho_4,1} = T^{\rho_3,1},\ \ T^{\rho_3,0} = T^{\rho_2,1},\ \ T^{\rho_2,0} = T^{\rho_1,0}. \tag{A.67}$$

It follows from (A.66), (A.67) and (A.58) that

$$T^{\hat{\rho}_1,1} = T^{\hat{\rho}_4,0},\ \ T^{\hat{\rho}_4,1} = T^{\hat{\rho}_3,1},\ \ T^{\hat{\rho}_3,0} = T^{\hat{\rho}_2,1},\ \ T^{\hat{\rho}_2,0} = T^{\hat{\rho}_1,0}. \tag{A.68}$$

It follows from (A.60), (A.61), (A.64), (A.65) and (A.68) that

$$T^{\hat{\rho}_1,0} \sim T^{\hat{\rho}_1,1} = T^{\hat{\rho}_4,0} \prec T^{\hat{\rho}_4,1} = T^{\hat{\rho}_3,1} \sim T^{\hat{\rho}_3,0} = T^{\hat{\rho}_2,1} \prec T^{\hat{\rho}_2,0} = T^{\hat{\rho}_1,0} \tag{A.69}$$

It follows that any relation $\precsim$ that rationalizes g is not a preorder. So g is not rationalizable. This completes the proof. □

A.4 The case in which for all $\omega, \psi \in \Psi$ with disjoint support, $\frac{\mathrm{d}}{\mathrm{d}\sigma}\big|_{\sigma=0} R\left(\tilde{T}_\sigma^{\omega,\psi}\right) = 0$.

During the course of the proof of Proposition A.5, I assumed that there exist $\omega, \psi \in \Psi$ with disjoint support such that $\frac{\mathrm{d}}{\mathrm{d}\sigma}\big|_{\sigma=0} R\left(\tilde{T}_\sigma^{\omega,\psi}\right) \neq 0$. As discussed during the course of the proof of Proposition A.5, in this appendix, I show that if we happen to start with a tax T such that for all $\omega, \psi \in \Psi$, $\frac{\mathrm{d}}{\mathrm{d}\sigma}\big|_{\sigma=0} R\left(\tilde{T}_\sigma^{\omega,\psi}\right) = 0$, then we can construct a new tax policy T_0 which satisfies all the essential properties in Proposition A.5, but for which there exist ω, ψ in Ψ with disjoint support, such that $\frac{\mathrm{d}}{\mathrm{d}\sigma}\big|_{\sigma=0} R\left(\tilde{T}_\sigma^{\omega,\psi}\right) \neq 0$. Thus we can use T_0 instead of T to establish Propositions A.5 and A.3.

Let us rename the tax policy T in Propositions A.5 and A.3 as T^* to free up T to be used as a variable representing an arbitrary tax policy. Now define

$$\mathscr{T}^* = \left\{T \in \hat{\mathscr{T}} : T = T^* \text{ on } [z_1', \bar{z}] \text{ and } \frac{\mathrm{d}^2}{\mathrm{d}z^2} T < 0 \text{ on } \left(z_{\underline{\phi}}, z_1'\right]\right\}.$$

Note that any tax policy $T \in \mathscr{T}^*$ would have been sufficient for the proofs of Propositions A.5 and A.3, since the proofs did not depend on the precise values of T^* beneath z_1' did not matter for the proof.

For any $\psi \in \Psi$, let $\operatorname{supp}(\psi)$ be the support of ψ. Choose any $T \in \mathscr{T}^*$ and $\omega, \psi \in \Psi$. Then there exists a function $\chi = \chi_T^{\omega,\psi}$ for which (A.43)-(A.44) are satisfied when T^* is replaced by T. For any $T \in \mathscr{T}^*, \omega, \psi \in \Psi$ with $\operatorname{supp}(\omega) \cap \operatorname{supp}(\psi) = \emptyset$ and σ sufficiently close to 0 to be in the domain of $\chi_T^{\omega,\psi}$, define the tax policy

$$S_\sigma^{T,\omega,\psi} = T + \sigma\omega - \chi_T^{\omega,\psi}(\sigma)\,\psi.$$

To deal with the case covered by this section it is sufficient to prove the following proposition.

Proposition A.4 *There exists $T \in \mathscr{T}^*, \omega, \psi \in \Psi$ with $\operatorname{supp}(\omega) \cap \operatorname{supp}(\psi) = \emptyset$ such that $\frac{\mathrm{d}}{\mathrm{d}\sigma}\big|_{\sigma=0} R\left(S_\sigma^{T,\omega,\psi}\right) \neq 0$.*

This proposition is sufficient because, once it is established, we can simply substitute the tax policy T guaranteed by the proposition for the policy T^* that we originally used in Propositions A.5 and A.3 and the arguments in both those propositions go through for T.

Proof of Proposition A.4. Assume for contradiction that

$$\text{for all } T \in \mathscr{T}^*, \omega, \psi \in \Psi \text{ with } \operatorname{supp}(\omega) \cap \operatorname{supp}(\psi) = \emptyset,\ \frac{\mathrm{d}}{\mathrm{d}\sigma}\bigg|_{\sigma=0} R\left(S_\sigma^{T,\omega,\psi}\right) = 0. \tag{A.70}$$

It follows from applying the implicit function theorem to first order condition for the agent's optimization problem

$$\max_z z - S_\sigma^{T,\omega,\psi}(z) - v_\phi(z)$$

that

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\bigg|_{\sigma=0} z_\phi\left(S_\sigma^{T,\omega,\psi}\right) = -\frac{\omega'\left(z_\phi(T)\right) - \left[\frac{\mathrm{d}}{\mathrm{d}\sigma}\big|_{\sigma=0}\chi_T^{\omega,\psi}(\sigma)\right]\psi'\left(z_\phi(T)\right)}{T''\left(z_\phi(T)\right) + v_\phi''\left(z_\phi(T)\right)}. \tag{A.71}$$

Next observe that

$$R\left(S_\sigma^{T,\omega,\psi}\right) = \int_{\underline{\phi}}^{\overline{\phi}} \left[T_0\left(z_\phi\left(S_\sigma^{T,\omega,\psi}\right)\right) + \sigma\omega\left(z_\phi\left(S_\sigma^{T,\omega,\psi}\right)\right) - \chi_T^{\omega,\psi}(\sigma)\,\psi\left(z_\phi\left(S_\sigma^{T,\omega,\psi}\right)\right)\right] f(\phi)\,\mathrm{d}\phi.$$

So the marginal effect on revenue of changing σ is

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\sigma}\bigg|_{\sigma=0} R\left(S_\sigma^{T,\omega,\psi}\right) &= \int_{\underline{\phi}}^{\overline{\phi}} \left[T'\left(z_\phi(T)\right)\frac{\mathrm{d}}{\mathrm{d}\sigma}\bigg|_{\sigma=0} z_\phi\left(S_\sigma^{T,\omega,\psi}\right) + \omega\left(z_\phi(T)\right) - \left[\frac{\mathrm{d}}{\mathrm{d}\sigma}\bigg|_{\sigma=0}\chi_T^{\omega,\psi}(\sigma)\right]\psi\left(z_\phi\left(\tilde{T}_\sigma\right)\right)\right] f(\phi)\,\mathrm{d}\phi \\
&= \int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi(T)\right)\frac{\omega'\left(z_\phi(T)\right) - \left[\frac{\mathrm{d}}{\mathrm{d}\sigma}\big|_{\sigma=0}\chi_T^{\omega,\psi}(\sigma)\right]\psi'\left(z_\phi(T)\right)}{T''\left(z_\phi(T)\right) + v_\phi''\left(z_\phi(T)\right)} + \omega\left(z_\phi(T)\right)\right. \\
&\quad \left. - \left[\frac{\mathrm{d}}{\mathrm{d}\sigma}\bigg|_{\sigma=0}\chi_T^{\omega,\psi}(\sigma)\right]\psi\left(z_\phi(T)\right)\right] f(\phi)\,\mathrm{d}\phi \\
&= \int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi(T)\right)\frac{\omega'\left(z_\phi(T)\right)}{T''\left(z_\phi(T)\right) + v_\phi''\left(z_\phi(T)\right)} + \omega\left(z_\phi(T)\right)\right] f(\phi)\,\mathrm{d}\phi \\
&\quad - \left[\frac{\mathrm{d}}{\mathrm{d}\sigma}\bigg|_{\sigma=0}\chi_T^{\omega,\psi}(\sigma)\right]\int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi(T)\right)\frac{\psi'\left(z_\phi(T)\right)}{T''\left(z_\phi(T)\right) + v_\phi''\left(z_\phi(T)\right)} + \psi\left(z_\phi(T)\right)\right] f(\phi)\,\mathrm{d}\phi,
\end{aligned} \tag{A.72}$$

where the second equality follows from (A.71).

Lemma A.2 *Assume (A.70). Let $T \in \mathscr{T}^*$. Suppose that there exists $\omega \in \Psi$ such that*

$$\int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi\left(T\right)\right) \frac{\omega'\left(z_\phi\left(T\right)\right)}{T''\left(z_\phi\left(T\right)\right) + v''_\phi\left(z_\phi\left(T\right)\right)} + \omega\left(z_\phi\left(T\right)\right) \right] f\left(\phi\right) \mathrm{d}\phi = 0. \tag{A.73}$$

Then for all $\psi \in \Psi$,

$$\int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi\left(T\right)\right) \frac{\psi'\left(z_\phi\left(T\right)\right)}{T''\left(z_\phi\left(T\right)\right) + v''_\phi\left(z_\phi\left(T\right)\right)} + \psi\left(z_\phi\left(T\right)\right) \right] f\left(\phi\right) \mathrm{d}\phi = 0. \tag{A.74}$$

Proof. Choose $T \in \mathscr{T}^*$, and consider $\omega \in \Psi$ satisfying (A.73). Then for any $\psi \in \Psi$, there exists a sequence of functions $\psi_0, \psi_1, \ldots, \psi_n$ in Ψ such that $\psi_0 = \omega, \psi_n = \psi$ and $\mathrm{supp}\left(\psi_{j-1}\right) \cap \mathrm{supp}\left(\psi_j\right) = \emptyset$ for $j = 1, \ldots, n$.[27] It now follows from (A.70) and (A.72) that

$$\begin{aligned}
0 = \left.\frac{\mathrm{d}}{\mathrm{d}\sigma}\right|_{\sigma=0} R\left(S_\sigma^{T,\psi_1,\omega}\right) &= \int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi\left(T\right)\right) \frac{\psi_1'\left(z_\phi\left(T\right)\right)}{T''\left(z_\phi\left(T\right)\right) + v''_\phi\left(z_\phi\left(T\right)\right)} + \psi_1\left(z_\phi\left(T\right)\right) \right] f\left(\phi\right) \mathrm{d}\phi \\
&\quad - \left[\left.\frac{\mathrm{d}}{\mathrm{d}\sigma}\right|_{\sigma=0} \chi^{\psi_1,\omega}\left(\sigma\right) \right] \int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi\left(T\right)\right) \frac{\omega'\left(z_\phi\left(T\right)\right)}{T''\left(z_\phi\left(T\right)\right) + v''_\phi\left(z_\phi\left(T\right)\right)} + \omega\left(z_\phi\left(T\right)\right) \right] f\left(\phi\right) \mathrm{d}\phi \\
&= \int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi\left(T\right)\right) \frac{\psi_1'\left(z_\phi\left(T\right)\right)}{T''\left(z_\phi\left(T\right)\right) + v''_\phi\left(z_\phi\left(T\right)\right)} + \psi_1\left(z_\phi\left(T\right)\right) \right] f\left(\phi\right) \mathrm{d}\phi
\end{aligned}$$

So ψ_1 satisfies (A.74). A similar argument shows that if ψ_j satisfies (A.74), then so does ψ_{j+1}. □

Define

$$\begin{aligned}
\mathscr{T}_0^* &= \left\{ T \in \mathscr{T}^* : \forall \psi \in \Psi, \int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi\left(T\right)\right) \frac{\psi'\left(z_\phi\left(T\right)\right)}{T''\left(z_\phi\left(T\right)\right) + v''_\phi\left(z_\phi\left(T\right)\right)} + \psi\left(z_\phi\left(T\right)\right) \right] f\left(\phi\right) \mathrm{d}\phi = 0 \right\}, \\
\mathscr{T}_1^* &= \left\{ T \in \mathscr{T}^* : \forall \psi \in \Psi, \int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi\left(T\right)\right) \frac{\psi'\left(z_\phi\left(T\right)\right)}{T''\left(z_\phi\left(T\right)\right) + v''_\phi\left(z_\phi\left(T\right)\right)} + \psi\left(z_\phi\left(T\right)\right) \right] f\left(\phi\right) \mathrm{d}\phi \neq 0 \right\}, \\
\mathscr{T}_2^* &= \left\{ T \in \mathscr{T}^* : \exists \psi \in \Psi, \int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi\left(T\right)\right) \frac{\psi'\left(z_\phi\left(T\right)\right)}{T''\left(z_\phi\left(T\right)\right) + v''_\phi\left(z_\phi\left(T\right)\right)} + \psi\left(z_\phi\left(T\right)\right) \right] f\left(\phi\right) \mathrm{d}\phi \neq 0 \right\}.
\end{aligned}$$

Corollary A.2 *Assume (A.70). Then $\mathscr{T}_0^* \cup \mathscr{T}_1^* = \mathscr{T}^*$ and $\mathscr{T}_1^* = \mathscr{T}_2^*$.*

This corollary is an immediate consequence of Lemma A.2 and (A.70).

Lemma A.3 *Assume (A.70). Then for all $T \in \mathscr{T}_1^*$, there exists $c\left(T\right) \in \mathbb{R} \setminus \{0\}$, such that for all $\psi \in \Psi$,*

$$\int_{\underline{\phi}}^{\overline{\phi}} g_\phi\left(T\right) \psi\left(z_\phi\left(T\right)\right) f\left(\phi\right) \mathrm{d}\phi = c\left(T\right) \int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi\left(T\right)\right) \frac{\psi'\left(z_\phi\left(T\right)\right)}{T''\left(z_\phi\left(T\right)\right) + v''_\phi\left(z_\phi\left(T\right)\right)} + \psi\left(z_\phi\left(T\right)\right) \right] f\left(\phi\right) \mathrm{d}\phi \tag{A.75}$$

[27] There are two cases to consider: (i) $\mathrm{supp}\left(\omega\right) \cup \mathrm{supp}\left(\psi\right) = \left[z_{\underline{\phi}}, z_1'\right]$ and (ii) $\mathrm{supp}\left(\omega\right) \cup \mathrm{supp}\left(\psi\right) \neq \left[z_{\underline{\phi}}, z_1'\right]$. First consider case (i). Then because Ψ is defined so that the support of any function in Ψ is strictly contained in $\left[z_{\underline{\phi}}, z_1'\right]$, it follows that $\mathrm{supp}\left(\omega\right) \backslash \mathrm{supp}\left(\psi\right) \neq \emptyset$. So we can pick $\psi_0 = \omega$, ψ_1 so that $\mathrm{supp}\left(\psi_1\right) \cap \mathrm{supp}\left(\omega\right) = \emptyset$, ψ_2 so that $\mathrm{supp}\left(\psi_2\right) \subseteq \mathrm{supp}\left(\omega\right) \backslash \mathrm{supp}\left(\psi\right)$ and $\psi_3 = \psi$. Next consider case (ii). Then we can choose $\psi_0 = \omega$, ψ_2 so that $\mathrm{supp}\left(\psi_2\right) \cap \left(\mathrm{supp}\left(\omega\right) \cup \mathrm{supp}\left(\psi\right)\right) = \emptyset$, and $\psi_2 = \psi$. So we can find an appropriate sequence in both cases. Notice that in both cases, we can take $n \leq 3$.

Proof. If $T \in \mathscr{T}_1^*$, it follows from (A.70) and (A.72) that for all $\omega, \psi \in \Psi$ with disjoint support,

$$\left.\frac{\mathrm{d}}{\mathrm{d}\sigma}\right|_{\sigma=0} \chi_T^{\omega,\psi}(\sigma) = \frac{\int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi(T)\right) \frac{\omega'(z_\phi(T))}{T''(z_\phi(T))+v''_\phi(z_\phi(T))} + \omega\left(z_\phi(T)\right)\right] f(\phi)\,\mathrm{d}\phi}{\int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi(T)\right) \frac{\psi'(z_\phi(T))}{T''(z_\phi(T))+v''_\phi(z_\phi(T))} + \psi\left(z_\phi(T)\right)\right] f(\phi)\,\mathrm{d}\phi} \tag{A.76}$$

Next observe that in the case that $\sigma = 0$, (A.44) can equivalently be rewritten as:

$$\left.\frac{\mathrm{d}}{\mathrm{d}\sigma}\right|_{\sigma=0} \chi_T^{\omega,\psi}(\sigma) = \frac{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T)\,\omega\left(z_\phi(T)\right) f(\phi)\,\mathrm{d}\phi}{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T)\,\psi\left(z_\phi(T)\right) f(\phi)\,\mathrm{d}\phi}. \tag{A.77}$$

It follows from (A.76) and (A.77) that

$$\frac{\int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi(T)\right) \frac{\omega'(z_\phi(T))}{T''(z_\phi(T))+v''_\phi(z_\phi(T))} + \omega\left(z_\phi(T)\right)\right] f(\phi)\,\mathrm{d}\phi}{\int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi(T)\right) \frac{\psi'(z_\phi(T))}{T''(z_\phi(T))+v''_\phi(z_\phi(T))} + \psi\left(z_\phi(T)\right)\right] f(\phi)\,\mathrm{d}\phi} = \frac{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T)\,\omega\left(z_\phi(T)\right) f(\phi)\,\mathrm{d}\phi}{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T)\,\psi\left(z_\phi(T)\right) f(\phi)\,\mathrm{d}\phi},$$

or equivalently, for all $\omega, \psi \in \Psi$ with disjoint support,

$$\frac{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T)\,\omega\left(z_\phi(T)\right) f(\phi)\,\mathrm{d}\phi}{\int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi(T)\right) \frac{\omega'(z_\phi(T))}{T''(z_\phi(T))+v''_\phi(z_\phi(T))} + \omega\left(z_\phi(T)\right)\right] f(\phi)\,\mathrm{d}\phi} \tag{A.78}$$

$$= \frac{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T)\,\psi\left(z_\phi(T)\right) f(\phi)\,\mathrm{d}\phi}{\int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi(T)\right) \frac{\psi'(z_\phi(T))}{T''(z_\phi(T))+v''_\phi(z_\phi(T))} + \psi\left(z_\phi(T)\right)\right] f(\phi)\,\mathrm{d}\phi} \tag{A.79}$$

Now choose some specific $\omega \in \Psi$, and define

$$c(T) = \frac{\int_{\underline{\phi}}^{\overline{\phi}} g_\phi(T)\,\omega\left(z_\phi(T)\right) f(\phi)\,\mathrm{d}\phi}{\int_{\underline{\phi}}^{\overline{\phi}} \left[-T'\left(z_\phi(T)\right) \frac{\omega'(z_\phi(T))}{T''(z_\phi(T))+v''_\phi(z_\phi(T))} + \omega\left(z_\phi(T)\right)\right] f(\phi)\,\mathrm{d}\phi}. \tag{A.80}$$

Choose any other $\psi^* \in \Psi$. Then there exists a sequence $\psi_0, \psi_1, \ldots, \psi_n$ in Ψ such that $\psi_0 = \omega, \psi_n = \psi^*$ and ψ_{j-1} and ψ_j have disjoint support. (See footnote 27). Then setting $\psi_1 = \psi$, in (A.78), and using (A.80) with $\psi = \psi_1$, we derive (A.75) for ψ_1. Next assuming that (A.75) holds for ψ_j, we can derive it for ψ_{j+1} using a similar argument. □

Lemma A.4 *Assume (A.70). For all $T \in \mathscr{T}^*$ and $\phi \in \left[\underline{\phi}, \overline{\phi}\right]$, define*

$$a_T(\phi) = \left(\frac{g_\phi(T)}{c(T)} - 1\right) f(\phi),$$
$$b_T(\phi) = -\frac{T'\left(z_\phi(T)\right)}{\frac{\partial^2}{\partial\phi\partial z} v\left(z_\phi(T), \phi\right)} f(\phi)$$

For all $T \in \mathscr{T}_1^*$, $a_T = b_T'$ on $\left[\underline{\phi}, \phi_1'\right]$. For all $T \in \mathscr{T}_0^*$, $f = b_T'$.

Proof. Fix $T \in \mathscr{T}_1^*$. We can rewrite (A.75) as

$$\forall \psi \in \Psi,$$
$$\int_{\underline{\phi}}^{\overline{\phi}} \left(\frac{g_\phi(T)}{c(T)} - 1 \right) \psi(z_\phi(T)) f(\phi) \, \mathrm{d}\phi = \int_{\underline{\phi}}^{\overline{\phi}} -T'(z_\phi(T)) \frac{\psi'(z_\phi(T))}{T''(z_\phi(T)) + v''_\phi(z_\phi(T))} f(\phi) \, \mathrm{d}\phi \tag{A.81}$$

If for any $\psi \in \Psi$, we define $\psi_T(\phi) := \psi(z_\phi(T))$, then $\psi'_T(\phi) = \psi'(z_\phi(T)) \frac{\mathrm{d}}{\mathrm{d}\phi} z_\phi(T)$. Then observe also that

$$\begin{aligned} b_T(\phi) &= -\frac{T'(z_\phi(T))}{\frac{\partial^2}{\partial \phi \partial z} v(z_\phi(T), \phi)} f(\phi) \\ &= -\frac{T'(z_\phi(T))}{T''(z_\phi(T)) + v''_\phi(z_\phi(T))} \frac{T''(z_\phi(T)) + \frac{\partial^2}{\partial z^2} v(z_\phi(T), \phi)}{\frac{\partial^2}{\partial \phi \partial z} v(z_\phi(T), \phi)} f(\phi) \\ &= \frac{T'(z_\phi(T))}{T''(z_\phi(T)) + v''_\phi(z_\phi(T))} \frac{1}{\frac{\mathrm{d}}{\mathrm{d}\phi} z_\phi(T)} f(\phi), \end{aligned}$$

where the last equality follows from applying the implicit function theorem to the first-order condition for the agent's optimization problem. It now follows that we can rewrite (A.81) as

$$\forall \psi \in \Psi, \int_{\underline{\phi}}^{\phi'_1} a_T(\phi) \psi_T(\phi) \, \mathrm{d}\phi = -\int_{\underline{\phi}}^{\phi'_1} b_T(\phi) \psi'_T(\phi) \, \mathrm{d}\phi, \tag{A.82}$$

where the upper bounds of integration can be taken to be ϕ'_1 rather than $\overline{\phi}$ because all $\psi \in \Psi$ have support in $\left[\underline{\phi}, \phi'_1\right]$. The proof of Lemma A.4 for the case where $T \in \mathscr{T}_1^*$ is now completed by Lemma A.5 (below). The proof in the case that $T \in \mathscr{T}_0^*$. □

Lemma A.5 *(A.82) holds if and only if holds if and only if $a_T = b'_T$.*

Proof. First suppose that $a_T = b'_T$, and let B_T be an antiderivative of b_T. Then using integration by parts, we have that for any $\psi \in \Psi$,

$$\begin{aligned} \int_{\underline{\phi}}^{\phi'_1} b_T(\phi) \psi'_T(\phi) \, \mathrm{d}\phi &= \left[B_T(\phi) \psi_T(\phi) \right]_{\underline{\phi}}^{\phi'_1} - \int_{\underline{\phi}}^{\phi'_1} a_T(\phi) \psi_T(\phi) \, \mathrm{d}\phi \\ &= -\int_{\underline{\phi}}^{\phi'_1} a_T(\phi) \psi_T(\phi) \, \mathrm{d}\phi, \end{aligned}$$

where the second equality follows from the fact that for any $\psi \in \Psi$, $\psi_T\left(\underline{\phi}\right) = \psi_T(\phi'_1) = 0$.

Going in the other direction, suppose that (A.82) holds. It also follows from integration by parts that

$$\forall \psi \in \Psi, \int_{\underline{\phi}}^{\phi'_1} b'_T(\phi) \psi_T(\phi) \, \mathrm{d}\phi = -\int_{\underline{\phi}}^{\phi'_1} b_T(\phi) \psi'_T(\phi) \, \mathrm{d}\phi. \tag{A.83}$$

It follows from (A.82) and (A.83) that

$$\forall \psi \in \Psi, \int_{\underline{\phi}}^{\phi'_1} \left(b'_T(\phi) - a_T(\phi) \right) \psi_T(\phi) \, \mathrm{d}\phi = 0. \tag{A.84}$$

Since both b'_T and a_T are continuously differentiable, if $b'_T \neq a_T$, then it would be possible to find $\psi \in \Psi$ whose support is contained in either $\left\{ z_\phi(T) : \phi \in \left[\underline{\phi}, \phi'_1\right], b'_T(\phi) > a_T(\phi) \right\}$ or $\left\{ z_\phi(T) : \phi \in \left[\underline{\phi}, \phi'_1\right], b'_T(\phi) < a_T(\phi) \right\}$ and in that case the integral in (A.84) would be nonzero, a contradiction. So we must have $b'_T = a_T$. □

The requirement that $a_T = b'_T$ when $T \in \mathscr{T}_1^*$ is the requirement that:

$$\left(\frac{g_\phi(T)}{c(T)} - 1\right) f(\phi) = -\frac{\mathrm{d}}{\mathrm{d}\phi}\left[\frac{T'(z_\phi(T))}{\frac{\partial^2}{\partial\phi\partial z} v(z_\phi(T), \phi)} f(\phi)\right], \forall \phi \in \left[\underline{\phi}, \phi_1'\right], \forall T \in \mathscr{T}_1^*.$$

or equivalently that

$$g_\phi(T) f(\phi) = c(T) h(T, \phi), \forall \phi \in \left[\underline{\phi}, \phi_1'\right], \forall T \in \mathscr{T}_1^* \tag{A.85}$$

where

$$h(T, \phi) := \left(-\frac{\mathrm{d}}{\mathrm{d}\phi}\left[\frac{T'(z_\phi(T))}{\frac{\partial^2}{\partial\phi\partial z} v(z_\phi(T), \phi)} f(\phi)\right] + f(\phi)\right). \tag{A.86}$$

Similarly the requirement that $-f = b'_T$ when $T \in \mathscr{T}^*$ can be written as

$$0 = h(T, \phi), \forall \phi \in \left[\underline{\phi}, \phi_1'\right], \forall T \in \mathscr{T}_0^*. \tag{A.87}$$

Lemma A.6 *Assume (A.70). Let T_0 and T_1 be two tax policies in $\mathscr{T}_1^*$ satisfying (A.85) such that there is some interval I containing both 0 and $z^* := z_{\phi^*}(T_0) > 0$ for some $\phi^* \in \left(\underline{\phi}, \phi_1^*\right)$, $T_0 = T_1$ on I. Then $c(T_0) = c(T_1)$.*

Proof. The assumptions of the lemma imply that $z_{\phi^*}(T_0) = z_{\phi^*}(T_1)$ and $U_{\phi^*}(T_0) = U_{\phi^*}(T_1)$. It follows that

$$\begin{aligned} c(T_0) h(T_0, \phi^*) &= g_{\phi^*}(U_{\phi^*}(T_0), z_{\phi^*}(T_0)) f(\phi^*) = g_{\phi^*}(U_{\phi^*}(T_1), z_{\phi^*}(T_1)) f(\phi^*) \\ &= c(T_1) h(T_1, \phi^*) = c(T_1) h(T_0, \phi^*). \end{aligned}$$

where the first and third equalities follow from (A.85) and the last equality follows from (A.86) and the fact that T_0 and T_1 agree on an interval containing z^*. It now follows from the fact that $g_{\phi^*}(U_{\phi^*}(T_0), z_{\phi^*}(T_0)) f(\phi^*) > 0$ that $c(T_0) = c(T_1)$. □

Observe that by our assumptions $\min_{\phi \in \left[\underline{\phi}, \overline{\phi}\right], z \in Z} \frac{\partial^2}{\partial z^2} v(z, \phi) > 0$, $\min_{\phi \in \left[\underline{\phi}, \overline{\phi}\right], z \in Z} \left|\frac{\partial^2}{\partial\phi\partial z} v(z, \phi)\right| > 0$, and $\max_{\phi \in \left[\underline{\phi}, \overline{\phi}\right], z \in Z} \cdot \left|\frac{\partial^3}{\partial\phi\partial z^2} v(z_{\phi^*}(T_0), \phi^*)\right| < \infty$. It follows that there exists $\delta > 0$ sufficiently small that

$$\forall \phi \in \left[\underline{\phi}, \overline{\phi}\right], \forall T \in \mathscr{T}^*, \frac{\partial^2}{\partial z^2} v(z_\phi(T), \phi) + \frac{\frac{\partial^3}{\partial\phi\partial z^2} v(z_\phi(T), \phi)}{\frac{\partial^2}{\partial\phi\partial z} v(z_\phi(T), \phi)} \delta > 0 \tag{A.88}$$

Since the properties of $\mathscr{T}^*$ imply that for all $T \in \mathscr{T}^*$, $T'\left(z_{\underline{\phi}}(T)\right) = 0$, $T'\left(z_{\phi_1'}(T)\right) = \frac{\mathrm{d}}{\mathrm{d}z} T^*\left(z_{\phi_1^*}(T^*)\right)$, we can fix some $\delta \in \left(0, \frac{\mathrm{d}}{\mathrm{d}z} T^*\left(z_{\phi_1^*}(T^*)\right)\right)$ which is also sufficiently small that (A.88) is satisfied and for all $T \in \mathscr{T}^*, \exists z \in \left(z_{\underline{\phi}}, z_1'\right), T'(z) = \delta$. Hence for all $T_0 \in \mathscr{T}^*$,

$$\exists \phi^* \in \left[\underline{\phi}, \phi_1'\right], T_0'(z_{\phi^*}(T_0)) = \delta \tag{A.89}$$

So choose some such T_0 and ϕ^*. Observe moreover that $z_{\phi^*}(T_0) > 0$. Then choose exists $T_1 \in \mathscr{T}^*$ such that for some $z^\circ > 0$ and some ϕ°, $z^\circ = z_{\phi^\circ}(T_0)$ and T_0, T_1 agree on $[0, z^\circ]$, and such that $z_{\phi^*}(T_0) = z_{\phi^*}(T_1)$, $T_0'(z_{\phi^*}(T_0)) = T_1'(z_{\phi^*}(T_1)) = \delta$, $T_0''(z_{\phi^*}(T_0)) \neq T_1''(z_{\phi^*}(T_1))$. Define $T_\varepsilon := (1 - \varepsilon) T_0 + \varepsilon T_1$. By Corollary A.2, there are two possibilities: Either (i) $\mathscr{T}_1^* \neq \emptyset$ or (ii) all $\mathscr{T}^* = \mathscr{T}_0^*$. In case (i), then we may assume that for all

$\varepsilon \in [0,1]$, $T_\varepsilon \in \mathscr{T}_1^*$. To see this, observe that it follows from Corollary A.2 that if $T_1 \in \mathscr{T}^*$ and T_1 is sufficiently close to T_0, then $T_\varepsilon \in \mathscr{T}_1^*$ for all $\varepsilon \in [0,1]$.[28] So if $T_0 \in \mathscr{T}_1^*$, we can simply choose T_1 sufficiently close. In case (ii), it is immediate that $T_\varepsilon \in \mathscr{T}_0^*$ for all $\varepsilon \in [0,1]$. In case (i), it follows from Lemma A.6 that there exists $c \in \mathbb{R} \setminus \{0\}$ such that $c(T_\varepsilon) = c, \forall \varepsilon \in [0,1]$, and that there exists $U \in \mathbb{R}$ and $z > 0$ such that $z_{\phi^*}(T_\varepsilon) = z, U_{\phi^*}(T_\varepsilon) = U$, and $T_\varepsilon'(z_{\phi^*}(T_\varepsilon)) = \delta, \forall \varepsilon \in [0,1]$. It follows that there exists $r \in \mathbb{R}_+$ such that $g_{\phi^*}(T_\varepsilon) f(\phi^*) = r, \forall \varepsilon \in [0,1]$. It follows that $r = ch(T_\varepsilon, \phi^*), \forall \varepsilon \in [0,1]$. So $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} h(T_\varepsilon, \phi) = 0$. Define $k(T,\phi)$ by $-k(T,\phi) + f(\phi) = h(T,\phi)$. It follows that $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} k(T_\varepsilon, \phi^*) = 0$. In case (ii), it follows from (A.87) that $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} k(T_\varepsilon, \phi^*) = 0$. So, in both cases, $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} k(T_\varepsilon, \phi^*) = 0$.

I will now attain a contradiction by showing that $\frac{\mathrm{d}}{\mathrm{d}\varepsilon} k(T_\varepsilon, \phi^*) \neq 0$. We have

$$
\begin{aligned}
& k(T_\varepsilon, \phi^*) \\
=& \frac{\mathrm{d}}{\mathrm{d}\phi}\left[\frac{T_\varepsilon'(z_{\phi^*}(T_\varepsilon))}{\frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_\varepsilon), \phi^*)} f(\phi^*)\right] \\
=& \left[\left(T_\varepsilon''(z_{\phi^*}(T_\varepsilon)) \left.\frac{\mathrm{d}}{\mathrm{d}\phi}\right|_{\phi=\phi^*} [z_\phi(T_\varepsilon)] f(\phi^*) + T_\varepsilon'(z_{\phi^*}(T_\varepsilon)) f'(\phi^*)\right) \frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_\varepsilon), \phi^*)\right. \\
& \left. - \left[\frac{\partial^3}{\partial\phi\partial z^2} v(z_{\phi^*}(T_\varepsilon), \phi^*) \left.\frac{\mathrm{d}}{\mathrm{d}\phi}\right|_{\phi=\phi^*} [z_\phi(T_\varepsilon)] + \frac{\partial^3}{\partial\phi^2\partial z} v(z_{\phi^*}(T_\varepsilon), \phi^*)\right] T_\varepsilon'(z_{\phi^*}(T_\varepsilon)) f(\phi^*)\right] \\
& \Bigg/ \left[\frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_\varepsilon), \phi^*)\right]^2 \\
=& \left\{\left[T_\varepsilon''(z_{\phi^*}(T_\varepsilon)) \frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_\varepsilon), \phi^*) - \frac{\partial^3}{\partial\phi\partial z^2} v(z_{\phi^*}(T_\varepsilon), \phi^*) T_\varepsilon'(z_{\phi^*}(T_\varepsilon))\right] \left.\frac{\mathrm{d}}{\mathrm{d}\phi}\right|_{\phi=\phi^*} [z_\phi(T_\varepsilon)] f(\phi^*)\right. \\
& \left. + \left[\frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_\varepsilon), \phi^*) f'(\phi^*) + \frac{\partial^3}{\partial\phi^2\partial z} v(z_{\phi^*}(T_\varepsilon), \phi^*) f(\phi^*)\right] T_\varepsilon'(z_{\phi^*}(T_\varepsilon))\right\} \\
& \Bigg/ \left[\frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_\varepsilon), \phi^*)\right]^2 \\
=& \left\{-\left[T_\varepsilon''(z_{\phi^*}(T_\varepsilon)) \frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_\varepsilon), \phi^*) - \frac{\partial^3}{\partial\phi\partial z^2} v(z_{\phi^*}(T_\varepsilon), \phi^*) T_\varepsilon'(z_{\phi^*}(T_\varepsilon))\right]\right. \\
& \times \frac{\frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_\varepsilon), \phi^*)}{T_\varepsilon''(z_{\phi^*}(T_\varepsilon)) + \frac{\partial^2}{\partial z^2} v(z_{\phi^*}(T_\varepsilon), \phi^*)} f(\phi^*) \\
& \left. + \left[\frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_\varepsilon), \phi^*) f'(\phi^*) + \frac{\partial^3}{\partial\phi^2\partial z} v(z_{\phi^*}(T_\varepsilon), \phi^*) f(\phi^*)\right] T_\varepsilon'(z_{\phi^*}(T_\varepsilon))\right\} \\
& \Bigg/ \left[\frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_\varepsilon), \phi^*)\right]^2 \\
=& \left\{-\left[T_\varepsilon''(z_{\phi^*}(T_0)) \frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_0), \phi^*) - \frac{\partial^3}{\partial\phi\partial z^2} v(z_{\phi^*}(T_0), \phi^*) \delta\right]\right. \\
& \times \frac{\frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_0), \phi^*)}{T_\varepsilon''(z_{\phi^*}(T_0)) + \frac{\partial^2}{\partial z^2} v(z_{\phi^*}(T_0), \phi^*)} f(\phi^*) \\
& \left. + \left[\frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_0), \phi^*) f'(\phi^*) + \frac{\partial^3}{\partial\phi^2\partial z} v(z_{\phi^*}(T_0), \phi^*) f(\phi^*)\right] \delta\right\} \\
& \Bigg/ \left[\frac{\partial^2}{\partial\phi\partial z} v(z_{\phi^*}(T_0), \phi^*)\right]^2
\end{aligned}
$$

[28] Closeness is measured relative to the norm ρ defined in Section 2.

$$
\begin{aligned}
&= \frac{-T_\varepsilon''\left(z_{\phi^*}\left(T_0\right)\right) + \frac{\frac{\partial^3}{\partial\phi\partial z^2}v\left(z_{\phi^*}(T_0),\phi^*\right)}{\frac{\partial^2}{\partial\phi\partial z}v\left(z_{\phi^*}(T_0),\phi^*\right)}\delta}{T_\varepsilon''\left(z_{\phi^*}\left(T_0\right)\right) + \frac{\partial^2}{\partial z^2}v\left(z_{\phi^*}\left(T_0\right),\phi^*\right)} f\left(\phi^*\right) + C \\
&= \frac{-T_0''\left(z_{\phi^*}\left(T_0\right)\right) - \varepsilon\left(T_1''\left(z_{\phi^*}\left(T_0\right)\right) - T_0''\left(z_{\phi^*}\left(T_0\right)\right)\right) + \frac{\frac{\partial^3}{\partial\phi\partial z^2}v\left(z_{\phi^*}(T_0),\phi^*\right)}{\frac{\partial^2}{\partial\phi\partial z}v\left(z_{\phi^*}(T_0),\phi^*\right)}T_0'\left(z_{\phi^*}\left(T_0\right)\right)}{T_0''\left(z_{\phi^*}\left(T_0\right)\right) + \varepsilon\left(T_1''\left(z_{\phi^*}\left(T_0\right)\right) - T_0''\left(z_{\phi^*}\left(T_0\right)\right)\right) + \frac{\partial^2}{\partial z^2}v\left(z_{\phi^*}\left(T_0\right),\phi^*\right)} f\left(\phi^*\right) + C,
\end{aligned}
$$

where the third equality substitutes an expression equivalent to $\frac{\mathrm{d}}{\mathrm{d}\phi}\Big|_{\phi=\phi^*}\left[z_\phi\left(T_\varepsilon\right)\right]$ applying the implicit function theorem to the agent's first-order condition, and C is a constant that does not depend on ε.

It follows that

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} k\left(T_\varepsilon,\phi^*\right) = &\left[-\left(T_1''\left(z_{\phi^*}\left(T_0\right)\right) - T_0''\left(z_{\phi^*}\left(T_0\right)\right)\right)\left[T_0''\left(z_{\phi^*}\left(T_0\right)\right) + \frac{\partial^2}{\partial z^2}v\left(z_{\phi^*}\left(T_0\right),\phi^*\right)\right] f\left(\phi^*\right)\right. \\
&\left. - \left(T_1''\left(z_{\phi^*}\left(T_0\right)\right) - T_0''\left(z_{\phi^*}\left(T_0\right)\right)\right)\left[-T_0''\left(z_{\phi^*}\left(T_0\right)\right) + \frac{\frac{\partial^3}{\partial\phi\partial z^2}v\left(z_{\phi^*}\left(T_0\right),\phi^*\right)}{\frac{\partial^2}{\partial\phi\partial z}v\left(z_{\phi^*}\left(T_0\right),\phi^*\right)}\delta\right]\right] f\left(\phi^*\right) \\
&\Bigg/\left[T_0''\left(z_{\phi^*}\left(T_0\right)\right) + \frac{\partial^2}{\partial z^2}v\left(z_{\phi^*}\left(T_0\right),\phi^*\right)\right]^2 \\
= &-\frac{\left(T_1''\left(z_{\phi^*}\left(T_0\right)\right) - T_0''\left(z_{\phi^*}\left(T_0\right)\right)\right)\left[\frac{\partial^2}{\partial z^2}v\left(z_{\phi^*}\left(T_0\right),\phi^*\right) + \frac{\frac{\partial^3}{\partial\phi\partial z^2}v\left(z_{\phi^*}(T_0),\phi^*\right)}{\frac{\partial^2}{\partial\phi\partial z}v\left(z_{\phi^*}(T_0),\phi^*\right)}\delta\right]}{\left[T_0''\left(z_{\phi^*}\left(T_0\right)\right) + \frac{\partial^2}{\partial z^2}v\left(z_{\phi^*}\left(T_0\right),\phi^*\right)\right]^2} f\left(\phi^*\right) \\
\neq &0,
\end{aligned}
$$

where the last non-equality follows from (A.88). As we established above that $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\big|_{\varepsilon=0} k\left(T_\varepsilon,\phi^*\right) = 0$, this leads to the desired contradiction, implying that (A.70) must be false and so completes the proof of Proposition A.4. □

A.5 The other direction

In this section I prove the other direction of the main theorem. In particular I prove:

Proposition A.5 *Under the assumptions of section 7, if g depends only on utility then g is rationalizable.*

Proof. If welfare weights depend only on utility, then we can write welfare weights as a function of a single argument $y = c - v_\phi(z)$, that is, of the form $g_\phi\left(c - v_\phi(z)\right)$. Now for each ϕ define the utility function

$$
u_\phi^g\left(\hat{y}\right) = \int_0^{\hat{y}} g_\phi\left(y\right)\mathrm{d}y
$$

Then if we define

$$
U_\phi^g\left(T\right) = z_\phi\left(T\right) - T\left(z_\phi\left(T\right)\right) - v_\phi\left(z_\phi\left(T\right)\right) \tag{A.90}
$$

Then consider the social welfare function

$$
W\left(T\right) = \int_{\underline{\phi}}^{\overline{\phi}} U_\phi^g\left(T\right) f\left(\phi\right)\mathrm{d}\phi
$$

Then define $\precsim$ as in (17), and the proof is essentially the same as the proof of Proposition 4. □