Consistently Comparing Tax Policies with Generalized Social Marginal Welfare Weights*

Katy Bergstrom† William Dodds‡

March 26, 2026

[Preliminary Draft: Do Not Distribute]

Abstract

This paper illustrates how to consistently compare tax policies while incorporating non-utilitarian ethical considerations by using the generalized social marginal welfare weight (GSMWW) approach of Saez and Stantcheva (2016). Our first contribution is to derive necessary and sufficient conditions for GSMWW to be consistent (i.e., that they do not generate "preference cycles"). Our second contribution is to illustrate that the result of Sher (2024), which says that all consistent GSMWW are fundamentally utilitarian, relies crucially on a functional form assumption for GSMWW; once you relax that assumption, there are many GSMWW which are consistent and non-utilitarian. Our third contribution is to illustrate that consistent GSMWW can always be constructed from an arbitrary set of "mechanical weights" that capture societal preferences over the direct distributional impacts of tax reforms holding behavior fixed. This result provides a bridge from survey-based evidence on redistributive preferences to a consistent framework for evaluating large tax reforms that can incorporate non-utilitarian ethical considerations. Finally, our results are robust to multidimensional tax schedules and heterogeneity, behavioral agents, general equilibrium effects, and non-smooth behavioral responses such as bunching.

Keywords: welfare weights, generalized social marginal welfare weights, consistency, transitivity

JEL: D60, D63, D71, H21, H23, I31

*The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors.

†Tulane University. Email: kbergstrom@tulane.edu

‡Tulane University. Email: wdodds@tulane.edu

1 Introduction

The vast majority of optimal tax papers assume a policymaker who is a utilitarian and therefore seeks to maximize a (weighted) sum of utilities (Mirrlees, 1971; Saez, 2001). While utilitarian frameworks are tractable and have various desirable properties (Kaplow and Shavell, 2001), there are a variety of ethical principles — such as egalitarianism, libertarianism, or equality of opportunity — which people care about yet are not encapsulated by a utilitarian objective (Weinzierl, 2014, 2017; Fleurbaey and Maniquet, 2018). Saez and Stantcheva (2016) proposed a novel way to incorporate different societal norms for fairness into the theory of optimal taxation: evaluate tax reforms and characterize the optimal tax schedule in terms of *generalized social marginal welfare weights* (GSMWW) which directly encode societal preferences for redistribution. This approach not only avoids the need to specify a social welfare function, but also allows great flexibility in choosing welfare weights to encode different ethical principles. Despite the intuitive appeal of the GSMWW approach, Sher (2024) argues that there is an important drawback of this approach: GSMWW generically generate inconsistent (i.e., non-transitive) rankings of tax policies. More specifically, Sher (2024) argues that GSMWW which are not *structurally utilitarian*, meaning that they are not derived from a (weighted) utilitarian social welfare function, will *always* yield inconsistent global rankings over tax policies. Hence, under any non-utilitarian GSMWW, there exist tax policies T_1, T_2, T_3 such that T_1 is more desirable than T_2 which is more desirable than T_3 which is more desirable than T_1. The core argument of this paper is that after relaxing the functional form assumption on GSMWW made by Saez and Stantcheva (2016) and Sher (2024), the GSMWW approach *can* be used to consistently compare tax policies with non-utilitarian ethical considerations; moreover, we illustrate how to *construct* consistent GSMWW starting from any given "mechanical weights" which capture societal preferences for redistribution ignoring behavioral responses to tax reforms.

Our paper therefore aims to illustrate that the GSMWW approach of Saez and Stantcheva (2016) is more flexible than previously recognized: the inconsistency result of Sher (2024) is not a fundamental limitation of the framework, but rather results from a restriction on how welfare weights are allowed to vary with the tax schedule. However, once we broaden the definition of GSMWW and allow them to depend more flexibly on the tax schedule, then they can be both consistent and non-utilitarian. Broadly, the present paper has three contributions. First, we provide a general characterization of consistency for GSMWW in Theorem 1. This general characterization relies on two key insights: (1) GSMWW are the values of an infinite dimensional vector field (which maps a tax function to a set of weights) and (2) consistency of comparisons implied by GSMWW is equivalent to an (infinite dimensional) path independence requirement (loosely, integrating GSMWW between two different tax schedules yields the same result regardless of how one moves

from the first schedule to the second schedule). This allows us to prove that a set of GSMWW are consistent if and only if they are the Gateaux derivative of some objective functional or, equivalently, if the Gateaux derivative of the GSMWW is symmetric.

The second contribution of our paper is to use this result to prove that there are GSMWW which yield consistent global comparisons that are *not* structurally utilitarian; the key insight is that this requires broadening the definition of GSMWW relative to Saez and Stantcheva (2016) and Sher (2024). Saez and Stantcheva (2016) and Sher (2024) define a set of GSMWW $g(c, z, n)$ where $-g(c, z, n)$ determines how much society values giving \$1 to a person of type n who earns income z and consumes c. We show that the consistency condition in Theorem 1 implies that any consistent $g(c, z, n)$ must satisfy a partial differential equation (PDE); solving this PDE yields that any consistent $g(c, z, n)$ takes the form $g(c, z, n) = u_c(c, z, n) \cdot \psi(u(c, z, n), n)$, and this is exactly the same conclusion as Sher (2024). Welfare weights of this form are "structurally utilitarian" in the sense that they depend on individual characteristics only through the individual's marginal utility scaled by a function of their utility. However, we then show that this result relies *critically* on the assumption that GSMWW must take the form $g(c, z, n)$, which restricts GSMWW in two important ways: (1) while the GSMWW for type n can depend on the tax level and marginal tax rate for type n (as these can be written in terms of c and z), it cannot depend on higher derivatives of the tax schedule and (2) the GSMWW for type n cannot depend on the tax level or marginal tax rates at incomes chosen by other types. We prove in Proposition 1 that when we relax these assumptions, there *are* consistent GSMWW which are not structurally utilitarian and we provide several examples. Intuitively, consistent GSMWW require this richer structure because (1) the envelope theorem does not apply with non-utilitarian social norms so that behavioral responses to tax changes (which depend on the second derivative of the tax schedule via the implicit function theorem) impact the desirability of a tax reform and (2) various principles of fairness rely on properties of the income distribution that are non-local to a specific type n (e.g., inequality aversion).

The third contribution of this paper is to show how to construct consistent GSMWW starting from a given set of "mechanical weights", which encode society's views on the desirability of a tax reform if no one changed their behavior in response to the tax reform. The key idea is that an arbitrary set of GSMWW conflates the desirability of *mechanical* changes and *behavioral* changes of tax reforms, and one cannot choose the desirability of mechanical and behavioral changes independently (in a consistent fashion) because they must be linked via some (unknown) objective function by Theorem 1. However, Proposition 2 illustrates that we *can* freely choose how to assess the desirability of tax reforms in absence of behavioral responses and then characterize the corresponding set of consistent GSMWW. In other words, Proposition 2 illustrates how society must value

behavioral responses to a tax change given societal values on the mechanical effects of a tax reform in order for GSMWW to be consistent. Viewed alternatively, Proposition 2 shows that if we interpret any given set of weights as reflecting societal preferences in absence of behavioral responses (which is often implicit in surveys that seek to recover social preferences over redistribution as in Appendix C of Saez and Stantcheva (2016) or Capozza and Srinivasan (2025)), then we can always turn these into a consistent set of GSMWW as long as one correctly includes the implied desirability of behavioral changes.

The rest of the paper proceeds as follows. Section 2 presents some mathematical preliminaries, defines GSMWW, and presents our first main result characterizing consistency. Section 3 illustrates how the most general result from Sher (2024) (including income effects and private information) follows from Theorem 1. Section 4 shows that this result no longer holds (so that there are consistent, non-utilitarian GSMWW) when GSMWW are allowed to depend on the tax schedule in more complex ways. Section 4 also provides examples of consistent GSMWW which incorporate libertarian social preferences and inequality aversion. Section 5 illustrates how to construct consistent GSMWW starting from any given "mechanical weights" which assess desirability of tax reforms in absence of behavioral responses; Section 5 also provides an example of this construction using poverty alleviation mechanical weights. Finally, Section 6 illustrates how our main results carry over to settings with multidimensional tax schedules, behavioral agents who make optimization errors, general equilibrium effects of taxation, and non-smooth behavioral responses. Section 7 concludes.

2 Generalized Welfare Weights and Consistency

2.1 Notation and Mathematical Preliminaries

Our over-arching goal is to illustrate how to compare the desirability of tax policies in a consistent fashion with non-utilitarian ethical principles. We work with a set $\mathcal{T}$ of admissible tax schedules, where the choice of $\mathcal{T}$ encodes the regularity conditions imposed on the tax schedules being compared. Different results in the paper require different levels of regularity: our characterization of consistency in Theorem 1 requires only that $\mathcal{T}$ be convex, while results that rely on individuals responding smoothly to tax reforms — such as Proposition Sher and Proposition 2 — require $\mathcal{T}$ to be restricted to smooth tax schedules under which all individuals have a unique optimum. We are explicit about which assumptions on $\mathcal{T}$ are required for each result throughout the paper.

We consider a population of individuals indexed by a type vector $\mathbf{n} = (n_1, n_2, ..., n_K) \in \mathbf{N}$ on compact $\mathbf{N}$ distributed according to distribution $F(\mathbf{n})$ with density $f(\mathbf{n})$. Individuals choose $\mathbf{z} = (z_1, z_2, ..., z_J) \in A(\mathbf{n}) \subseteq \mathbb{R}^J$ to maximize a utility function subject to a

budget constraint given a tax schedule, $T(\mathbf{z})$, which is a function of choice variables $\mathbf{z}$:

$$\begin{aligned} &\max_{\mathbf{z}\in A(\mathbf{n})} \; u(c, \mathbf{z}; \mathbf{n}) \\ &\text{s.t. } \; c = y(\mathbf{z}) - T(\mathbf{z}) \end{aligned} \tag{1}$$

where c is numeraire consumption and is a function of choices $\mathbf{z}$ as well as the tax schedule $T(\mathbf{z})$. For example, z_i might represent income from a particular source (e.g., labor or savings) or consumption of a particular good or the z_i's could represent incomes in various time periods.

Let $\mathbf{z}(\mathbf{n})$ denote optimal choices for type $\mathbf{n}$ under the given tax schedule and let $\mathbf{Z} = \{\mathbf{z}(\mathbf{n}) | \mathbf{n} \in \mathbf{N}\}$ denote the set of $\mathbf{z}$ that are chosen by some type $\mathbf{n}$ under the given tax schedule. While we omit additional arguments to make expressions more readable, it is very important to note that $\mathbf{z}(\mathbf{n})$ also depends on the tax schedule T.

Next, we will frequently utilize the concept of a *Gateaux derivative* which we define as in the Encyclopedia of Mathematics:

Definition 1. *Let $M : \mathcal{T} \mapsto \mathbb{R}$ be a functional on a normed vector space $\mathcal{T}$. We say that M is Gateaux differentiable at a $T \in \mathcal{T}$ if $\exists$ a bounded linear functional, DM_T, which we call the Gateaux derivative, such that for any $\tau \in \mathcal{T}$:*[1]

$$\lim_{\epsilon \to 0} \frac{M(T + \epsilon\tau) - M(T)}{\epsilon} = DM_T(\tau)$$

We refer to $\lim_{\epsilon\to 0} \frac{M(T+\epsilon\tau)-M(T)}{\epsilon}$ as the Gateaux variation of M at T in the direction of τ.

Definition 2. *$B : \mathcal{T} \mapsto \mathbb{R}$ is a bounded linear functional if $B(a_1T_1 + a_2T_2) = a_1B(T_1) + a_2B(T_2)$ $\forall a_1, a_2 \in \mathbb{R}, T_1, T_2 \in \mathcal{T}$ and $\exists P$ s.t. for any $T \in \mathcal{T}$ then $|B(T)| < P \|T\|_{\mathcal{T}}$.*

To build intuition, when $\mathcal{T} = \mathbb{R}^k$, the Gateaux variation is the directional derivative and is always equal to the Gateaux derivative (the gradient) multiplied by the direction vector. For an infinite dimensional example, when $\mathcal{T}$ represents the set of compactly supported continuous functions (endowed with the supnorm), every bounded linear functional can be represented as follows for some measure Γ (this is the Riesz-Markov-Kakutani Representation Theorem, Theorem 6.19 of Rudin (1974)):

$$B(T) = \int_{\mathbf{S}} T(\mathbf{s}) d\Gamma(\mathbf{s}) \tag{2}$$

Next, let us define "generalized welfare weights" as in Saez and Stantcheva (2016) as well as the notion of "consistency" as in Sher (2024).

Definition 3. *The generalized welfare weight for an individual of type $\mathbf{n}$ equals $g(\mathbf{n}; T)$,*

[1] Gateaux differentiability in Definition 1 is stronger than existence of the Gateaux variation (which does not require linearity and boundedness) but weaker than Frechet differentiability because we do not require uniform convergence in all directions τ. We define these objects as in the Encyclopedia of Mathematics, but note that what we call the "Gateaux variation" is sometimes referred to by other authors as the Gateaux derivative.

where the function g depends on characteristics $\mathbf{n}$ and the tax schedule T.[2] *A small tax reform from $T(\mathbf{z})$ to $T(\mathbf{z}) + \epsilon\tau(\mathbf{z})$ is desirable if and only if*

$$\int_{\mathbf{N}} g(\mathbf{n}; T)\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})d\mathbf{n} > 0$$

The idea behind generalized welfare weights is to encapsulate society's (or a policymaker's) valuation of small tax reforms. The desirability criterion in Definition 3 has a natural interpretation: $-g(\mathbf{n}; T)$ represents the marginal social value of giving \$1 to an individual $\mathbf{n}$ when the tax schedule is T (i.e., $-g(\mathbf{n}; T)$ represents the social value of changing their tax liability, $\tau(\mathbf{z}(\mathbf{n}))$, by $-\$1$). The key innovation of Saez and Stantcheva (2016) is therefore to evaluate the desirability of tax reforms without explicitly appealing to (or even writing down) an objective function. In other words, generalized welfare weights allow one to measure the societal value of marginal tax changes without writing down a particular objective function. Practically, eliciting societal preferences over tax *changes* may be significantly easier than eliciting societal preferences over arbitrary states of the world so that eliciting $g(\mathbf{n}; T)$ is likely to be significantly easier than eliciting entire objective functions. Finally, the structure of Definition 3 is identical to Saez and Stantcheva (2016) and Sher (2024) with one key departure: Definition 3 allows $g(\mathbf{n}; T)$ to depend arbitrarily on the *entire* tax schedule T, rather than just on the income and consumption for type n that arise when they optimize under the tax schedule T; hence, $g(\mathbf{n}; T)$ should *not* be understood as a weight $g(\mathbf{n}, T(\mathbf{z}(\mathbf{n}))$ that depends on type $\mathbf{n}$ and the tax level at the choice $\mathbf{z}$ chosen by type $\mathbf{n}$. This generalization turns out to be essential for obtaining consistent non-utilitarian rankings, as we show in Section 4.

While Saez and Stantcheva (2016) limit their analysis to local reforms, our goal (as in Sher (2024)) is to use generalized welfare weights to make global comparisons of tax policies by integrating up the desirability criterion for successive small reforms. Formally, we will consider families of tax schedules $T^{(\epsilon)}(\mathbf{z})$ parametrized by ϵ where $T^{(\epsilon)}(\mathbf{z})$ is differentiable in ϵ. In essence, the parameter ϵ characterizes some path between various tax schedules.

Definition 4. *A tax policy $T^{(\epsilon_2)}(\mathbf{z})$ is more desirable than a tax policy $T^{(\epsilon_1)}(\mathbf{z})$ if and only if:*

$$\Delta W(T^{(\epsilon_2)}, T^{(\epsilon_1)}) \equiv -\int_{\epsilon_1}^{\epsilon_2}\int_{\mathbf{N}} g(\mathbf{n}; T^{(\epsilon)})T_{\epsilon}^{(\epsilon)}(\mathbf{z}(\mathbf{n}))f(\mathbf{n})d\mathbf{n}d\epsilon > 0$$

where $T_{\epsilon}^{(\epsilon)}(\mathbf{z}) \equiv \left[\frac{d}{dv}T^{(v)}(\mathbf{z})\right]\Big|_{v=\epsilon}$. Hence, $\Delta W(T^{(\epsilon_2)}, T^{(\epsilon_1)})$ denotes the "social desirability" of moving from tax schedule $T^{(\epsilon_1)}$ to tax schedule $T^{(\epsilon_2)}$ generated by integrating the generalized welfare weights along the path parametrized by $T^{(\epsilon)}(\mathbf{z})$. One nuance to notice with Definition 3 and Definition 4 is that they are entirely independent of the budgetary consequences of a tax schedule; in other words, Definitions 3 and 4 allow us to compare

[2] Saez and Stantcheva (2016) call these objects "generalized social marginal welfare weights", but we drop the "social marginal" for brevity.

the societal value of a tax policy in absence of budgetary consequences.[3]

Next, let us define "consistency" for a set of generalized welfare weights:

Definition 5. *Generalized welfare weights* $g(\mathbf{n}; T)$ *are* inconsistent *if there exists a set of tax policies* $T^{(\epsilon)}(\mathbf{z})$ *with* $\epsilon \in [0,1]$ *and* $T^{(0)}(\mathbf{z}) = T^{(1)}(\mathbf{z})$ *such that for some* $\epsilon_1, \epsilon_2 \in (0,1)$*:*

$$\Delta W(T^{(\epsilon_1)}(\mathbf{z}), T^{(0)}(\mathbf{z})) > 0 \text{ and } \Delta W(T^{(\epsilon_2)}(\mathbf{z}), T^{(\epsilon_1)}(\mathbf{z})) > 0 \text{ and } \Delta W(T^{(1)}(\mathbf{z}), T^{(\epsilon_2)}(\mathbf{z})) > 0 \tag{3}$$

Generalized welfare weights are consistent *if there do not exist any such tax schedules satisfying Equation* (3)*.*

Definition 5 is the same notion of consistency defined in Sher (2024). Thus, inconsistent generalized welfare weights $g(\mathbf{n}; T)$ generate "preference cycles" wherein a policy $T^{(\epsilon_1)}(\mathbf{z})$ is preferred to $T^{(0)}(\mathbf{z})$, $T^{(\epsilon_2)}(\mathbf{z})$ preferred to $T^{(\epsilon_1)}(\mathbf{z})$, yet $T^{(1)}(\mathbf{z}) = T^{(0)}(\mathbf{z})$ is preferred to $T^{(\epsilon_2)}(\mathbf{z})$; in contrast, consistent generalized welfare weights do not admit any such preference cycles.

Remark 1. *The consistency requirement in Definition 5 is analogous to transitivity in standard social choice theory (Arrow, 1951). A policymaker whose welfare weights are inconsistent in the sense of Definition 5 violates the weak axiom of revealed preference; hence, consistency is not merely a technical convenience but a basic rationality requirement for normative policy evaluation.*

2.2 Characterization of Consistent Generalized Welfare Weights

Next, we will illustrate how to characterize consistent generalized welfare weights in a very general sense. There are two key insights. First, the weights $g(\mathbf{n}; T)$ are effectively the "values" of an infinite dimensional vector field. Second, consistency of $g(\mathbf{n}; T)$ is equivalent to the "path independence" property:

Lemma 1. *Suppose that the set of tax functions* $\mathcal{T}$ *is path connected. Then* $g(\mathbf{n}; T)$ *is consistent if and only if* G *is* path independent *so that for any two paths* $T^{(\epsilon)}(\mathbf{z}), \hat{T}^{(\epsilon)}(\mathbf{z}) : [0,1] \mapsto \mathcal{T}$ *with* $T^{(0)}(\mathbf{z}) = \hat{T}^{(0)}(\mathbf{z})$ *and* $T^{(1)}(\mathbf{z}) = \hat{T}^{(1)}(\mathbf{z})$*:*

$$\int_0^1 \int_{\mathbf{N}} g(\mathbf{n}; T^{(\epsilon)}) T_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n})) f(\mathbf{n}) d\mathbf{n} d\epsilon = \int_0^1 \int_{\mathbf{N}} g(\mathbf{n}; \hat{T}^{(\epsilon)}) \hat{T}_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n})) f(\mathbf{n}) d\mathbf{n} d\epsilon$$

Proof. See Appendix A.1. □

Lemma 1 is the infinite-dimensional analogue of the classical result from vector calculus that a vector field is conservative if and only if its line integrals are path independent. The proof follows the same logic as the finite-dimensional case. The assumption that $\mathcal{T}$ is path connected is necessary to ensure that any two tax schedules can be connected by a path that stays within $\mathcal{T}$.

[3] This is analogous to how a utility function allows one to compare different alternatives regardless of whether they are affordable.

Given this new abstract formulation of consistency, we can develop the following straight-forward characterization of consistency for generalized welfare weights:

Theorem 1. *Suppose that $g(\mathbf{n};T)$ is continuous in T and that the set of tax functions $\mathcal{T}$ is convex. Then the following three conditions are equivalent:*

1. *A set of generalized welfare weights $g(\mathbf{n};T)$ is consistent.*
2. *$\int_{\mathbf{N}} g(\mathbf{n};T)\tau(\mathbf{z}(\mathbf{n}))d\mathbf{n}$ is the Gateaux derivative of some objective functional.*
3. *The Gateaux variation of the functional $\int_{\mathbf{N}} g(\mathbf{n};T)\tau(\mathbf{z}(\mathbf{n}))d\mathbf{n}$ in the direction γ is symmetric in τ and γ (assuming this Gateaux variation exists and is continuous).*

Proof. See Appendix A.2. □

Theorem 1 provides a characterization of consistent generalized welfare weights in terms of two conditions that are each independently useful. Condition 2 — that consistent welfare weights must arise as the Gateaux derivative of some objective functional — gives consistency a clear economic interpretation: a policymaker whose generalized welfare weights are consistent must be behaving *as if* they were maximizing some well-defined objective function, even if they never explicitly wrote one down. Condition 3 — the symmetry of the Gateaux variation — allows one to *verify* whether a set of generalized welfare weights $g(\mathbf{n};T)$ are consistent without requiring one to find the objective functional explicitly. In words, Condition 3 requires that the *change*, as we move in the direction γ, of the desirability of a tax reform in direction τ is the same as the *change*, as we move in the direction τ, of the desirability of a tax reform in the direction γ.

The core idea behind Theorem 1 is the recognition that (1) $g(\mathbf{n};T)$ are effectively the values of an infinite dimensional vector field and that (2) consistency is equivalent to this infinite dimensional vector field possessing the path independence property from Lemma 1.[4] Given these two insights, the three conditions in Theorem 1 follow as infinite-dimensional analogues of three equivalent characterizations of conservative vector fields in classical vector calculus: path independence of line integrals, existence of a scalar potential, and vanishing curl (i.e., symmetry of the Jacobian). Hence, the primary contribution of Theorem 1 is translating these results into the framework of generalized welfare weights and showing that they can be used to provide a tractable and economically interpretable characterization of consistent generalized welfare weights.[5]

[4]Technically, the relevant infinite dimensional vector field is a mapping $G : \mathcal{T} \mapsto \mathcal{T}^*$, where $\mathcal{T}^*$ is the (dual) space of bounded linear functionals on $\mathcal{T}$. G takes a tax schedule T as an argument and then spits out a linear functional which, when evaluated in the direction τ, is given by $\int_{\mathbf{N}} g(\mathbf{n};T)\,\tau(\mathbf{z}(\mathbf{n}))\,f(\mathbf{n})\,d\mathbf{n}$.

[5]While we could not find exact infinite dimensional analogues to the finite dimensional results written down anywhere, these ideas are almost surely known to mathematicians and should therefore not be construed as the main contribution of Theorem 1: for instance, the fact that Condition 3 implies Condition 2 in Theorem 1 (which is is an infinite dimensional version of the Poincare Lemma) is proved in an infinite dimensional setting under stronger smoothness assumptions and weaker assumptions on the domain in Kriegl and Michor (1997).

Remark 2. *Theorem 1 is stated without any assumptions on the dimensionality of* $\mathbf{z}$ *or* $\mathbf{n}$*, on the smoothness of individual behavior, or on the form of the tax schedule. In particular, it applies equally to settings with multidimensional heterogeneity and tax schedules, behavioral agents who make optimization errors, and non-smooth behavioral responses such as bunching. This generality is possible precisely because Theorem 1 characterizes consistency purely in terms of properties of* $g(\mathbf{n}; T)$ *— that is, in terms of how generalized welfare weights vary with the tax schedule — rather than in terms of the structural primitives of the individual optimization problem. Section 6 will exploit the fact that this characterization applies unchanged to these richer environments.*

3 Relationship to Sher (2024)

Section 2 has argued that generalized welfare weights are consistent if and only if they are the gradient of some objective function or, equivalently, if they satisfy a symmetry condition that the Gateaux derivative of $\int_{\mathbf{N}} g(\mathbf{n}; T)\tau(\mathbf{z}(\mathbf{n}))d\mathbf{n}$ in the direction of γ is symmetric in γ and τ. While Theorem 1 does not place any restrictions on the dimensions of the type space $\mathbf{N}$ or the choice space $\mathbf{Z}$, for the purposes of relating our results back to Sher (2024), we will now suppose that $n \in N$ is a scalar and $z \in Z$ is a scalar as well. Sher (2024) argues that if generalized welfare weights take the form $g(c, z, n)$, then they are consistent if and only if weights are "structurally utilitarian":[6]

Definition 6. *Generalized welfare weights are "structurally" utilitarian if they take the following form:*

$$g(c, z, n) = u_c(c, z, n)\psi(u(c, z, n), n) \tag{4}$$

for some function $\psi(u, n)$*.*

Weights that take the form in Equation (4) can be derived from a utilitarian objective and are therefore consistent by Theorem 1.[7] Next, let us impose a few regularity conditions on the set of tax policies that we will be comparing, $\mathcal{T}$, both of which are standard in the optimal tax literature (Mirrlees, 1971; Saez, 2001) and are also made in Sher (2024):

Assumption 1. *Suppose that* $\mathcal{T}$ *is a convex set of smooth tax schedules that is open within the set of compactly supported smooth functions. Suppose that for all* $T \in \mathcal{T}$ *that all individuals have a unique optimum and second order conditions are satisfied strictly (i.e., utility is strictly concave under their optimal choice of* z*) and that the mapping* $n \mapsto z$ *is monotonic and differentiable.*[8]

[6] Most of Sher (2024) is developed for the quasi-linear case wherein $u(c, z, n) = c - v(z, n)$, but Section VI of his paper discusses how the results carry over to more general utility functions under additional assumptions.

[7] To see this, take the Gateaux derivative of $\int_N \Psi\left(u\left(z(n) - T(z(n)) - \epsilon\tau(z(n)), z(n), n\right), n\right) f(n)dn$ and appeal to the envelope theorem so that this Gateaux derivative equals: $-\int_N \Psi_c(u(c(n), z(n), n), n)u_c(c(n), z(n), n)\tau(z(n))f(n)dn$ and set $\psi(u(c(n), z(n), n), n) \equiv \Psi_c(u(c(n), z(n), n), n)$.

[8] Note, implicitly, we have also assumed that all types n have unrestricted choice sets over all $z \in \mathbb{R}_+$

Next, we want to show that any set of generalized welfare weights of the form $g(c, z, n)$ that are consistent (when comparing tax schedules in $\mathcal{T}$ that satisfies Assumption 1) must be structurally utilitarian. Theorem 1 Condition 3 tells us that for a set of generalized welfare weights to be consistent on a convex set $\mathcal{T}$, then the Gateaux variation of

$$\int_N g(c(z(n)), z(n), n)\, \gamma(z(n))\, f(n)\, dn \tag{5}$$

in the direction of $\tau(z)$ must be symmetric in τ and γ. Now suppose that we consider how Equation (5) varies if we perturb the tax schedule from $T(z)$ to $T(z) + \epsilon\tau(z)$, recalling that $z(n)$ is a function of the tax schedule. To begin with, note that for any given tax schedule $T(z) + \epsilon\tau(z)$, individuals' optima satisfy the following first order condition:

$$u_1(z - T(z) - \epsilon\tau(z), z/n; v)\,(1 - T'(z) - \epsilon\tau'(z)) + \frac{1}{n}u_2(z - T(z) - \epsilon\tau(z), z/n; v) = 0$$

Under Assumption 1 we can apply the implicit function theorem to determine the impacts of a tax perturbation on each individual's optimal income:

$$\begin{aligned}\frac{\partial z(n)}{\partial \epsilon} &= \frac{u_c\tau'(z) + [u_{cc}(1 - T'(z)) + u_{cz}]\,\tau(z)}{-u_cT''(z) + u_{cc}(1 - T'(z))^2 + 2(1 - T'(z))u_{cz} + u_{zz}} \\ &\equiv \xi(n)\tau'(z(n)) + \eta(n)\tau(z(n))\end{aligned} \tag{6}$$

where $\xi(n) \equiv \frac{u_c}{-u_cT''(z)+u_{cc}(1-T'(z))^2+2(1-T'(z))u_{cz}+u_{zz}}$ and
$\eta(n) \equiv \frac{[u_{cc}(1-T'(z))+u_{cz}]}{-u_cT''(z)+u_{cc}(1-T'(z))^2+2(1-T'(z))u_{cz}+u_{zz}}$.

Next, we differentiate Equation (5) with respect to ϵ, recognizing that the derivative of z w.r.t. ϵ is given by Equation (6):

$$\begin{aligned}&\frac{d}{d\epsilon}\int_N g(c(n), z(n), n)\, \gamma(z(n))\, f(n)\, dn = -\int_N g_c\, \gamma(z(n))\tau(z(n))\, f(n)\, dn \\ &+ \int_N \left[g_c\, c'(z(n))\gamma(z(n)) + g_z\gamma(z(n)) + g\,\gamma'(z(n))\right]\frac{dz(n)}{d\epsilon}\, f(n)\, dn\end{aligned} \tag{7}$$

Noting that the first integral in Equation (7) is already symmetric in τ and γ, we focus on the second integral. Substituting $\frac{dz(n)}{d\epsilon} = \xi(n)\,\tau'(z(n)) + \eta(n)\,\tau(z(n))$, we have:

$$\begin{aligned}&\int_N \left[g_c\, c'(z(n))\gamma(z(n)) + g_z\gamma(z(n)) + g\,\gamma'(z(n))\right]\frac{dz(n)}{d\epsilon}\, f(n)\, dn \\ &= \int_N \Big(\xi(n)\,\tau'(z(n)) + \eta(n)\,\tau(z(n))\Big)\Big[g_c\,\gamma(z(n)) \\ &+ g_z\,\gamma(z(n)) + g\,\gamma'(z(n))\Big]\, f(n)\, dn\end{aligned} \tag{8}$$

and that the utility function $u(c, z, n)$ is sufficiently smooth (i.e., twice continuously differentiable) in all arguments; these are again all standard assumptions in the optimal tax literature.

Symmetry requires that Equation (8) is equal to:

$$\int_N \Big(\xi(n)\,\gamma'(z(n)) + \eta(n)\,\gamma(z(n))\Big)\Big[g_c\,\tau(z(n)) + g_z\,\tau(z(n)) + g\,\tau'(z(n))\Big]\,f(n)\,dn \tag{9}$$

Canceling terms, we get that symmetry requires:

$$\int_N \{\xi(n)\,[g_c\,c'(z(n)) + g_z] + \eta(n)g\} \times [\tau'(z(n))\gamma(z(n)) - \tau(z(n))\gamma'(z(n))]\,f(n)dn \tag{10}$$

Because τ and γ are arbitrary, we can set $\gamma(z) = 1$ and choose $\tau(z)$ arbitrarily so that $(\tau'(z)\gamma(z) - \tau(z)\gamma'(z)) = \tau'(z)$ for an arbitrarily chosen $\tau'(z)$.[9] Hence, by the fundamental lemma of the calculus of variations, Equation (10) holds if and only if:

$$\xi(n)\,[g_c(c,z,n)\,c'(z) + g_z(c,z,n)] + \eta(n)g(c,z,n) = 0 \tag{11}$$

Using the expressions for $\xi(n)$ and $\eta(n)$ from Equation (6) along with the fact that $u_c c'(z) + u_z = 0$, consistency therefore requires that for all c, z, n that arise optimally under a $T \in \mathcal{T}$:

$$u_c\left[g_c(c,z,n)\,\frac{-u_z}{u_c} + g_z(c,z,n)\right] - \left[u_{cc}\frac{-u_z}{u_c} + u_{cz}\right]\,g(c,z,n) = 0 \tag{12}$$

Thus, in order to satisfy symmetry of the second Gateaux derivative (which is necessary and sufficient for consistency by Theorem 1), Equation (12) must hold. Equation (12) has a natural interpretation: the first term captures the directional derivative of $g(c,z,n)$ along the indifference curve of u through (c,z): the vector $(-u_z/u_c, 1)$ is tangent to the level set of u in the (c,z) plane, so $-u_z g_c/u_c + g_z$ measures how g varies as one moves along an indifference curve while holding n fixed. The second term captures the level of g itself, weighted by income effects. Hence, Equation (12) is a first-order PDE that restricts how $g(c,z,n)$ can vary across indifference curves: loosely, g must vary across indifference curves in proportion to itself, which forces g to be proportional to u_c scaled by a function of utility $u(c,z,n)$ and n. More formally:

Proposition Sher. *Suppose Assumption 1 holds. The solution to the partial differential equation (12) over the set $\{(c,z)\}^n_{T\in\mathcal{T}}$ is given by $g(c,z,n) = u_c(c,z,n)\cdot\psi(u(c,z,n),n)$. Hence, by Theorem 1, generalized welfare weights $g(c,z,n)$ are consistent (when comparing tax policies $T \in \mathcal{T}$) if and only if they are equal to $g(c,z,n) = u_c(c,z,n)\cdot\psi(u(c,z,n),n)$ for some function ψ.*

Proof. See Appendix A.3. □

Proposition Sher is not new — this result was proved by Sher (2024) using a different argument. We present it here to illustrate how it follows as a special case of the general consistency condition in Theorem 1. The takeaway is that if generalized welfare weights

[9] Note, we have used the assumption that $T + \epsilon\tau \in \mathcal{T}$ for any smooth τ and sufficiently small ϵ by openness of $\mathcal{T}$ in the set of smooth functions.

are restricted to be a function $g(c, z, n)$, then the only possible consistent weights are structurally utilitarian as in Definition 6.[10] To build intuition, consider the quasi-linear case where $u(c, z, n) = c - v(z, n)$ so that $u_c = 1$ and there are no income effects (i.e., $\eta(n) = 0$). In this case, the symmetry condition (12) simplifies to:

$$g_c(c, z, n)(1 - T'(z)) + g_z(c, z, n) = 0$$

Intuitively then, $g(c, z, n)$ must be a function that satisfies the first order condition for each type n under every tax schedule $T \in \mathcal{T}$ and, heuristically, this is only possible if $g(c, z, n)$ is a function that depends solely on $u(c, z, n)$ and n.

Finally, Proposition Sher is slightly weaker than the main result of Sher (2024) in one additional respect: Sher (2024) shows that inconsistency arises even when attention is restricted to tax schedules that generate the same aggregate revenue level. We establish this using our framework in Appendix B.2; the core idea is that there are still enough degrees of freedom in the set of tax schedules that generate the same revenue level to infer that any consistent generalized welfare weights must satisfy Equation (12).

4 Consistent Generalized Welfare Weights That are Not Structurally Utilitarian

Proposition Sher relies on two assumptions: (1) individuals respond smoothly to tax changes (Assumption 1) and (2) generalized welfare weights take the form $g(c, z, n)$. As we will discuss in Section 6.4, the first assumption is inessential and Proposition Sher extends without it. In contrast, one of the key insights of the present paper is that the second assumption, that generalized welfare weights take the form $g(c, z, n)$, is *critical* to proving Proposition Sher. Welfare weights of the form $g(c, z, n)$ can depend on the tax level $T(z) = z - c(z)$ and the marginal tax rate $1 - T'(z)$ faced by type n (as $1 - T'(z) = \frac{u_c(c,z,n)}{u_z(c,z,n)}$ by the FOC), but are restricted in two important ways: they cannot depend on higher derivatives of the tax schedule (such as $T''(z)$, which impacts behavioral responses via the implicit function theorem), and the weight assigned to type n cannot depend on the tax schedule at incomes chosen by other types n'. Once either restriction is relaxed, consistent non-utilitarian welfare weights exist:

Proposition 1. *Suppose Assumption 1 holds. There are consistent generalized welfare*

[10] One subtlety in Proposition Sher is that the restriction to structurally utilitarian weights is derived by requiring the symmetry condition (12) to hold over the set $\{(c, z)\}^n_{T \in \mathcal{T}}$ that arise optimally for type n under tax schedules in $\mathcal{T}$ of suitably "well-behaved" tax schedules. This has an important implication for the scope of the consistency result: Proposition Sher establishes that $g(c, z, n)$ must be structurally utilitarian at the allocations that arise under tax schedules in the set $\mathcal{T}$, but does not directly restrict g at allocations that lie outside this set. In other words, the inconsistency result of Sher (2024) applies to global rankings of tax policies *within* a well-behaved set $\mathcal{T}$, rather than to rankings of arbitrary tax policies. In Appendix B.9, we show that under a single-crossing condition, the symmetry condition extends to all (c, z) chosen under *any* continuous tax schedule — not just those that are well-behaved — so that the restriction to structurally utilitarian weights holds more broadly.

weights $g(n;T)$ which are not structurally utilitarian.

The proof of Proposition 1 is constructive: we present two explicit examples of consistent, non-utilitarian generalized welfare weights. Each is derived from a well-defined objective function that captures normative principles frequently discussed in the public finance literature: (1) libertarian concerns about the share of income extracted by the government and (2) distributional concerns about inequality as measured by the Gini coefficient. In both cases, the generalized welfare weights depend on the tax schedule in a complex way and therefore cannot be expressed using the simple functional form $g(c,z,n)$ yet are consistent by Theorem 1 precisely because they arise as the Gateaux derivative of a well-defined objective functional.

4.1 Libertarian-Utilitarian Weights

Our first example will provide an example wherein consistent generalized welfare weights $g(n;T)$ depend not only on $T(z)$ and $T'(z)$ (which can be expressed as functions of c and z via $c(z) = z - T(z)$ and the first order condition) but also on $T''(z)$, which has, in general, no fixed relationship to c, z, and n. Consider the following objective function wherein the policymaker cares about maximizing a weighted sum of utilities $\phi(n)\, u\left(z(n) - T(z(n)) - \epsilon\tau(z(n)), z, n\right)$ but further weights these utilities by a factor of $\left[\frac{z(n)}{z(n)-T(z(n))}\right]$. Hence, the policymaker has the following objective function:

$$\int_N \left[\frac{z(n)}{z(n)-T(z(n))}\right] \phi(n)\, u\left(z(n) - T(z(n)) - \epsilon\tau(z(n)), z, n\right)\, f(n)dn \tag{13}$$

This second weighting factor $\frac{z(n)}{z(n)-T(z(n))}$ is simply the income to consumption ratio and thereby encodes the "libertarian principle" that individuals become more deserving of tax cuts the higher proportion of their income that is already being taxed *above and beyond the utility impact of a tax cut* (Saez and Stantcheva, 2016). Using the language of Sher (2024), this is a "libertarian-utilitarian mix" objective. Let us illustrate what the generalized welfare weights derived from this objective function look like under the simplifying assumption that $u(c,z,n) = c - \frac{(z/n)^{1+k}}{1+k}$. First, let us calculate the Gateaux derivative of this objective function assuming all individuals have a unique optimum under the smooth tax schedule T and that $f(n) \to 0$ as $n \to \underline{n}$ or $n \to \overline{n}$:

$$\begin{aligned}
&\frac{d}{d\epsilon}\left(\int_N \left[\frac{z(n)}{z(n)-T(z(n))-\epsilon\tau(z(n))}\right] \phi(n) \left(z(n) - T(z(n)) - \epsilon\tau(z(n)) - \frac{(z(n)/n)^{1+k}}{1+k}\right) f(n)dn\right) \\
&= -\int_N \left[\frac{z(n)}{z(n)-T(z(n))}\right] \phi(n)\,\tau(z(n))\, f(n)dn \\
&+ \int_N \left[\frac{z(n)}{[z(n)-T(z(n))]^2}\right] \phi(n)\, u(c(n),z(n),n)\tau(z(n))f(n)dn \\
&+ \int_N \phi(n)\frac{z(n)T'(z(n)) - T(z(n))}{(z(n)-T(z(n)))^2}\frac{\tau'(z(n))}{T''(z(n)) + \frac{kz(n)^{k-1}}{n^{k+1}}} u(c(n),z(n),n)f(n)dn
\end{aligned} \tag{14}$$

where we have applied (1) the first order condition to infer $\left[(1-T'(z(n))-\left(\frac{z(n)}{n}\right)^k\frac{1}{n}\right]\frac{dz}{d\epsilon}=0$ and (2) the implicit function theorem (which is valid under the assumption that all individuals have a unique maximum and T is smooth) to the first order condition $1-T'(z)-\epsilon\tau'(z)-\frac{z^k}{n^{1+k}}$ to determine:

$$\frac{dz}{d\epsilon}=\frac{-\tau'(z(n))}{T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}} \tag{15}$$

Finally, we need to get rid of the $\tau'(z(n))$ term. We will do this by multiplying and dividing by:

$$\frac{dz}{dn}=\frac{(1+k)z(n)^k/n^{k+2}}{T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}} \tag{16}$$

and then integrating by parts:

$$\begin{aligned}&\int_N \phi(n)\frac{z(n)T'(z(n))-T(z(n))}{(z(n)-T(z(n)))^2}\frac{u(c(n),z(n),n)f(n)}{T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}}\left(\frac{dz}{dn}\right)^{-1}\tau'(z(n))\frac{dz}{dn}dn=\\&-\int_N\frac{d}{dn}\left[\phi(n)\frac{z(n)T'(z(n))-T(z(n))}{(z(n)-T(z(n)))^2}\frac{u(c(n),z(n),n)}{(1+k)z(n)^k/n^{k+2}}f(n)\right]\tau(z(n))dn\end{aligned} \tag{17}$$

where we have used the assumption that $f(n)\to 0$ as $n\to\underline{n}$ or $n\to\overline{n}$ so as to ignore the boundary term that results from integration by parts.

Thus, the generalized welfare weights associated with the libertarian-inspired objective function in Equation (13) are given by the following (complicated) expression:[11]

$$\begin{aligned}g(n;T)=&-\left[\frac{z(n)}{z(n)-T(z(n))}\right]\phi(n)\;+\left[\frac{z(n)}{[z(n)-T(z(n))]^2}\right]\phi(n)\,u(c(n),z(n),n)\\&-\frac{d}{dn}\left[\phi(n)\frac{z(n)T'(z(n))-T(z(n))}{(z(n)-T(z(n)))^2}\frac{u(c(n),z(n),n)}{(1+k)z(n)^k/n^{k+2}}f(n)\right]\Big/f(n)\end{aligned} \tag{18}$$

Crucially, the generalized welfare weights given by Equation (18) are a function of $T(z)$, $T'(z)$, and $T''(z)$ (via the third $\frac{d}{dn}$ term). While $T(z)=c(z)-z$ and $T'(z)=1-\frac{z^k}{n^{1+k}}$ can be expressed as functions of c, z, and n, there *is no general expression* that relates c, z, and n to $T''(z)$ under every possible tax schedule. Hence, while the generalized welfare weights given by Equation (18) are consistent by Theorem 1, they cannot be expressed as $g(c,z,n)$ and therefore do not violate Proposition Sher. The dependence of $g(n;T)$ on $T''(z)$ in the third term of Equation (18) has a clear economic interpretation. Under the libertarian-utilitarian objective, the policymaker cares not only about the mechanical effect of a tax reform on individual consumption, but also about how behavioral responses to the reform affect the income-to-consumption ratio $z/(z-T(z))$ that determines the libertarian component of the welfare weight. Since behavioral responses depend on $T''(z)$ via the

[11] Note the final term in Equation (18) is divided by $f(n)$ so that the Gateaux derivative of W equals $\int_N g(n;T)f(n)dn$.

implicit function theorem (see Equation (15)), the welfare weights necessarily inherit this dependence. This is precisely the channel that is shut down when welfare weights are restricted to take the form $g(c, z, n)$: such weights can only capture the mechanical effect of a tax reform on individual welfare, not the effect of behavioral responses on the normative weight assigned to each individual.

4.2 GDP and Inequality Objective

Next, we provide an example wherein the generalized welfare weight for a particular type n depends on higher order derivatives of the tax schedule at their own income level as well as the tax schedule at other income levels (via its dependence on population level statistics). In particular, we will construct generalized welfare weights for an objective function that depends on aggregate income (GDP) and inequality (as measured by the Gini coefficient):

$$W(T) = \alpha GDP(T) - (1-\alpha)Gini(T)$$

where

$$GDP(T) = \int_N z(n)f(n)dn$$

and

$$Gini(T) = \frac{1}{GDP}\int_Z H(z)(1-H(z))dz = \frac{\int_N F(n)(1-F(n))\frac{dz}{dn}dn}{GDP(T)}$$

where $H(z)$ is the CDF of income z. Again, for simplicity, we will assume that $z(n)$ solves:

$$\max_z \quad z - T(z) - \frac{(z/n)^{1+k}}{1+k}$$

We show in Appendix B.1 that the generalized welfare weights associated with the objective function $W(T) = \alpha GDP(T) - (1-\alpha)Gini(T)$ are given by:

$$\begin{aligned} g(n;T) = {} & \alpha\left[\frac{f'(n)n + f(n)}{(1+k)(1-T'(z(n)))} + \frac{f(n)T''(z(n))}{(1-T'(z(n)))\left[T''(z(n)) + \frac{kz(n)^{k-1}}{n^{k+1}}\right]}\right] \\ & - (1-\alpha)\Bigg\{\frac{1}{GDP(T)}\frac{d}{dn}\left[A(n)\frac{n\left[T''(z(n)) + \frac{kz(n)^{k-1}}{n^{k+1}}\right]}{(1+k)(1-T'(z(n)))}\right] \\ & - \frac{Gini(T)}{GDP(T)}\left[\frac{f'(n)n + f(n)}{(1+k)(1-T'(z(n)))} + \frac{f(n)T''(z(n))}{(1-T'(z(n)))\left[T''(z(n)) + \frac{kz(n)^{k-1}}{n^{k+1}}\right]}\right]\Bigg\} \end{aligned} \tag{19}$$

where:

$$A(n) \equiv F(n)(1-F(n))\frac{1+k}{n}\left(\frac{T''(z(n))}{\left[T''(z(n)) + k\frac{z^{k-1}}{n^{k+1}}\right]^2} + \frac{\frac{z^k}{n^{k+1}}\left[T'''(z(n)) + k(k-1)\frac{z(n)^{k-2}}{n^{k+1}}\right]}{\left[T''(z(n)) + k\frac{z^{k-1}}{n^{k+1}}\right]^3}\right)$$

There are two key takeaways from Equation (19). First, as in our previous example with a libertarian-inspired objective function, the generalized welfare weights consistent with a GDP and Gini objective function depend on higher order derivatives of the tax schedule (in this case, the generalized welfare weights depend on T, T', T'', T''' and T'''' via $\frac{d}{dn}[A(n)]$). Second, the generalized welfare weight for type n depends on non-local

aspects of the tax schedule: the generalized welfare weight for type n depends on the level of both GDP and the Gini coefficient, which in turn depend on *the entire tax schedule*. Hence, these generalized welfare weights, which are consistent by Theorem 1, depend in a very complex way on the tax schedule via their dependence on higher order derivatives at the z chosen by a given n as well as via their dependence on population level statistics which, in turn, depend on the entire tax schedule.

The two examples in this section illustrate the key message of Proposition 1: the structurally utilitarian conclusion of Proposition Sher depends crucially on the assumption that generalized welfare weights are of the form $g(c, z, n)$. Once welfare weights are allowed to depend on the tax schedule in the richer ways required by non-utilitarian normative objectives — through higher derivatives of T or through population-level statistics — consistent non-utilitarian welfare weights do in fact exist.

5 Construction of Consistent Generalized Welfare Weights

Theorem 1 establishes that consistent generalized welfare weights must arise as the Gateaux derivative of some objective functional, and the examples in Section 4 exploited this fact by starting with an explicit objective function and computing the associated generalized welfare weights. However, the key innovation of Saez and Stantcheva (2016) is to evaluate the desirability of tax reforms *without* explicitly writing down an objective function. A natural question then, in the spirit of Saez and Stantcheva (2016), is whether one can construct consistent welfare weights *without* writing down an objective function — instead starting from directly elicited societal preferences over tax reforms. This Section shows that the answer is yes, provided one starts from "mechanical weights" that capture societal preferences over the direct distributional impacts of tax reforms, holding behavior fixed. First, let us define "mechanical weights" $m(\mathbf{n}; T)$:

Definition 7. *A set of "mechanical weights" $m(\mathbf{n}; T)$ depend on type $\mathbf{n}$ and the tax schedule T. A small tax reform in the direction of τ starting from T is desirable in absence of behavioral responses to the tax reform (i.e., holding $\mathbf{z}$ fixed for each $\mathbf{n}$ at the value chosen optimally under T) if and only if:*

$$\int_{\mathbf{N}} m(\mathbf{n}; T)\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})d\mathbf{n} > 0$$

Thus, $-m(\mathbf{n}; T)$ captures the value that society places on giving \$1 to individuals of type $\mathbf{n}$ given tax schedule T, holding their choices $\mathbf{z}$ fixed. While behavioral effects do not change the desirability of a small tax reform under utilitarian (or structurally utilitarian) welfare weights due to the envelope condition, behavioral responses *will* in general lead to first order impacts on the desirability of tax reforms under alternative generalized welfare weights. We saw this in the examples of Section 4: the generalized welfare weights that arise as the gradients of these particular objective functions depend

on the extent to which behavior changes with a tax reform. Hence, when we evaluate the desirability of a small tax reform by

$$\int_{\mathbf{N}} g(\mathbf{n};T)\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})d\mathbf{n}$$

or the desirability of a large tax reform by

$$\int_{\epsilon_1}^{\epsilon_2}\int_{\mathbf{N}} g(\mathbf{n};T^{(\epsilon)})T_{\epsilon}^{(\epsilon)}(\mathbf{z}(\mathbf{n}))f(\mathbf{n})d\mathbf{n}d\epsilon$$

then $g(\mathbf{n};T)$ implicitly *combines* social preferences over both the mechanical impacts of a tax reform and the behavioral impacts of a tax reform. However, Theorem 1 imposes a restriction on how these two components must be related: since consistent welfare weights must be the Gateaux derivative of some objective functional, the valuation of behavioral responses is not a free parameter but is instead *determined* to some degree by the valuation of mechanical transfers via the structure of the underlying objective functional. Proposition 2 formalizes this insight by showing, for any arbitrary set of mechanical weights of the form $m(c,z,n)$, how to construct consistent generalized welfare weights that are compatible with the given mechanical weights.[12]

Proposition 2. *Suppose Assumption 1 holds and that $f(n)\to 0$ as $n\to \underline{n}$ and as $n\to\overline{n}$. Next, suppose that one specifies mechanical weights $m(c,z,n)$. Let us denote $M(c,z,n)$ as $\int m(c,z,n)dc$ and suppose that $M(c,z,n)$ is twice differentiable in all arguments. If there are no income effects then:*

$$m(c(n),z(n),n)f(n)+\frac{d}{dn}\left[[M_c(1-T'(z(n)))+M_z]\xi(n)f(n)\left(\frac{dz}{dn}\right)^{-1}\right]$$

is a consistent set of generalized welfare weights (when comparing tax policies $T\in\mathcal{T}$). If there are income effects then:

$$m(c(n),z(n),n)f(n)-[M_c(1-T'(z(n)))+m_z]\eta(n)f(n)+\frac{d}{dn}\left[[M_c(1-T'(z(n)))+M_z]\xi(n)f(n)\left(\frac{dz}{dn}\right)^{-1}\right]$$

is a consistent set of generalized welfare weights.

Proof. Suppose that one starts with the objective function $W(T)=-\int_N M(c(n),z(n),n)f(n)dn$ with $M(c,z,n)=\int m(c,z,n)dc$ and $m(c,z,n)$ an arbitrary smooth function. Consider taking the Gateaux derivative of W starting at an initial tax schedule T:

$$DW_T(\tau)=\int_N M_c(c(n),z(n),n)\tau(z(n))f(n)dn-\int_N[M_c(1-T'(z(n))+M_z]\frac{dz}{d\epsilon}f(n)dn$$

Next, note that that by Equation (6):

$$\int_N[M_c(1-T'(z(n))+M_z]\frac{dz}{d\epsilon}f(n)dn=\int_N[M_c(1-T'(z(n)))+M_z][\eta(n)\tau(z(n))+\xi(n)\tau'(z(n))]f(n)dn$$

[12]For simplicity, Proposition 2 assumes that type $n\in N$ is a scalar and $z\in Z$ is a scalar representing income and that the mechanical weights take the functional form $m(c,z,n)$; in Section 6.4, Proposition 2 General will relax both of these assumptions.

Multiplying and dividing by $\frac{dz}{dn}$ and then integrating by parts (using the fact that $f(n) \to 0$ as $n \to \underline{n}, \overline{n}$) then yields:

$$\int_N [M_c(1 - T'(z(n))) + M_z]\xi(n)\tau'(z(n))f(n)dn$$
$$= -\int_N \frac{d}{dn}\left[[M_c(1 - T'(z(n))) + M_z]\xi(n)f(n)\left(\frac{dz}{dn}\right)^{-1}\right]\tau(z(n))dn$$

Thus:

$$DW_T(\tau) = \int_N \Bigg\{m(c(n), z(n), n)f(n) - [M_c(1 - T'(z(n))) + m_z]\eta(n)f(n)$$
$$+ \frac{d}{dn}\left[[M_c(1 - T'(z(n))) + M_z]\xi(n)f(n)\left(\frac{dz}{dn}\right)^{-1}\right]\Bigg\}\tau(z(n))dn$$

Hence:

$$m(c(n), z(n), n)f(n) - [M_c(1 - T'(z(n))) + m_z]\eta(n)f(n) + \frac{d}{dn}\left[[M_c(1 - T'(z(n))) + M_z]\xi(n)f(n)\left(\frac{dz}{dn}\right)^{-1}\right]$$

are consistent generalized welfare weights by Theorem 1 because they form the Gateaux derivative of the functional $W(T)$. □

Proposition 2 provides an affirmative answer to the question of whether one can construct consistent generalized welfare weights without explicitly writing down an objective function. The key insight is that specifying mechanical weights $m(c, z, n)$ is equivalent to specifying an objective function up to integration: the objective functional is $W(T) = -\int_N M(c(n), z(n), n)f(n)dn$ where $M(c, z, n) = \int m(c, z, n)dc$. The consistent generalized welfare weights then arise as the Gateaux derivative of this functional, and decompose naturally into two parts: the mechanical weights $m(c, z, n)$ themselves, which capture the direct social value of transferring resources to individuals of each type, and an additional term, $\frac{d}{dn}[[M_c(1 - T'(z(n))) + M_z]\xi(n)f(n)]$, that captures the social value of behavioral responses to the tax reform, which is *implied* by the mechanical weights $m(c, z, n)$ via the consistency requirement of Theorem 1.

It is worth emphasizing that Proposition 2 does not yield a *unique* set of consistent generalized welfare weights given a set of mechanical weights $m(c, z, n)$. The function $M(c, z, n) = \int m(c, z, n)dc$ is only determined up to the constant of integration and this constant of integration is allowed to vary freely with z and n; hence, different choices of the integration constant $K(z, n)$ yield different consistent generalized welfare weights. Intuitively, the mechanical weights pin down society's valuation of mechanical transfers but leave open the valuation of purely behavioral responses that have no direct mechanical counterpart. One can therefore interpret the family of consistent generalized welfare weights compatible with a given $m(c, z, n)$ as characterizing the *set* of all normative objectives that agree with society's valuation of mechanical transfers but differ in how

they value behavioral responses.

Finally, Proposition 2 has a direct empirical interpretation that connects our theoretical framework to the growing literature on survey-based measurement of redistributive preferences (Saez and Stantcheva, 2016; Capozza and Srinivasan, 2025). Surveys that ask respondents to evaluate the desirability of redistributive transfers to individuals of different types (which often implicitly hold behavior fixed by asking respondents to consider individual specific taxes/transfers) are naturally recovering mechanical weights in the sense of Definition 7. Proposition 2 shows that these empirically elicited mechanical weights can always be turned into a consistent set of generalized welfare weights by adding the implied valuation of behavioral responses. This provides a bridge from survey evidence on redistributive preferences to a consistent framework for evaluating large tax reforms, without requiring respondents to reason about the complex behavioral consequences of tax changes.

5.1 Example: Poverty Alleviation

Next, we illustrate Proposition 2 via an example by constructing consistent generalized welfare weights using "poverty alleviation" *mechanical welfare weights*. Saez and Stantcheva (2016) and Sher (2024) both present examples with welfare weights which incorporate a poverty alleviation desire. Suppose that a policymaker evaluates the desirability of a mechanical tax perturbation (i.e., ignoring behavioral responses) via the following mechanical welfare weights:

$$-\int_N \left(\frac{\bar{c}}{c}-1\right)\mathbb{1}[c\leq \bar{c}]\tau(z(n))f(n)dn \tag{20}$$

where $\bar{c}$ is the poverty line. The mechanical weights $-\left(\frac{\bar{c}}{c}-1\right)\mathbb{1}[c\leq\bar{c}]$ capture a poverty alleviation motive insofar as a tax cut is valued only for individuals below the poverty line $\bar{c}$, and the social value of a given tax cut is higher the further an individual's consumption falls below $\bar{c}$, reflecting the idea that the poorest individuals are the most deserving of transfers.[13] As in Section 4, we adopt the quasi-linear utility specification so that individuals solve:

$$\max_z \quad z - T(z) - \frac{(z/n)^{1+k}}{1+k}$$

To construct consistent generalized welfare weights via Proposition 2, we set $M(c,z,n) = \int m(c,z,n)\,dc$ by integrating the mechanical weights with respect to c:[14]

$$M(c,z,n) \equiv [\bar{c}\log(c) - c - (\bar{c}\log(\bar{c}) - \bar{c})]\mathbb{1}[c\leq\bar{c}]$$

The next step is to calculate the Gateaux derivative of:

$$\int_N [\bar{c}\log(c(n)) - c(n) - (\bar{c}\log(\bar{c}) - \bar{c})]\mathbb{1}[c(n)\leq\bar{c}]f(n)dn$$

[13] These are essentially the same "poverty alleviation" weights used in Sher (2024).

[14] Note that $M(c,z,n)$ is only pinned down up to an additive function of (z,n); here we subtract the constant $\bar{c}\log(\bar{c}) - \bar{c}$ so that $M(c,z,n)$ is continuous at the poverty line $\bar{c}$. One can still construct generalized welfare weights without this normalization using the more general Proposition 2 General.

Consider a tax perturbation from T to $T+\epsilon\tau$, recognizing that $c(n)$ depends on the tax schedule via its dependence $z(n)$. Differentiating under the integral (which is admissible because both $M(c,z,n)$ and $M_c(c,z,n)=\left(\frac{\bar{c}}{c}-1\right)\mathbb{1}[c\leq\bar{c}]$ are continuous), we get that:

$$\begin{aligned}&\frac{d}{d\epsilon}\int_N[\bar{c}\log(c(n))-c(n)-(\bar{c}\log(\bar{c})-\bar{c})]\mathbb{1}[c(n)\leq\bar{c}]f(n)dn\\&=-\int_N\left(\frac{\bar{c}}{c}-1\right)\mathbb{1}[c\leq\bar{c}]\tau(z(n))f(n)dn\\&+\int_N\left(\frac{\bar{c}}{c}-1\right)\mathbb{1}[c\leq\bar{c}](1-T'(z(n)))\frac{dz}{d\epsilon}f(n)dn\end{aligned}\tag{21}$$

Focusing on the second term in Equation (21) and assuming all individuals have a unique optimum under the smooth tax schedule T and that $f(n)\to 0$ as $n\to\underline{n}$ and as $n\to\overline{n}$:

$$\begin{aligned}&\int_N\left(\frac{\bar{c}}{c}-1\right)\mathbb{1}[c\leq\bar{c}](1-T'(z(n)))\frac{dz}{d\epsilon}f(n)dn\\&=\int_N\left(\frac{\bar{c}}{c}-1\right)\mathbb{1}[c\leq\bar{c}](1-T'(z(n)))\frac{-\tau'(z(n))}{T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}}f(n)dn\\&=\int_N\left(\frac{\bar{c}}{c}-1\right)\mathbb{1}[c\leq\bar{c}](1-T'(z(n)))\frac{-1}{T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}}f(n)\left(\frac{dz}{dn}\right)^{-1}\tau'(z(n))\frac{dz}{dn}dn\\&=\int_N\frac{d}{dn}\left[\left(\frac{\bar{c}}{c}-1\right)\mathbb{1}[c\leq\bar{c}](1-T'(z(n)))\frac{1}{T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}}f(n)\left(\frac{dz}{dn}\right)^{-1}\right]\tau(z(n))dn\\&=\int_N\frac{d}{dn}\left[\left(\frac{\bar{c}}{c}-1\right)\mathbb{1}[c\leq\bar{c}]\frac{n}{(1+k)}f(n)\right]\tau(z(n))dn\\&=\int_N\left[\left(\frac{\bar{c}}{c}-1\right)\mathbb{1}[c\leq\bar{c}]\frac{f(n)+nf'(n)}{(1+k)}-\frac{\bar{c}}{c^2}\mathbb{1}[c\leq\bar{c}]\frac{nf(n)z(n)^k/n^{k+2}}{T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}}\right]\tau(z(n))dn\end{aligned}\tag{22}$$

where $\frac{dz}{d\epsilon}$ is given by Equation (15) and the expression for $\frac{dz}{dn}$ is given by Equation (16).[15] Thus, if a policymaker evaluates the mechanical desirability of a small tax perturbation via Equation (20), then by Proposition 2 one can evaluate large policy changes via the *consistent* generalized welfare weights:

$$g(n;T)=\left\{-\left(\frac{\bar{c}}{c}-1\right)+\left[\left(\frac{\bar{c}}{c}-1\right)\frac{f(n)+nf'(n)}{(1+k)}-\frac{\bar{c}}{c^2}\frac{nf(n)z(n)^k/n^{k+2}}{T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}}\right]\right\}\mathbb{1}[c\leq\bar{c}]\tag{23}$$

Equation (23) illustrates the general idea of Proposition 2 in a concrete setting. The first term, $\left(\frac{\bar{c}}{c}-1\right)\mathbb{1}[c\leq\bar{c}]$, is simply the (negative) mechanical weight: it captures the direct social value of transferring resources to individuals below the poverty line in absence of behavioral responses. The remaining two terms capture the implied social value of behavioral responses to a tax reform given the specified mechanical weights. More specifically, these two terms capture the fact that a tax reform induces individuals below

[15] We have also used the assumption that $f(n)\to 0$ as $n\to\underline{n}$ or $n\to\overline{n}$ so that the boundary term arising from integration by parts equals 0.

the poverty line to adjust their labor supply, which in turn changes their consumption and therefore alters the social value of transferring them money. Taken together, these terms show that a policymaker who cares about poverty alleviation and wishes to evaluate large tax reforms in a consistent manner must account not only for the direct transfers to the poor but also for the induced behavioral responses that shift individuals' consumption levels.[16]

Finally, it is instructive to compare Equation (23) with the poverty alleviation weights discussed in Sher (2024). Sher (2024) shows that $g(c,z,n) = \left(\frac{\bar{c}}{c} - 1\right) \mathbb{1}[c \leq \bar{c}]$, which correspond to our *mechanical* weights, are inconsistent generalized welfare weights.[17] Equation (23) can therefore be interpreted as the "corrected" version of these poverty alleviation weights that restores consistency by incorporating the implied valuation of behavioral responses as prescribed by Proposition 2.

6 Richer Environments

As discussed in Remark 2, Theorem 1 is very general and can therefore accommodate many interesting and realistic components: multidimesional tax schedules and heterogeneity, optimization errors, and non-smooth behavioral responses such as bunching or jumping between optima. However, in contrast, Propositions Sher and 2 were both derived under a set of simplifying assumptions: tax schedules are unidimensional, individuals respond smoothly to tax reforms, general equilibrium effects are absent, and (for Proposition 2) mechanical weights take the specific functional form $m(c,z,n)$. This section shows that these assumptions can all be relaxed to varying degrees. We proceed in four steps. First, we extend Propositions Sher and 2 to settings with multidimensional tax schedules and heterogeneity. Second, we show that Proposition 2 can be extended when agents make optimization errors and have distinct "decision" and "experience"

[16] The careful reader will recall that Proposition 2 does not pin down a *unique* set of consistent generalized welfare weights given mechanical welfare weights because the function $M(c,z,n)$ is only pinned down up to a function $v(z,n)$ (intuitively, the desirability of behavioral changes is not pinned down exactly if one only knows mechanical welfare weights). Hence, one can use nearly identical logic to show that *any* $g(n;T)$ equal to Equation (23) plus $\frac{d}{dn}\left[v_z(z(n),n)\frac{n}{(1+k)(1-T'(z(n)))}\right]$ for any function $v(z,n)$ are a consistent set of generalized welfare weights.

[17] To see this via Theorem 1, one can perturb Equation (20) in the direction of $T + \epsilon\gamma$, noting that z varies with ϵ to yield that the Gateaux derivative of Equation (20) equals:

$$\int_N \left\{ \frac{1}{\bar{c}} \mathbb{1}[c \leq \bar{c}] \gamma(z(n)) \tau(z(n)) + \frac{1}{\bar{c}} \mathbb{1}[c \leq \bar{c}][1 - T'(z(n))] \tau(z(n)) \frac{-\gamma'(z(n))}{T''(z(n)) + \frac{kz(n)^{k-1}}{n^{k+1}}} \right.$$
$$\left. + \left(\frac{\bar{c}}{c} - 1\right) \mathbb{1}[c \leq \bar{c}] \tau'(z(n)) \frac{-\gamma'(z(n))}{T''(z(n)) + \frac{kz(n)^{k-1}}{n^{k+1}}} \right\} f(n) dn$$

To satisfy the symmetry requirement of Theorem 1, we require that $\frac{\frac{1}{\bar{c}} \mathbb{1}[c \leq \bar{c}][1 - T'(z(n))]}{T''(z(n)) + \frac{kz(n)^{k-1}}{n^{k+1}}} [\tau'(z(n))\gamma(z(n)) - \tau(z(n))\gamma'(z(n))] = 0$. But this only holds (by the fundamental theorem of the calculus of variations) if $[1 - T'(z(n))] = 0 \ \forall c, z, n$, which clearly does not hold generically.

utility functions; Proposition Sher can also be applied to this setting, however, we show that any consistent generalized welfare weights $g(c, z, n)$ must be must be structurally utilitarian in *decision utility* rather than in the usual experience utility. Third, we discuss how Proposition 2 extends to settings with general equilibrium effects of taxation on wages and prices. Finally, we present a fully general version of Proposition 2 — Proposition 2 General — that simultaneously relaxes all of the smoothness and functional form assumptions maintained in the earlier sections, at the cost of a higher level of abstraction.

6.1 Multidimensional Tax Schedules and Heterogeneity

Let us begin by discussing how our main results extends when the tax schedule depends on several choice variables $\mathbf{z} = (z_1, z_2, ..., z_J)$ and individuals differ in terms of $\mathbf{n} = (n_1, n_2, ..., n_K)$. First, note that the necessary and sufficient conditions for consistency of generalized welfare weights given in Theorem 1 were already written allowing for multidimensional heterogeneity and multidimensional tax schedules. Second, let us now illustrate how Proposition Sher from Sher (2024) extends to the multidimensional case. We will need a multidimensional analogue of Assumption 1 that incorporates standard smoothness assumptions from the multidimensional optimal taxation literature (Golosov et al., 2014; Spiritus et al., 2022; Bergstrom and Dodds, 2026):

Assumption 2. *Suppose that $\mathcal{T}$ is a convex set of smooth tax schedules that is open within the set of compactly supported smooth functions. Suppose that for all $T \in \mathcal{T}$ that all individuals have a unique optimum and second order conditions are satisfied strictly (i.e., utility is strictly concave under their optimal choice of $\mathbf{z}$) and that the mapping $\mathbf{n} \mapsto \mathbf{z}$ is bijective and differentiable.*

Under Assumption 2, we can prove a generalization of the main result from Sher (2024) when tax schedules are multidimensional:

Proposition Sher Multidimensional. *Suppose Assumption 2 holds. Generalized welfare weights $g(c, \mathbf{z}, \mathbf{n})$ are consistent (when comparing tax policies $T \in \mathcal{T}$) if and only if they are equal to $g(c, \mathbf{z}, \mathbf{n}) = u_c(c, \mathbf{z}, \mathbf{n}) \cdot \psi(u(c, \mathbf{z}, \mathbf{n}), \mathbf{n})$ for some function ψ.*

Proof. The proof is qualitatively exactly the same as the unidimensional case, see Appendix B.3 for details. □

Proposition 2 can also be modified to hold in multidimensional contexts:

Proposition 2 Multidimensional. *Suppose Assumption 2 holds and that $f(\mathbf{n}) \to 0$ as $\mathbf{n} \to \partial\mathbf{N}$ where $\partial\mathbf{N}$ denotes the boundary of $\mathbf{N}$. Next, suppose that one specifies mechanical weights $m(c, \mathbf{z}, \mathbf{n})$. Let us denote $M(c, \mathbf{z}, \mathbf{n})$ as $\int m(c, \mathbf{z}, \mathbf{n})dc$. If there are no income effects then:*

$$m(c(\mathbf{n}), \mathbf{z}(\mathbf{n}), \mathbf{n})f(\mathbf{n}) + \nabla_{\mathbf{n}} \cdot \left[[M_c(1 - T'(\mathbf{z}(\mathbf{n}))) + M_{\mathbf{z}}]\mathbf{X}(\mathbf{n})f(\mathbf{n}) \left(\frac{d\mathbf{z}}{d\mathbf{n}} \right)^{-1} \right]$$

is a consistent set of generalized welfare weights where $\nabla_{\mathbf{n}} \cdot [[M_c(1 - T'(\mathbf{z}(\mathbf{n}))) + M_{\mathbf{z}}]\xi(\mathbf{n})f(\mathbf{n})]$ *captures the value of behavioral changes pinned down by the valuation of mechanical changes given by* $m(c, \mathbf{z}, \mathbf{n})$*. If there are income effects then:*

$$m(c(\mathbf{n}), \mathbf{z}(\mathbf{n}), \mathbf{n})f(\mathbf{n}) - [M_c(1 - T'(\mathbf{z}(\mathbf{n}))) + m_{\mathbf{z}}]\vec{\eta}(\mathbf{n})f(\mathbf{n}) + \nabla_{\mathbf{n}} \cdot \left[[M_c(1 - T'(\mathbf{z}(\mathbf{n}))) + M_{\mathbf{z}}]\mathbf{X}(\mathbf{n})f(\mathbf{n}) \left(\frac{d\mathbf{z}}{d\mathbf{n}} \right)^{-1} \right]$$

is a consistent set of generalized welfare weights.

Proof. The proof is qualitatively exactly the same as the unidimensional case, see Appendix B.4 for details. □

Propositions Sher Multidimensional and 2 Multidimensional confirm that Propositions Sher and 2 are not artifacts of the unidimensional setting. Proposition Sher Multidimensional shows that the structurally utilitarian characterization of Proposition Sher extends without modification to settings with multidimensional tax schedules: consistent welfare weights of the form $g(c, \mathbf{z}, \mathbf{n})$ must still take the form $u_c \cdot \psi(u, \mathbf{n})$, regardless of the dimensionality of $\mathbf{z}$. Proposition 2 Multidimensional similarly extends the constructive result of Proposition 2: given any mechanical weights $m(c, \mathbf{z}, \mathbf{n})$, one can construct a consistent set of generalized welfare weights by augmenting the mechanical weights with a term that captures the implied social value of behavioral responses.

6.2 Behavioral Agents

Over the past decade, public economists have been increasingly interested in exploring how policy analysis changes when accounting for the fact that households may be irrational and make mistakes when making decisions (see Bernheim and Taubinsky (2018) for a review). One way to incorporate this idea is to allow individuals to have a "decision utility" function $u(c, \mathbf{z}, \mathbf{n})$ which governs their decision making and a separate "experience utility" function $\hat{u}(c, \mathbf{z}, \mathbf{n})$ which governs their well-being. Hence, by optimizing $u(c, \mathbf{z}, \mathbf{n})$, agents are acting sub-optimally relative to their own well-being $\hat{u}(c, \mathbf{z}, \mathbf{n})$. Next, let us briefly discuss how our results on consistency of generalized welfare weights extend when agents make optimization errors in this sense.

First, note that in Theorem 1, the concept of "utility" is never even mentioned explicitly: utility only enters implicitly via individuals choices $\mathbf{z}$ which vary with T by maximizing their utility function. Thus, Theorem 1 carries over immediately in the presence of optimization errors, recognizing that $\mathbf{z}(\mathbf{n})$ is chosen to maximize $u(c, \mathbf{z}, \mathbf{n})$ (but this does not impact any of the arguments in any way). Similarly, Proposition 2 also holds without modification in the presence of behavioral agents as long as one recognizes that the behavioral responses ξ and η are derived from the decision utility function $u(c, z, n)$. However, the most interesting implication of extending our theory to behavioral agents concerns Proposition Sher. It turns out that the entirety of the

argument for Proposition Sher goes through as long as $u(c, z, n)$ represents *decision utility*. In other words, generalized welfare weights of the form $g(c, z, n)$ are consistent if and only if $g(c, z, n) = u_c(c, z, n)\psi(u(c, z, n), n)$ for some function ψ: thus, when agents make optimization errors, generalized welfare weights $g(c, z, n)$ are only consistent if they are "structurally utilitarian" in decision utility, *not* the experience utility that governs well-being. This is normatively unappealing: a policymaker who cares about household well-being would naturally want to weight individuals according to their experience utility, not their decision utility. The fact that consistent weights $g(c, z, n)$ cannot accommodate this reflects a fundamental limitation of the functional form restriction identified throughout the paper: an objective function that involves experience utility $\hat{u}(c, z, n)$ will generate generalized welfare weights that depend on $T''(z)$; this is because the individual first-order condition is not satisfied under $\hat{u}(c, z, n)$, so that generalized welfare weights will depend on the size of behavioral responses, which in turn depend on $T''(z)$ due to the implicit function theorem.[18] In this sense, the behavioral extension reinforces the paper's broader message: the functional form restriction $g(c, z, n)$ is not merely a technical limitation but encodes a substantive — and often undesirable — normative restriction on how generalized welfare weights can depend on individual circumstances.

6.3 General Equilibrium

Next, we will discuss how consistency of generalized welfare weights changes when we allow for general equilibrium effects of taxes on wages/prices. We augment the model from Section 2.1 to allow individual utility to also depend on a vector, $\mathbf{w}$, of "general equilibrium" parameters: $\mathbf{w}$ might consist of prices, wages, externalities, or other quantities that are impacted in some way by aggregate individual decisions (and hence depend on taxes). Thus, individuals maximize the following utility function:

$$\begin{aligned} U(n; T, \mathbf{w}) = \max_{\mathbf{z}} \; & u\left(c, \mathbf{z}; \mathbf{n}, \mathbf{w}\right) \\ \text{s.t. } & c = y(\mathbf{z}, \mathbf{w}) - T(\mathbf{z}) \end{aligned} \tag{24}$$

First, Theorem 1 again goes through essentially without modification: $g(\mathbf{n}; T)$ just now represents the social value of increasing taxes on type $\mathbf{n}$ inclusive of any general equilibrium effects that are generated by this tax increase.[19] However, proving an extension of Sher's result, Proposition Sher, with general equilibrium effects proved impossible. Intuitively, with general equilibrium effects, a tax reform at income level z affects wages and prices faced by all individuals, so the welfare weight assigned to type n depends on the tax schedule at every income level, not just at $z(n)$. This non-locality means that the PDE system characterizing consistent weights of the form $g(c, z, n)$ is intractable without

[18] We illustrate this with an example in Appendix B.5.

[19] Note, calculating Gateaux derivatives to verify the conditions of Theorem 1 is more involved with general equilibrium effects; see Appendix B.7 for an example.

additional structure on how equilibrium objects $\mathbf{w}$ respond to tax changes. On the other hand, we *can* extend Proposition 2 to allow for general equilibrium effects, but we will combine this extension with a very broad generalization of Proposition 2 in the following subsection.

6.4 Relaxing Assumption 1

Finally, while Theorem 1 requires no assumptions about individuals having a unique optimum or responding smoothly in response to tax reforms (or even about the choice sets of $\mathbf{z}$ available to individuals), both Proposition Sher and Proposition 2 both relied on this smoothness via dependence on Assumption 1. Moreover, Propositions 2 and 2 Multidimensional also assume that the type density goes to zero on the boundary of the type space and that mechanical weights take the form $m(c, \mathbf{z}, \mathbf{n})$. We illustrate in Appendix B.9 that Proposition Sher holds when making comparisons across the entire set of continuous tax schedules under a single crossing condition: the only consistent welfare weights of the form $g(c, z, n)$ are structurally utilitarian across the entire set of continuous tax schedules, including those that generate bunching or people with multiple optima. The key intuition is that any (c, z, n) that arises optimally under an arbitrary continuous tax schedule also arises optimally under some linear tax schedule. Since the symmetry condition of Theorem 1 must hold for all continuous tax schedules — and in particular for all linear schedules — it is sufficient to derive the structurally utilitarian restriction on $g(c, z, n)$ from linear tax schedules alone, and this restriction then applies to all (c, z, n) that arise under any continuous tax schedule.

Finally, we will now illustrate how smoothness assumptions and boundary assumptions can be relaxed in Proposition 2 (with the cost of this relaxation being a very high level of abstraction). To do this, we need to introduce some new notation. Recall that in Definition 7 we defined a set of mechanical weights $m(n; T)$ and in Proposition 2 we assumed that mechanical weights took the specific functional form $m(c, z, n)$. Now, we will expand on this notation to define mechanical weights $m(\mathbf{n}; T, \mathbf{z}(\cdot, T), \mathbf{w}(T))$ which vary with $\mathbf{n}$ and depend on (1) the tax schedule, (2) the choice function $\mathbf{z}(\mathbf{n}; T)$, and (3) general equilibrium objects $\mathbf{w}(T)$.[20] More precisely, we will now illustrate how we can construct consistent generalized welfare weights by specifying (arbitrary) mechanical weights and then inferring the generalized welfare weights that account for both the mechanical and behavioral effects of tax reforms in a consistent manner:

Proposition 2 General. *Suppose that one specifies mechanical weights* $m(\mathbf{n}; T, \mathbf{z}(\cdot, T), \mathbf{w}(T))$ *over a convex set of tax policies* $\mathcal{T}$*. Suppose that these mechanical weights satisfy the sym-*

[20] Note, the mechanical weights $m(\mathbf{n}; T, \mathbf{z}(\cdot, T), \mathbf{w}(T))$ for a particular type $\mathbf{n}$ can depend on the entire choice function $\mathbf{z}(\cdot, T)$ rather than on the single $\mathbf{z}(\mathbf{n}, T)$ chosen by type $\mathbf{n}$.

metry condition:

$$\frac{d}{dv}\left\{\int_{\mathbf{N}} m(\mathbf{n}; T+v\tau, \mathbf{z}(\cdot,T), \mathbf{w}(T))\gamma(\mathbf{z}(\mathbf{n}))\,d\mathbf{n}\right\}\Big|_0 = \frac{d}{dv}\left\{\int_{\mathbf{N}} m(\mathbf{n}; T+v\gamma, \mathbf{z}(\cdot,T), \mathbf{w}(T))\tau(\mathbf{z}(\mathbf{n}))\,d\mathbf{n}\right\}\Big|_0 \tag{25}$$

Fix a T_0 *and define for any* $T, \hat{T} \in \mathcal{T}$*:*

$$M(\mathbf{n}; T, \mathbf{z}(\cdot,\hat{T}), \mathbf{w}(\hat{T})) = \int_0^1 m(\mathbf{n}; T_0 + v(T-T_0), \mathbf{z}(\cdot,\hat{T}), \mathbf{w}(\hat{T}))(T(\mathbf{z}(\mathbf{n})) - T_0(\mathbf{z}(\mathbf{n})))dv$$

Suppose that the Gateaux derivative of M *with respect to* $\hat{T}$*, denoted* $DM_{\hat{T}}$*, exists* $\forall \hat{T} = T \in \mathcal{T}$*. Then evaluating the desirability of large tax reforms from* $T^{(\epsilon_1)}(\mathbf{z})$ *to* $T^{(\epsilon_2)}(\mathbf{z})$ *via:*

$$\begin{aligned}&\Delta W(T^{(\epsilon_2)}(\mathbf{z}), T^{(\epsilon_1)}(\mathbf{z})) \equiv \\ &\int_{\epsilon_1}^{\epsilon_2}\left[\int_{\mathbf{N}} m(\mathbf{n}; T^{(\epsilon)}, \mathbf{z}(\cdot,T^{(\epsilon)}), \mathbf{w}(T^{(\epsilon)}))f(\mathbf{n})T_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n}))d\mathbf{n} + DM_{\hat{T}}[T_\epsilon^{(\epsilon)}]\big|_{\hat{T}=T=T^{(\epsilon)}}\right]d\epsilon\end{aligned} \tag{26}$$

is a consistent way to evaluate large tax reforms (in the sense of Definition 5).

Proof. See Appendix A.4. □

Proposition 2 General is the most general version of the constructive result in this paper, simultaneously relaxing the smoothness assumptions of Assumption 1, the boundary conditions on $f(\mathbf{n})$, the functional form restriction $m(c, \mathbf{z}, \mathbf{n})$ on mechanical weights, the smoothness assumptions on $M(c, z, n)$, as well as the partial equilibrium restriction. The cost of this generality is a higher level of abstraction: the consistent welfare weights in Equation (26) are expressed as the sum of a mechanical component

$$\int_{\mathbf{N}} m(\mathbf{n}; T^{(\epsilon)}, \mathbf{z}(\cdot,T^{(\epsilon)}), \mathbf{w}(T^{(\epsilon)}))f(\mathbf{n})T_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n}))d\mathbf{n}$$

and a correction term $DM_{\hat{T}}[T_\epsilon^{(\epsilon)}]\big|_{\hat{T}=T=T^{(\epsilon)}}$ that captures the implied social value of behavioral and general equilibrium responses. The latter term is expressed quite abstractly rather than as an explicit formula involving derivatives of the tax schedule, reflecting the fact that in the fully general setting, behavioral responses may be non-smooth or non-local and cannot be reduced to a simple expression.[21]

Taken together, the results of this section establish that the paper's main conclusions — the characterization of consistent generalized welfare weights in Theorem 1, the structurally utilitarian restriction on weights of the form $g(c, \mathbf{z}, \mathbf{n})$, and the constructive recipe for consistent welfare weights in Proposition 2 — are robust to a wide range of departures

[21] The symmetry condition on mechanical weights required by Proposition 2 General is the analogue of the integrability condition discussed in Section 5: it ensures that the mechanical weights are themselves consistent (as a description of societal preferences over reforms which do not induce behavioral responses), which is a necessary condition for them to be completable into a consistent set of generalized welfare weights. This condition is automatically satisfied when mechanical weights take the functional form $m(c, \mathbf{z}, \mathbf{n})$ assumed in Propositions 2 and 2 Multidimensional, but must be verified separately in the fully general case.

from the baseline model. The generality of Theorem 1 is particularly notable: because it characterizes consistency purely in terms of the functional form of generalized welfare weights, without any reference to the structural primitives of individual optimization, it applies equally across all of the environments considered in this section. The constructive results are similarly robust, with Proposition 2 General providing a fully general recipe that applies even in the presence of bunching, optimization errors, and general equilibrium effects. The one result that requires more structure is the structurally utilitarian characterization of Proposition Sher, which relies on the functional form restriction $g(c, \mathbf{z}, \mathbf{n})$; as argued throughout the paper, however, this is a genuine restriction that rules out economically interesting welfare weights rather than just an innocuous technical assumption.

For the interested reader, we illustrate several example applications of Proposition 2 General in Appendices B.6, B.7, and B.8. We illustrate how to construct consistent generalized welfare weights when (1) mechanical weights are discontinuous as in the poverty alleviation example from Saez and Stantcheva (2016) (where individuals below the poverty line have a mechanical weight of 1 and those above have a mechanical weight of 0), (2) there are GE wage effects as in Sachs et al. (2020) or Bergstrom and Dodds (2026), and (3) tax schedules have kinks (inducing bunching) and individuals have multiple optima.

7 Conclusion

This paper has examined the consistency of generalized social marginal welfare weights as a tool for evaluating tax reforms and ranking tax policies. Our first main result, Theorem 1, provides a complete characterization of consistent generalized welfare weights: a set of welfare weights generates consistent global comparisons of tax policies if and only if it arises as the Gateaux derivative of some objective functional, or equivalently, if and only if its Gateaux variation is symmetric. Our second main result, Proposition 1, illustrates that the inconsistency result of Sher (2024) is not a fundamental limitation of the Saez and Stantcheva (2016) generalized welfare weight approach: consistent non-utilitarian welfare weights do exist once we broaden the definition of generalized welfare weights to allow them to depend on the tax schedule in the richer ways required by non-utilitarian normative objectives (i.e., through higher derivatives of the tax schedule or through population-level statistics). Our third main result, Proposition 2 and its generalization, Proposition 2 General, further show that consistent welfare weights can always be constructed from an arbitrary set of mechanical weights that capture societal preferences over the direct distributional impacts of tax reforms, thereby providing a constructive recipe for implementing the generalized welfare weight approach in a consistent manner.

Looking forward, our results have several implications for future work. First, the generalized welfare weight approach of Saez and Stantcheva (2016) can be used to eval-

uate large tax reforms while incorporating non-utilitarian social preferences, provided that the generalized welfare weights are chosen to be consistent via Theorem 1. Second, there is value in further exploring the quantitative implications of using consistent non-utilitarian generalized welfare weights for the shape of the optimal tax schedule: our examples illustrate that consistent generalized welfare weights incorporating libertarian or inequality-aversion motives can differ substantially from their structurally utilitarian counterparts, suggesting that non-utilitarian generalized welfare weights may yield optimal tax schedules that differ considerably from those derived under standard utilitarian welfare weights.

References

Kenneth J. Arrow. *Social Choice and Individual Values*. John Wiley & Sons, New York, 1951.

Katy Bergstrom and William Dodds. Optimal taxation with multiple dimensions of heterogeneity. *Journal of Public Economics*, 200:104442, 2021. ISSN 0047-2727. doi: https://doi.org/10.1016/j.jpubeco.2021.104442. URL https://www.sciencedirect.com/science/article/pii/S0047272721000785.

Katy Bergstrom and William Dodds. A general theory of inverse welfare functions. *Theoretical Economics (Forthcoming)*, 2026.

B. Douglas Bernheim and Dmitry Taubinsky. Behavioral public economics. In B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, editors, *Handbook of Behavioral Economics - Foundations and Applications 1*, volume 1, pages 381–516. North-Holland, 2018. doi: 10.1016/bs.hesbe.2018.07.007.

Joël Blot. The rank theorem in infinite dimension. *Nonlinear Analysis: Theory, Methods & Applications*, 10(3):245–250, 1986. doi: 10.1016/0362-546X(86)90085-4.

Francesco Capozza and Krishna Srinivasan. Who should get money? estimating welfare weights in the us. *URPP Equality of Opportunity Discussion Paper Series*, (50), 2025.

Marc Fleurbaey and François Maniquet. Optimal income taxation theory and principles of fairness. *Journal of Economic Literature*, 56(3):1029–1079, 2018. doi: 10.1257/jel.20171350.

Mikhail Golosov, Aleh Tsyvinski, and Nicolas Werquin. A variational approach to the analysis of tax systems. *NBER Working Paper 20780*, 48, 2014. URL https://www.nber.org/papers/w20780.

Louis Kaplow and Steven Shavell. Any non-welfarist method of policy assessment violates the pareto principle. *Journal of Political Economy*, 109(2):281–286, 2001. doi: 10.1086/319553.

Andreas Kriegl and Peter W. Michor. *The Convenient Setting of Global Analysis*, volume 53 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1997.

James Mirrlees. An exploration in the theory of optimal income taxation. *Review of Eco-*

nomic Studies, 38:175–208, 1971. URL http://aida.econ.yale.edu/~dirkb/teach/pdf/mirrlees/1971optimaltaxation.pdf.

Walter Rudin. *Real and complex analysis*. McGraw-Hill Book Company, 1974.

Dominik Sachs, Aleh Tsyvinski, and Nicolas Werquin. Nonlinear tax incidence and optimal taxation in general equilibrium. *Econometrica*, 88(2):469–493, 2020. doi: https://doi.org/10.3982/ECTA14681. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA14681.

Emmanuel Saez. Using elasticities to derive optimal income tax rates. *Review of Economic Studies*, 68:205–229, 2001. URL http://eml.berkeley.edu/~saez/derive.pdf.

Emmanuel Saez and Stefanie Stantcheva. Generalized social marginal welfare weights for optimal tax theory. *American Economic Review*, 106(1):24–45, 2016.

Itai Sher. Generalized social marginal welfare weights imply inconsistent comparisons of tax policies. *American Economic Review*, 114(11):3551–3577, 2024. doi: 10.1257/aer.20211025.

Kevin Spiritus, Etienne Lehmann, Sander Renes, and Floris Zoutman. Optimal taxation with multiple incomes and types. *SSRN Electronic Journal*, 2022.

Matthew Weinzierl. The promise of positive optimal taxation: Normative diversity and a role for equal sacrifice. *Journal of Public Economics*, 118:128–142, 2014. doi: 10.1016/j.jpubeco.2014.07.002.

Matthew Weinzierl. Popular acceptance of inequality due to innate brute luck and support for classical benefit-based taxation. *Journal of Public Economics*, 155:54–63, 2017. doi: 10.1016/j.jpubeco.2017.09.004.

A Appendix: Proofs

A.1 Proof of Lemma 1

Suppose that $g(\mathbf{n}; T)$ is path independent. Consider any path between schedules $T^{(0)}(\mathbf{z})$, $T^{(\epsilon_1)}(\mathbf{z})$, $T^{(\epsilon_2)}(\mathbf{z})$, and then back to $T^{(1)}(\mathbf{z}) = T^{(0)}(\mathbf{z})$ such that:

$$\int_0^{\epsilon_1} \int_{\mathbf{N}} g(\mathbf{n}; T^{(\epsilon)})\, T_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n}))\, f(\mathbf{n})\, dn\, d\epsilon > 0 \tag{27}$$

$$\int_{\epsilon_1}^{\epsilon_2} \int_{\mathbf{N}} g(\mathbf{n}; T^{(\epsilon)})\, T_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n}))\, f(\mathbf{n})\, dn\, d\epsilon > 0 \tag{28}$$

By additivity of integrals then:

$$\int_0^{\epsilon_2} \int_{\mathbf{N}} g(\mathbf{n}; T^{(\epsilon)})\, T_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n}))\, f(\mathbf{n})\, dn\, d\epsilon > 0$$

But then by path independence because $T^{(1)}(\mathbf{z}) = T^{(0)}(\mathbf{z})$:

$$\int_1^{\epsilon_2} \int_{\mathbf{N}} g(\mathbf{n}; T^{(\epsilon)})\, T_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n}))\, f(\mathbf{n})\, dn\, d\epsilon > 0$$

which implies:

$$\int_{\epsilon_2}^1 \int_{\mathbf{N}} g(\mathbf{n}; T^{(\epsilon)})\, T_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n}))\, f(\mathbf{n})\, dn\, d\epsilon < 0$$

So that $g(\mathbf{n}; T)$ are consistent.

Next, suppose that $g(\mathbf{n}; T)$ is not path independent so that there exists $T^{(\epsilon)}(\mathbf{z})$ and $\hat{T}^{(\epsilon)}(\mathbf{z})$ with $T^{(0)}(\mathbf{z}) = \hat{T}^{(0)}(\mathbf{z})$ and $T^{(1)}(\mathbf{z}) = \hat{T}^{(1)}(\mathbf{z})$ such that WLOG:

$$\int_0^1 \int_{\mathbf{N}} g(\mathbf{n}; T^{(\epsilon)}) T_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n})) f(\mathbf{n}) dn d\epsilon > \int_0^1 \int_{\mathbf{N}} g(\mathbf{n}; \hat{T}^{(\epsilon)}) \hat{T}_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n})) f(\mathbf{n}) dn d\epsilon \tag{29}$$

Consider $\tilde{T}^{(\epsilon)}(\mathbf{z}) = T^{(2\epsilon)}(\mathbf{z}) \mathbb{1}[\epsilon \leq 1/2] + \hat{T}^{(2-2\epsilon)}(\mathbf{z}) \mathbb{1}[\epsilon > 1/2]$, which is just a closed loop running from $T^{(0)}(\mathbf{z})$ to $T^{(1)}(\mathbf{z})$ back to $T^{(0)}(\mathbf{z})$ where the first half of the loop is given by $T^{(\epsilon)}(\mathbf{z})$ where ϵ runs from 0 to 1 and the second half of the loop is given by $\hat{T}^{(\epsilon)}(\mathbf{z})$ where ϵ runs from 1 to 0. Noting that by Equation (29):

$$\begin{aligned} \delta &\equiv \int_0^1 \int_{\mathbf{N}} g(\mathbf{n}; \tilde{T}^{(\epsilon)}) \tilde{T}_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n})) f(\mathbf{n}) dn d\epsilon \\ &= \int_0^1 \int_{\mathbf{N}} g(\mathbf{n}; T^{(\epsilon)}) T_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n})) f(\mathbf{n}) dn d\epsilon - \int_0^1 \int_{\mathbf{N}} g(\mathbf{n}; \hat{T}^{(\epsilon)}) \hat{T}_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n})) f(\mathbf{n}) dn d\epsilon > 0 \end{aligned} \tag{30}$$

But by the intermediate value theorem, there must be some $\epsilon_1 \in (0, 1)$ where

$$\int_0^{\epsilon_1} \int_{\mathbf{N}} g(\mathbf{n}; \tilde{T}^{(\epsilon)}) \tilde{T}_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n})) f(\mathbf{n}) dn d\epsilon = \delta/2$$

and some $\epsilon_2 \in (\epsilon_1, 1)$ (also by the intermediate value theorem) s.t.

$$\int_0^{\epsilon_2} \int_{\mathbf{N}} g(\mathbf{n}; \tilde{T}^{(\epsilon)}) \tilde{T}_\epsilon^{(\epsilon)}(\mathbf{z}(\mathbf{n})) f(\mathbf{n}) dn d\epsilon = \frac{3}{4}\delta$$

which implies that $g(\mathbf{n}; T)$ is not consistent.

A.2 Proof of Theorem 1

Proof. We prove these statements from easiest to hardest.

(2) $\implies$ (1)

Suppose $DW_T[\tau] \equiv \int_{\mathbf{N}} g(\mathbf{n}; T)\tau(\mathbf{z}(\mathbf{n}))d\mathbf{n}$ is the Gateaux derivative of a functional W. Then for any path $T^{(\epsilon)}(\mathbf{z})$ parametrized by ϵ between $T^{(0)}(\mathbf{z})$ and $T^{(1)}(\mathbf{z})$:

$$\int_0^1 DW_{T^{(\epsilon)}(\mathbf{z})}[T_\epsilon(\mathbf{z}(\mathbf{n}); \epsilon)]d\epsilon = \int_0^1 \frac{d}{d\epsilon}\left\{W(T^{(\epsilon)}(\mathbf{z}))\right\} d\epsilon = W(T^{(1)}(\mathbf{z})) - W(T^{(0)}(\mathbf{z}))$$

where the first equality follows by the chain rule for Gateaux derivatives and the second equality follows by the fundamental theorem of calculus.

(2) $\implies$ (3)

This follows by a reduction to the finite dimensional case. Fix arbitrary $\tau, \gamma \in \mathcal{T}$ and define $P(s,t) = W(T_0 + s\tau + t\gamma)$ for (s,t) in a neighborhood of $(0,0)$ in $\mathbb{R}^2$ (this is well-defined since $\mathcal{T}$ is convex). Then:

$$\frac{\partial^2 P}{\partial s \partial t}(0,0) = D^2 W_{T_0}[\tau, \gamma], \qquad \frac{\partial^2 P}{\partial t \partial s}(0,0) = D^2 W_{T_0}[\gamma, \tau]$$

where $D^2 W_{T_0}[\tau, \gamma]$ represents the second Gateaux variation of W in the directions τ and γ (i.e., the Gateaux variation of $\int_{\mathbf{N}} g(\mathbf{n}; T)\tau(\mathbf{z}(\mathbf{n}))d\mathbf{n}$ in the direction γ). Continuity of $T \mapsto D^2 W_T[\tau, \gamma]$ implies continuity of $(s,t) \mapsto \frac{\partial^2 P}{\partial s \partial t}(s,t)$, so by the classical Schwarz theorem applied to P:

$$D^2 W_{T_0}[\tau, \gamma] = \frac{\partial^2 P}{\partial s \partial t}(0,0) = \frac{\partial^2 P}{\partial t \partial s}(0,0) = D^2 W_{T_0}[\gamma, \tau]$$

(1) $\implies$ (2)

We will show that if a set of weights g is path independent, then they are the Gateaux derivative of some objective function W. First, let us define the proposed objective function for any given $\hat{T}$. Pick an arbitrary schedule T_0 and select a path such that $T^{(0)}(\mathbf{z}) = T_0(\mathbf{z})$ and $T^{(1)}(\mathbf{z}) = \hat{T}(\mathbf{z})$ (such a path exists assuming $\mathcal{T}$ is path connected, which is implied by convexity) and define:

$$W(\hat{T}) \equiv \int_0^1 \int_{\mathbf{N}} g(\mathbf{n}; T^{(\epsilon)})\, T_\epsilon^{(\epsilon)}(\mathbf{z})\, d\mathbf{n}\, d\epsilon$$

We want to show that

$$\int_{\mathbf{N}} g(\mathbf{n}; \hat{T})\, \tau(\mathbf{z}(\mathbf{n}))\, d\mathbf{n}$$

is the Gateaux derivative of the objective function W evaluated at $\hat{T}$ in the direction γ.

Consider any path $\tilde{T}^{(\epsilon)}(\mathbf{z})$ such that $\tilde{T}^{(0)}(\mathbf{z}) = \hat{T}(\mathbf{z})$ and $\tilde{T}^{(0)}_{\epsilon}(\mathbf{z}) = \tau(\mathbf{z})$. We have:

$$\begin{aligned} & W(\tilde{T}^{(v)}) - W(\hat{T}) \\ &= \int_0^1 \int_{\mathbf{N}} g(\mathbf{n}; T^{(\epsilon)})\, T^{(\epsilon)}_{\epsilon}(\mathbf{z})\, d\mathbf{n}\, d\epsilon \\ &\quad + \int_0^v \int_{\mathbf{N}} g(\mathbf{n}; \tilde{T}^{(\epsilon)}) \tilde{T}^{(\epsilon)}_{\epsilon}(\mathbf{z}(n)) d\mathbf{n} d\epsilon \\ &\quad - \int_0^1 \int_{\mathbf{N}} g(\mathbf{n}; T^{(\epsilon)})\, T^{(\epsilon)}_{\epsilon}(\mathbf{z})\, d\mathbf{n}\, d\epsilon \\ &= \int_0^v \int_{\mathbf{N}} g(\mathbf{n}; \tilde{T}^{(\epsilon)}) \tilde{T}^{(\epsilon)}_{\epsilon}(\mathbf{z}(n)) d\mathbf{n} d\epsilon \end{aligned} \tag{31}$$

where the first equality uses path independence by selecting the path that first integrates between T_0 and $\hat{T}$ and then integrates from $\hat{T}$ to $\tilde{T}^{(v)}$.

Next, let $\tilde{T}^{(\epsilon)} = \hat{T} + \epsilon\tau$. Dividing by v and taking the limit as $v \to 0$ yields that

$$DW_{\hat{T}}[\tau] = \int_{\mathbf{N}} g(\mathbf{n}; \hat{T})\, \tau(\mathbf{z}(\mathbf{n}))\, d\mathbf{n}$$

as long as $g(\mathbf{n}; T)$ is continuous a.e. in T at $\hat{T}$.[22]

(3) $\implies$ (2)

Let T, τ, γ represent arbitrary elements of $\mathcal{T}$. We show that if the Gateaux variation of the functional $\int_{\mathbf{N}} g(\mathbf{n}; T)\tau(\mathbf{z}(\mathbf{n}))d\mathbf{n}$ exists and is symmetric, then $\int_{\mathbf{N}} g(\mathbf{n}; T)\tau(\mathbf{z}(\mathbf{n}))d\mathbf{n}$ is the Gateaux derivative of an objective function W in the direction τ.

First, let us pick an arbitrary tax schedule T_0 and define the proposed objective function for any other tax schedule γ via:

$$W(\gamma) = \int_0^1 \int_{\mathbf{N}} g(\mathbf{n}; T_0 + \epsilon(\gamma - T_0))[\gamma(\mathbf{z}(\mathbf{n})) - T_0(\mathbf{z}(\mathbf{n}))] d\mathbf{n}\, d\epsilon.$$

Next, take the Gateaux variation of W at γ in the direction $\tau \in \mathcal{T}$:

$$DW_{\gamma}[\tau] \equiv \lim_{v \to 0} \frac{W(\gamma + v\tau) - W(\gamma)}{v}.$$

Note that:

$$W(\gamma + v\tau) = \int_0^1 \int_{\mathbf{N}} g(\mathbf{n}; T_0 + \epsilon(\gamma + v\tau - T_0))\, [\gamma(\mathbf{z}(\mathbf{n})) + v\tau(\mathbf{z}(\mathbf{n})) - T_0(\mathbf{z}(\mathbf{n}))]\, d\mathbf{n}\, d\epsilon.$$

We obtain:

$$\begin{aligned} DW_{\gamma}[\tau] = \int_0^1 \Bigg[& \int_{\mathbf{N}} g(\mathbf{n}; T_0 + \epsilon(\gamma - T_0))\tau(\mathbf{z}(\mathbf{n}))\, d\mathbf{n} \\ & + \frac{d}{dv} \left\{ \int_{\mathbf{N}} g(\mathbf{n}; T_0 + \epsilon(\gamma + v\tau - T_0))[\gamma(\mathbf{z}(\mathbf{n})) - T_0(\mathbf{z}(\mathbf{n}))]\, d\mathbf{n} \right\} \Bigg|_{v=0} \Bigg] d\epsilon \end{aligned}$$

[22] Technically, we also need that $|g(\mathbf{n}; \hat{T})| \leq h(\mathbf{n})$ for some function $h(\mathbf{n})$ to apply the dominated convergence theorem; we will assume this always holds for any given weights $g(\mathbf{n}; T)$.

Symmetry of the Gateaux variation of g implies that for any starting tax schedule $\hat{T}$:

$$\frac{d}{dv}\left\{\int_{\mathbf{N}} g(\mathbf{n};\hat{T}+\epsilon v\tau)[\gamma(\mathbf{z}(\mathbf{n}))-T_0(\mathbf{z}(\mathbf{n}))]\,d\mathbf{n}\right\}\Big|_{v=0} = \frac{d}{dv}\left\{\int_{\mathbf{N}} g(\mathbf{n};\hat{T}+\epsilon v[\gamma-T_0])\tau(\mathbf{z}(\mathbf{n}))\,d\mathbf{n}\right\}\Big|_{v=0}$$

Setting $\hat{T} = T_0 + \epsilon(\gamma - T_0)$ and applying the standard unidimensional chain rule:

$$\begin{aligned}
&\frac{d}{dv}\left\{\int_{\mathbf{N}} g(\mathbf{n};T_0+\epsilon(\gamma-T_0)+\epsilon v\tau)[\gamma(\mathbf{z}(\mathbf{n}))-T_0(\mathbf{z}(\mathbf{n}))]\,d\mathbf{n}\right\}\Big|_{v=0}\\
&= \frac{d}{dv}\left\{\int_{\mathbf{N}} g(\mathbf{n};T_0+\epsilon(\gamma-T_0)+\epsilon v(\gamma-T_0))\tau(\mathbf{z}(\mathbf{n}))\,d\mathbf{n}\right\}\Big|_{v=0}\\
&= \epsilon\frac{d}{d\epsilon}\left\{\int_{\mathbf{N}} g(\mathbf{n};T_0+\epsilon(\gamma-T_0))\tau(\mathbf{z}(\mathbf{n}))\,d\mathbf{n}\right\}
\end{aligned}$$

Thus,

$$\begin{aligned}
DW_\gamma[\tau] = \int_0^1 \Bigg[&\int_{\mathbf{N}} g(\mathbf{n};T_0+\epsilon(\gamma-T_0))\tau(\mathbf{z}(\mathbf{n}))\,d\mathbf{n}\\
&+\epsilon\frac{d}{d\epsilon}\left\{\int_{\mathbf{N}} g(\mathbf{n};T_0+\epsilon(\gamma-T_0))\tau(\mathbf{z}(\mathbf{n}))\,d\mathbf{n}\right\}\Bigg]d\epsilon
\end{aligned} \tag{32}$$

Performing integration by parts yields that this term equals:

$$\left[\epsilon\int_{\mathbf{N}} g(\mathbf{n};T_0+\epsilon(\gamma-T_0))\tau(\mathbf{z}(\mathbf{n}))\,d\mathbf{n}\right]_{\epsilon=0}^{\epsilon=1} - \int_0^1\left[\int_{\mathbf{N}} g(\mathbf{n};T_0+\epsilon(\gamma-T_0))\tau(\mathbf{z}(\mathbf{n}))\,d\mathbf{n}\right]d\epsilon.$$

Combining terms gives:

$$DW_\gamma[\tau] = \int_{\mathbf{N}} g(\mathbf{n};\gamma)\tau(\mathbf{z}(\mathbf{n}))\,d\mathbf{n}$$

Hence, the Gateaux derivative of W at γ in direction τ is

$$\int_{\mathbf{N}} g(\mathbf{n};\gamma)\tau(\mathbf{z}(\mathbf{n}))\,d\mathbf{n}$$

as claimed. □

A.3 Proof of Proposition Sher

Proof. We want to solve the following partial differential equation (PDE) for $g(c,z,n)$ on some open set of (c,z) for each n (noting that there are no g_n or u_n terms that enter the equation so that the solution varies with each value of n):[23]

$$-u_z\,g_c + u_c\,g_z = g\left(u_{cc}\left(-\frac{u_z}{u_c}\right)+u_{cz}\right)$$

where we have suppressed the arguments (c,z,n) for brevity. We will solve this PDE via the *method of characteristics*. The characteristic equations for this system are given by:

$$\frac{dc}{ds} = -u_z, \qquad \frac{dz}{ds} = u_c, \qquad \frac{dg}{ds} = g\left(u_{cc}\left(-\frac{u_z}{u_c}\right)+u_{cz}\right)$$

Next, we will solve the ODE given by $\frac{dg}{ds} = g\left(u_{cc}\left(-\frac{u_z}{u_c}\right)+u_{cz}\right)$. After dividing both

[23] We have not actually shown that for each n, the set of (c,z) chosen optimally under some tax schedule $T\in\mathcal{T}$ is an open set. We prove this (very) technical result at the end of Appendix B.3.

sides by g we get:

$$\frac{d}{ds}\log g = -\frac{u_{cc}\,u_z}{u_c} + u_{cz}$$

Using the equations $\frac{dc}{ds} = -u_z$ and $\frac{dz}{ds} = u_c$, we can rewrite the right-hand-side:

$$\frac{d}{ds}\log u_c = \frac{dc}{ds}\frac{u_{cc}}{u_c} + \frac{dz}{ds}\frac{u_{cz}}{u_c}$$

Therefore:

$$\frac{d}{ds}\log g = \frac{d}{ds}\log u_c$$

Integrating:

$$\log g = \log u_c + C$$

where C is a constant that can vary arbitrarily along each characteristic curve. Next, we claim that along each characteristic curve, $u(c, z, n)$ is constant. To see this, note that $\frac{du(c(s),z(s),n)}{ds} = u_c\frac{dc}{ds} + u_z\frac{dz}{ds} = -u_c u_z + u_c u_z = 0$. Hence, characteristic curves are iso-utility curves for a given n, yielding:

$$\log g = \log u_c + C(u)$$

Finally, note that there is a symmetry condition given by Equation (12) for each type n and because there are no g_n terms in the PDE, the constant $C(u)$ can also depend freely on n as well. Integrating:

$$g(c, z, n) = u_c(c, z, n) \cdot \psi(u(c, z, n), n)$$

for some arbitrary function ψ. □

A.4 Proof of Proposition 2 General

Proof. Suppose that one starts with the objective function

$$W(T) = \int_N M(\mathbf{n}; T, \mathbf{z}(\cdot, T), \mathbf{w}(T)) f(\mathbf{n}) d\mathbf{n}$$

Consider taking the Gateaux derivative of W in the direction of τ starting at an initial tax schedule T, but splitting this Gateaux derivative up as follows:

$$DW(T)[\tau] = \lim_{\epsilon\to 0} \frac{\int_N M(\mathbf{n}; T+\epsilon\tau, \mathbf{z}(\cdot, T), \mathbf{w}(T)) f(\mathbf{n}) d\mathbf{n} - \int_N M(\mathbf{n}; T, \mathbf{z}(\cdot, T), \mathbf{w}(T)) f(\mathbf{n}) d\mathbf{n}}{\epsilon}$$
$$+ \lim_{\epsilon\to 0} \frac{\int_N M(\mathbf{n}; T+\epsilon\tau, \mathbf{z}(\cdot, T+\epsilon\tau), \mathbf{w}(T+\epsilon\tau)) f(\mathbf{n}) d\mathbf{n} - \int_N M(\mathbf{n}; T+\epsilon\tau, \mathbf{z}(\cdot, T), \mathbf{w}(T)) f(\mathbf{n}) d\mathbf{n}}{\epsilon}$$
$$\equiv DM_T[\tau]\big|_{\hat{T}=T} + DM_{\hat{T}}[\tau]\big|_{\hat{T}=T}$$

Next, note that:

$$M(\mathbf{n}; T+\epsilon\tau, \mathbf{z}(\cdot, \hat{T}), \mathbf{w}(\hat{T})) = \int_0^1 m(\mathbf{n}; T_0 + v(T+\epsilon\tau-T_0), \mathbf{z}(\cdot, \hat{T}), \mathbf{w}(\hat{T}))(T(\mathbf{z}(\mathbf{n})) + \epsilon\tau(\mathbf{z}(\mathbf{n})) - T_0(\mathbf{z}(\mathbf{n})))dv$$

We obtain:

$$DM_T[\tau]\big|_{\hat{T}=T} = \int_0^1 \left[\int_{\mathbf{N}} m(\mathbf{n}; T_0 + v(T-T_0), \mathbf{z}(\cdot,T), \mathbf{w}(T))\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})d\mathbf{n} \right.$$
$$\left. + \frac{d}{d\epsilon}\left\{ \int_{\mathbf{N}} m(\mathbf{n}; T_0 + v(T+\epsilon\tau-T_0), \mathbf{z}(\cdot,T), \mathbf{w}(T))(T(\mathbf{z}(\mathbf{n})) - T_0(\mathbf{z}(\mathbf{n})))f(\mathbf{n})\, d\mathbf{n} \right\}\Big|_{\epsilon=0} \right] dv \tag{33}$$

Symmetry of the Gateaux derivative of m implies that:

$$\begin{aligned} &\frac{d}{d\epsilon}\left\{ \int_{\mathbf{N}} m(\mathbf{n}; T_0 + v(T-T_0) + v\epsilon\tau, \mathbf{z}(\cdot,T), \mathbf{w}(T))(T(\mathbf{z}(\mathbf{n})) - T_0(\mathbf{z}(\mathbf{n})))f(\mathbf{n})\, d\mathbf{n} \right\}\Big|_{\epsilon=0} \\ &= \frac{d}{d\epsilon}\left\{ \int_{\mathbf{N}} m(\mathbf{n}; T_0 + v(T-T_0) + \epsilon(T-T_0), \mathbf{z}(\cdot,T), \mathbf{w}(T))v\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})\, d\mathbf{n} \right\}\Big|_{\epsilon=0} \\ &= v\frac{d}{dv}\left\{ \int_{\mathbf{N}} m(\mathbf{n}; T_0 + v(T-T_0), \mathbf{z}(\cdot,T), \mathbf{w}(T))\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})\, d\mathbf{n} \right\} \end{aligned} \tag{34}$$

where the first equality uses the symmetry condition from Equation (25) (where $T_0 + v(T-T_0)$ takes the place of T and $(T-T_0)$ takes the place of γ in Equation (25)) and the last equality just uses the fact that:

$$\begin{aligned} &\frac{d}{d\epsilon}\left\{ \int_{\mathbf{N}} m(\mathbf{n}; T_0 + (v+\epsilon)(T-T_0), \mathbf{z}(\cdot,T), \mathbf{w}(T))\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})\, d\mathbf{n} \right\}\Big|_{\epsilon=0} \\ &= \frac{d}{dv}\left\{ \int_{\mathbf{N}} m(\mathbf{n}; T_0 + (v+\epsilon)(T-T_0), \mathbf{z}(\cdot,T), \mathbf{w}(T))\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})\, d\mathbf{n} \right\}\Big|_{\epsilon=0} \end{aligned} \tag{35}$$

Thus, we obtain:

$$\begin{aligned} DM_T[\tau]\big|_{\hat{T}=T} = &\int_0^1 \left[\int_{\mathbf{N}} m(\mathbf{n}; T_0 + v(T-T_0), \mathbf{z}(\cdot,T), \mathbf{w}(T))\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})d\mathbf{n} \right. \\ &\left. + v\frac{d}{dv}\left\{ \int_{\mathbf{N}} m(\mathbf{n}; T_0 + v(T-T_0), \mathbf{z}(\cdot,T), \mathbf{w}(T))\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})\, d\mathbf{n} \right\} \right] dv \end{aligned} \tag{36}$$

Performing integration by parts on the second term yields:

$$\begin{aligned} &v\int_{\mathbf{N}} m(\mathbf{n}; T_0 + v(T-T_0), \mathbf{z}(\cdot,T), \mathbf{w}(T))\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})\, d\mathbf{n}\Big|_{v=0}^{v=1} \\ &- \int_0^1 \int_{\mathbf{N}} m(\mathbf{n}; T_0 + v(T-T_0), \mathbf{z}(\cdot,T), \mathbf{w}(T))\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})d\mathbf{n}dv \end{aligned} \tag{37}$$

Combining terms gives:

$$DM_T[\tau]\big|_{\hat{T}=T} = \int_{\mathbf{N}} m(\mathbf{n}; T, \mathbf{z}(\cdot,T), \mathbf{w}(T))\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})\, d\mathbf{n}$$

Finally, we have:

$$DW(T)[\tau] = \int_{\mathbf{N}} m(\mathbf{n}; T, \mathbf{z}(\cdot,T), \mathbf{w}(T))\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})\, d\mathbf{n} + DM_{\hat{T}}[\tau]\big|_{\hat{T}=T}$$

Thus,

$$\int_{\mathbf{N}} m(\mathbf{n}; T, \mathbf{z}(\cdot,T), \mathbf{w}(T))\tau(\mathbf{z}(\mathbf{n}))f(\mathbf{n})\, d\mathbf{n} + DM_{\hat{T}}[\tau]\big|_{\hat{T}=T}$$

is the Gateaux derivative of an objective function and therefore evaluating the desirability

of tax reforms via Equation (26) is consistent in the sense of Definition 5 using the exact same logic as the proof of point 2 in Theorem 1. □

B Appendix: Additional Results

B.1 Derivation of Equation (19)

Assuming all individuals have a unique optimum under the smooth tax schedule T and that $f(n) \to 0$ as $n \to \underline{n}$ or $n \to \overline{n}$, the Gateaux derivative of GDP-per-capita is:

$$\begin{aligned}
DGDP_T[\tau] &= \int_N \frac{\partial z(n;\epsilon)}{\partial \epsilon}\Big|_{\epsilon=0} f(n)dn \\
&= \int_N \frac{-\tau'(z(n))}{T''(z(n)) + \frac{kz(n)^{k-1}}{n^{k+1}}} f(n)dn \\
&= \int_N \frac{-1}{T''(z(n)) + \frac{kz(n)^{k-1}}{n^{k+1}}} f(n) \left(\frac{dz}{dn}\right)^{-1} \tau'(z(n))\frac{dz}{dn}dn \\
&= \int_N \frac{d}{dn}\left[\frac{1}{T''(z(n)) + \frac{kz(n)^{k-1}}{n^{k+1}}} f(n) \left(\frac{dz}{dn}\right)^{-1}\right] \tau(z(n))dn \\
&= \int_N \frac{d}{dn}\left[\frac{1}{(1+k)z(n)^k/n^{k+2}} f(n)\right] \tau(z(n))dn \\
&= \int_N \left\{ \frac{f'(n)n + f(n)}{(1+k)(1-T'(z))} + \frac{f(n)T''(z(n))}{(1-T'(z(n)))\left[T''(z(n)) + \frac{kz(n)^{k-1}}{n^{k+1}}\right]} \right\} \tau(z(n))dn
\end{aligned}$$

where $\frac{dz}{d\epsilon}$ is given by Equation (15) and the expression for $\frac{dz}{dn}$ is given by Equation (16).[24] The final equality evaluates the $\frac{d}{dn}$ derivative after substituting $1 - T'(z) = \frac{z^k}{n^{1+k}}$.

Next, let us turn to calculating the Gateaux derivative of the Gini coefficient. Plugging in the expression for $\frac{dz}{dn}$ from Equation (16), the Gini coefficient can be written as:

$$\frac{1}{GDP(T)} \int_N F(n)(1-F(n)) \frac{(1+k)[1-T'(z(n))]}{n\left[T''(z(n)) + k\frac{z^{k-1}}{n^{k+1}}\right]} dn$$

Let us first focus on the $\int_N F(n)(1-F(n)) \frac{(1+k)[1-T'(z(n))]}{n\left[T''(z(n)) + k\frac{z^{k-1}}{n^{k+1}}\right]} dn$ term. The Gateaux derivative of this term is:

$$\int_N \frac{d}{dz}\left\{ F(n)(1-F(n)) \frac{(1+k)[1-T'(z(n))]}{n\left[T''(z(n)) + k\frac{z(n)^{k-1}}{n^{k+1}}\right]} \right\} \frac{dz}{d\epsilon} dn$$

where $\frac{dz}{d\epsilon}$ is given by Equation (15). Applying the quotient rule and plugging in the expression for $\frac{dz}{d\epsilon}$ yields a Gateaux variation of this term given by:

$$\underbrace{\int_N F(n)(1-F(n)) \frac{1+k}{n} \left(\frac{T''(z(n))}{\left[T''(z(n)) + k\frac{z^{k-1}}{n^{k+1}}\right]^2} + \frac{\frac{z^k}{n^{k+1}}\left[T'''(z(n)) + k(k-1)\frac{z(n)^{k-2}}{n^{k+1}}\right]}{\left[T''(z(n)) + k\frac{z^{k-1}}{n^{k+1}}\right]^3} \right) \tau'(z(n))\, dn}_{\equiv A(n)}$$

[24] We have also used the assumption that $f(n) \to 0$ as $n \to \underline{n}$ or $n \to \overline{n}$ so that the boundary term arising from integration by parts equals 0.

Multiplying and dividing by $\frac{dz}{dn}$ and then integrating by parts we find that the Gateaux derivative of this term equals:

$$\int_N \frac{d}{dn}\left[A(n)\frac{n[T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}]}{(1+k)(1-T'(z(n)))}\right]\tau(z(n))dn$$

Finally, this yields a Gateaux derivative of the Gini coefficient of:

$$\frac{GDP(T)\int_N \frac{d}{dn}\left[A(n)\frac{n\left[T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}\right]}{(1+k)(1-T'(z(n)))}\right]\tau(z(n))dn - DGDP_T[\tau]Gini(T)GDP(T)}{GDP(T)^2}$$

$$= \frac{1}{GDP(T)}\int_N \frac{d}{dn}\left[A(n)\frac{n\left[T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}\right]}{(1+k)(1-T'(z(n)))}\right]\tau(z(n))dn$$

$$-\frac{Gini(T)}{GDP(T)}\int_N\left[\frac{f'(n)n+f(n)}{(1+k)(1-T'(z(n)))}+\frac{f(n)T''(z(n))}{(1-T'(z(n)))\left[T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}\right]}\right]\tau(z(n))dn \tag{38}$$

B.2 Proving Proposition Sher within Constant Revenue Tax Schedules

When deriving Proposition Sher, we applied Theorem 1 to infer that for any consistent set of weights $g(c,z,n)$, the Gateaux derivative of $\int_{\mathbf{N}} g(c(z(n)),z(n),n)\,\gamma(z(n))\,f(n)\,dn$ in the direction of $\tau(z)$ must be symmetric in τ and γ. However, this statement is only true if the domain of tax functions $\mathcal{T}$ is convex - and the set of well-behaved tax schedules (as in Assumption 1) that also yield the same level of government revenue $\int_N T(z(n))f(n)dn = \hat{R}$ for some level $\hat{R}$ is not necessarily convex. However, under the milder assumption that the set of tax functions that yield the same level of government revenue is *path connected*, then any set of consistent generalized welfare weights $g(c,z,n)$ are such that $w(T) = \int_{\mathbf{N}} g(c(z(n)),z(n),n)\,\gamma(z(n))\,f(n)\,dn$ is the Frechet derivative of some objective function $W(T)$ as long as $w(T)$ is continuous. To see this note that Equation (31) in the proof to Theorem 1 can be applied to any vectors $\tau(\mathbf{z}) = \tilde{T}_\epsilon^{(0)}(\mathbf{z})$ that are tangent to the path-connected manifold. Hence, assuming that the $w(T)$ is continuously Frechet differentiable, then $W(T)$ is twice continuously Frechet differentiable, which means that it's second cross partial derivatives are symmetric. Hence, if $w(T)$ is continuously Frechet differentiable then any consistent set of generalized welfare weights must satisfy the Condition 3 of Theorem 1 for all directions that are tangent to the path-connected domain.

From here, suppose that government revenue $R(T) \equiv \int_{\mathbf{N}} T(\mathbf{z}(\mathbf{n}))dF(\mathbf{n})$ is Gateaux

differentiable with derivative in the direction τ given by:[25]

$$DR_T[\tau] = \int_Z \tau(z)dP(z)$$

Recall that the symmetry condition for generalized welfare weights $g(c, z, n)$ is given by:

$$\int_N \{\xi(n)\left[g_c\, c'(z(n)) + g_z\right] + \eta(n)g\} \times \left[\tau'(z(n))\gamma(z(n)) - \tau(z(n))\gamma'(z(n))\right] f(n)dn \quad (39)$$

However, when restricted to the set of tax schedules that yield a fixed amount of revenue, this symmetry condition only needs to hold for τ and γ satisfying $\int_Z \tau(z)dP(z) = 0$ and $\int_Z \gamma(z)dP(z) = 0$ (i.e., in the tangent space to the given manifold of revenue constant tax schedules).

Claim 1. *For any function $W(z)$ such that $W(z)$ is supported on an interval $[z_l, z_h]$ with $z_l > \underline{z}$ and $z_h < \overline{z}$, then there exist $\tau(z)$ and $\gamma(z)$ s.t. $(\tau'(z)\gamma(z) - \tau(z)\gamma'(z)) = W(z)$, $\int_Z \tau(z)dP(z) = 0$, and $\int_Z \gamma(z)dP(z) = 0$.*

Proof. First, normalize $W(z)$ so that $\int_Z W(z)dz = 1$. Then set:

$$\tau(z) = \begin{cases} 0 & \text{if } z < z_l \\ \int_0^z W(t)dt & \text{if } z \in [z_l, z_h] \\ \gamma(z) & \text{if } z > z_h \end{cases}$$

And pick $\gamma(z)$ so that $\gamma(z) = 1$ on $[z_l, z_h]$. Finally, we choose $\gamma(z)$ for the rest of the domain so that:

$$\int_{\underline{z}}^{z_l} \gamma(z)dP(z) + \int_{z_l}^{z_h} dP(z) + \int_{z_h}^{\overline{z}} \gamma(z)dP(z) = 0$$
$$\int_{\underline{z}}^{z_l} 0dP(z) + \int_{z_l}^{z_h} \int_0^z W(t)dtdP(z) + \int_{z_h}^{\overline{z}} \gamma(z)dP(z) = 0$$

Then we just pick $\gamma(z)$ on $[z_h, \overline{z}]$ to satisfy the second equation and then choose $\gamma(z)$ on $[\underline{z}, z_l]$ to satisfy the first equation given our choice of $\gamma(z)$ on $[z_h, \overline{z}]$. As long as $\int_{z_h}^{\overline{z}} \gamma(z)dP(z)$ and $\int_{\underline{z}}^{z_l} \gamma(z)dP(z)$ are not identically zero $\forall\gamma(z)$ then this procedure works. If either one is identically zero, then just redefine the domain $[\underline{z}, \overline{z}]$ by shrinking it inwards so that this is not the case and reapply the construction above. On portions of the domain where $\int_{z_h}^{\overline{z}} \gamma(z)dP(z) = 0$ or $\int_{\underline{z}}^{z_l} \gamma(z)dP(z) = 0$ then to find τ and γ s.t. $(\tau'(z)\gamma(z) - \tau(z)\gamma'(z)) = W(z)$, $\int_Z \tau(z)dP(z) = 0$, and $\int_Z \gamma(z)dP(z) = 0$, we can just set $\tau(z) = \int_0^z W(t)dt$ and $\gamma(z) = 1$. □

[25] We can always express the Gateaux derivative in this manner under Assumption 1 (see Section 3.1 of Bergstrom and Dodds (2026)). More generally, the Gateaux derivative of government revenue can be expressed as $\int_Z \tau(z)dP(z)$ under fairly mild regularity conditions; see Proposition 1 of Bergstrom and Dodds (2026).

By Claim 1, we know that the following must hold for any compactly supported $W(z)$:

$$\int_N \left\{\xi(n)\left[g_c\, c'(z(n)) + g_z\right] + \eta(n)g\right\} W(z(n))f(n)dn \tag{40}$$

But to apply the fundamental lemma of the calculus of variations, all we need is that Equation (40) holds for all compactly supported $W(z)$ which means that for all (c, z, n) that arise under some $T \in \mathcal{T} \cap \{T : \int_Z \tau(z)dP(z) = 0\}$:

$$\xi(n)\left[g_c(c, z, n)\, c'(z) + g_z(c, z, n)\right] + \eta(n)g(c, z, n) = 0$$

Hence, we have shown that the same symmetry condition used in Proposition Sher must hold even if we restrict the set of admissible tax perturbations to be those that hold revenue constant (under the slightly stronger assumption of continuous Frechet differentiability of $\int_{\mathbf{N}} g(c(z(n)), z(n), n)\, \gamma(z(n))\, f(n)\, dn$ rather than existence of a continuous Gateaux variation). From here Proposition Sher illustrates that consistent generalized welfare weights $g(c, z, n)$ must be structurally utilitarian even if we only mandate that $g(c, z, n)$ be consistent when restricted to tax schedules that do not change government revenue.[26]

B.3 Proof of Proposition Sher Multidimensional

Proof. To begin, consider how $\int_{\mathbf{N}} g(c(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}), \mathbf{n})\, \gamma(\mathbf{z}(\mathbf{n}))\, f(\mathbf{n})\, d\mathbf{n}$ varies if we perturb the tax schedule from $T(\mathbf{z})$ to $T(\mathbf{z})+\epsilon\tau(\mathbf{z})$, noting that $\mathbf{z}(\mathbf{n})$ and $c(\mathbf{z}(\mathbf{n}))$ are both functions of the tax schedule. In particular, for any given tax schedule $T(\mathbf{z}) + \epsilon\tau(\mathbf{z})$, individuals' optima satisfy the following first order condition:

$$\begin{aligned} &u_c(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n})\left(y_{z_1}(\mathbf{z}) - T_{z_1}(\mathbf{z}) - \epsilon\tau_{z_1}(\mathbf{z})\right) + u_{z_1}(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) = 0 \\ &\qquad\qquad \vdots \\ &u_c(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n})\left(y_{z_J}(\mathbf{z}) - T_{z_J}(\mathbf{z}) - \epsilon\tau_{z_J}(\mathbf{z})\right) + u_{z_J}(y(\mathbf{z}) - T(\mathbf{z}) - \epsilon\tau(\mathbf{z}), \mathbf{z}; \mathbf{n}) = 0 \end{aligned} \tag{41}$$

Under Assumption 2, we can determine the derivative of $\mathbf{z}(\mathbf{n})$ with respect to ϵ for any given function $\tau(\mathbf{z})$ via the implicit function theorem (again, recall that $\mathbf{z}(\mathbf{n})$ is also a function of the tax schedule):

$$\begin{aligned} &\frac{\partial \mathbf{z}}{\partial \epsilon}(\mathbf{n}) = -\mathbf{H}^{-1}(\mathbf{n}) FOC(\mathbf{n})_\epsilon|_{\epsilon=0} = -\mathbf{H}^{-1}(\mathbf{n})[\mathbf{a}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{B}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z})] \\ &\equiv \vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z}) \end{aligned} \tag{42}$$

where $FOC_\epsilon|_{\epsilon=0}$ is the vector of derivatives of the first order conditions 41 with respect to ϵ. The second equality in Equation 42 follows for a vector $\mathbf{a}(\mathbf{n})$ with the j^{th} component given by $-[y_{z_1}(\mathbf{z}) - T_{z_1}(\mathbf{z})]u_{cc}(c, \mathbf{z}, \mathbf{n}) - u_{cz_j}(c, \mathbf{z}, \mathbf{n})$ and a diagonal matrix $\mathbf{B}(\mathbf{n})$ where the j^{th} diagonal element is given by $-u_c(c, \mathbf{z}, \mathbf{n})$. The third equality in Equation 42 follows

[26] While proved using an entirely different setup, the ideas in Claim 1 also the fundamental ideas behind the corresponding result (Lemma 5) in Sher (2024).

by defining $\vec{\eta}(\mathbf{n}) \equiv -\mathbf{H}^{-1}(\mathbf{n})\mathbf{a}(\mathbf{n})$ and $\mathbf{X}(\mathbf{n}) \equiv -\mathbf{H}^{-1}(\mathbf{n})\mathbf{B}(\mathbf{n})$.

Next, we want to explore when the Gateaux variation of

$$\int_{\mathbf{N}} g(c(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}), \mathbf{n})\, \gamma(\mathbf{z}(\mathbf{n}))\, f(\mathbf{n})\, d\mathbf{n} \tag{43}$$

in the direction of $\tau(\mathbf{z})$ is symmetric in γ and τ. Considering a perturbation from T to $T + \epsilon\tau$, we then differentiate Equation (43) with respect to ϵ, recognizing that the derivative of $\mathbf{z}$ w.r.t. ϵ is given by Equation (42):

$$\begin{aligned}\frac{d}{d\epsilon}\int_{\mathbf{N}} g(c(\mathbf{z}(\mathbf{n})), \mathbf{z}(\mathbf{n}), \mathbf{n})\, \gamma(\mathbf{z}(\mathbf{n}))\, f(\mathbf{n})\, d\mathbf{n} = -\int_{\mathbf{N}} g_c\, \gamma(\mathbf{z}(\mathbf{n}))\tau(\mathbf{z}(\mathbf{n}))\, f(\mathbf{n})\, d\mathbf{n} + \\ \int_{\mathbf{N}} \left[g_c\, \nabla_{\mathbf{z}} c(\mathbf{z}(\mathbf{n}))\, \frac{d\mathbf{z}(\mathbf{n})}{d\epsilon}\, \gamma(\mathbf{z}(\mathbf{n})) + \nabla_{\mathbf{z}} g\, \frac{d\mathbf{z}(\mathbf{n})}{d\epsilon}\, \gamma(\mathbf{z}(\mathbf{n})) + g\, \nabla_{\mathbf{z}}\gamma(\mathbf{z}(\mathbf{n}))\, \frac{d\mathbf{z}(\mathbf{n})}{d\epsilon} \right] f(\mathbf{n})\, d\mathbf{n}\end{aligned} \tag{44}$$

Noting that the first integral in Equation (44) is already symmetric in τ and γ, we focus on the second integral. Substituting $\frac{d\mathbf{z}(\mathbf{n})}{d\epsilon} = \vec{\eta}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{X}(\mathbf{n}) \cdot \nabla_{\mathbf{z}}\tau(\mathbf{z}) = -\mathbf{H}^{-1}(\mathbf{n})[\mathbf{a}(\mathbf{n})\tau(\mathbf{z}) - u_c(\mathbf{n})\nabla_{\mathbf{z}}\tau(\mathbf{z})]$:

$$\begin{aligned}&\int_{\mathbf{N}} \left[g_c\, \nabla_{\mathbf{z}} c(\mathbf{z}(\mathbf{n}))\, \frac{d\mathbf{z}(\mathbf{n})}{d\epsilon}\, \gamma(\mathbf{z}(\mathbf{n})) + \nabla_{\mathbf{z}} g\, \frac{d\mathbf{z}(\mathbf{n})}{d\epsilon}\, \gamma(\mathbf{z}(\mathbf{n})) + g\, \nabla_{\mathbf{z}}\gamma(\mathbf{z}(\mathbf{n}))\, \frac{d\mathbf{z}(\mathbf{n})}{d\epsilon} \right] f(\mathbf{n})\, d\mathbf{n} \\ &= \int_{\mathbf{N}} \left[g_c\, \nabla_{\mathbf{z}} c(\mathbf{z}(\mathbf{n}))\, \gamma(\mathbf{z}(\mathbf{n})) + \nabla_{\mathbf{z}} g\, \gamma(\mathbf{z}(\mathbf{n})) + g\, \nabla_{\mathbf{z}}\gamma(\mathbf{z}(\mathbf{n})) \right] \\ &\times \Big(-\mathbf{H}^{-1}(\mathbf{n})[\mathbf{a}(\mathbf{n})\tau(\mathbf{z}(\mathbf{n})) - u_c(\mathbf{n})\nabla_{\mathbf{z}}\tau(\mathbf{z}(\mathbf{n}))] \Big) f(\mathbf{n})\, d\mathbf{n}\end{aligned} \tag{45}$$

Symmetry requires that Equation (45) is equal to:

$$\begin{aligned}&\int_{\mathbf{N}} \left[g_c\, \nabla_{\mathbf{z}} c(\mathbf{z}(\mathbf{n}))\, \tau(\mathbf{z}(\mathbf{n})) + \nabla_{\mathbf{z}} g\, \tau(\mathbf{z}(\mathbf{n})) + g\, \nabla_{\mathbf{z}}\tau(\mathbf{z}(\mathbf{n})) \right] \\ &\times \Big(-\mathbf{H}^{-1}(\mathbf{n})[\mathbf{a}(\mathbf{n})\gamma(\mathbf{z}(\mathbf{n})) - u_c(\mathbf{n})\nabla_{\mathbf{z}}\gamma(\mathbf{z}(\mathbf{n}))] \Big) f(\mathbf{n})\, d\mathbf{n}\end{aligned} \tag{46}$$

Canceling terms and rearranging we get that symmetry requires:[27]

$$\begin{aligned}&\int_{\mathbf{N}} \{u_c(\mathbf{n})\, [g_c\, \nabla_{\mathbf{z}} c(\mathbf{z}(\mathbf{n})) + \nabla_{\mathbf{z}} g] + g\mathbf{a}(\mathbf{n})\} \\ &\times \Big(\mathbf{H}^{-1}(\mathbf{n})[\gamma(\mathbf{z}(\mathbf{n}))\nabla_{\mathbf{z}}\tau(\mathbf{z}(\mathbf{n})) - \tau(\mathbf{z}(\mathbf{n}))\nabla_{\mathbf{z}}\gamma(\mathbf{z}(\mathbf{n}))] \Big) f(\mathbf{n})\, d\mathbf{n} = 0\end{aligned} \tag{47}$$

By the fundamental lemma of the calculus of variations applied to vector valued functions, the following vector must be identically zero (where we have substituted in $\nabla_{\mathbf{z}} c(\mathbf{z}) = \nabla_{\mathbf{z}} y(\mathbf{z}) - \nabla_{\mathbf{z}} T(\mathbf{z}) = -\nabla_{\mathbf{z}} u / u_c$ by the FOCs):[28]

$$\{u_c(\mathbf{n})\, [-g_c(c, \mathbf{z}, \mathbf{n})\, \nabla_{\mathbf{z}} u / u_c + \nabla_{\mathbf{z}} g(c, \mathbf{z}, \mathbf{n})] + g(c, \mathbf{z}, \mathbf{n})\mathbf{a}(\mathbf{n})\}\, \mathbf{H}^{-1}(\mathbf{n}) = 0\ \forall \mathbf{n}$$

[27] Note, we have used the fact that $\mathbf{H}^{-1}$ is a symmetric matrix so that $x\mathbf{H}^{-1}y = y\mathbf{H}^{-1}x$ for two $1 \times J$ vectors x and y, recognizing that we omit the transpose symbol to improve clarity of the expressions above.

[28] For clarity, recall that $\{u_c(\mathbf{n})\, [-g_c(c, \mathbf{z}, \mathbf{n})\, \nabla_{\mathbf{z}} u / u_c + \nabla_{\mathbf{z}} g(c, \mathbf{z}, \mathbf{n})] + g(c, \mathbf{z}, \mathbf{n})\mathbf{a}(\mathbf{n})\}$ is a $1 \times J$ vector and $\mathbf{H}^{-1}(\mathbf{n})$ is a $J \times J$ marix.

Multiplying by $\mathbf{H}(\mathbf{n})$, we get that the following system of equations must hold for all possible tax schedules $T \in \mathcal{T}$, which in turn implies that the following must hold, for each n, on some open set (c, z):[29]

$$[g_c(c, \mathbf{z}, \mathbf{n})\, \nabla_{\mathbf{z}} u - u_c \nabla_{\mathbf{z}} g(c, \mathbf{z}, \mathbf{n})] - g(c, \mathbf{z}, \mathbf{n})[-u_{cc}\, \nabla_{\mathbf{z}} u / u_c - \nabla_{\mathbf{z}} u_c] = 0 \tag{48}$$

Thus, in order to satisfy symmetry of the second Gateaux derivative (which is necessary and sufficient for consistency by Theorem 1), Equation (48) must hold. We again use the method of characteristics to solve the system of PDEs given by Equation (48). Each of the PDEs in System (48) is equal to:

$$\left[g_c(c, \mathbf{z}, \mathbf{n}) u_{z_j} - u_c g_{z_j}(c, \mathbf{z}, \mathbf{n})\right] - g(c, \mathbf{z}, \mathbf{n})[-u_{cc} u_{z_j} / u_c - u_{cz_j}] = 0$$

Solving each such equation exactly as in the proof to Proposition Sher yields that for each point $\mathbf{z}_{-j}$ (where $\mathbf{z}_{-j}$ denotes coordinates for all $z_1, z_2, ..., z_J$ other than z_j) and each $\mathbf{n}$, the solution g must satisfy

$$\log g = \log u_c + C$$

where C is a constant that can vary arbitrarily along each characteristic curve, again noting that along each characteristic curve, $u(c, z, n)$ is constant by the exact same arguments as in the proof to Proposition Sher. Thus, the constant C depends on utility u and the constant C can also depend freely on $\mathbf{n}$ as in Proposition Sher. However, because the characteristic curves hold $\mathbf{z}_{-j}$ constant, the constant can also depend freely on $\mathbf{z}_{-j}$ so that the along any given iso-utility curve in (c, z_j) space:

$$g(c, \mathbf{z}, \mathbf{n}) = u_c(c, \mathbf{z}, \mathbf{n}) \cdot \psi(u(c, \mathbf{z}, \mathbf{n}), \mathbf{n}, \mathbf{z}_{-j})$$

But this must hold for all of the various z_j, which in turn implies that $\psi(u(c, \mathbf{z}, \mathbf{n}), \mathbf{n}, \mathbf{z}_{-j})$ cannot depend directly on $\mathbf{z}_{-j}$ (because if it did then the solution for some $z_k \neq z_j$ would have to take the form $u_c(c, \mathbf{z}, \mathbf{n}) \cdot \psi(u(c, \mathbf{z}, \mathbf{n}), \mathbf{n}, \mathbf{z}_{-j,k}, z_k)$ which is not a solution to the PDE by the arguments above).

Finally, we return to the technical result that the set of $(c, \mathbf{z})$ chosen optimally under $T \in \mathcal{T}$ is open for each n. First, we note that the mapping from $(\tau(\mathbf{z}), \nabla_{\mathbf{z}} \tau(\mathbf{z})) \mapsto \left(\frac{\partial c}{\partial \epsilon}(\mathbf{n}), \frac{\partial \mathbf{z}}{\partial \epsilon}(\mathbf{n})\right)$ given by Equation (42) and the relationship that $\frac{\partial c}{\partial \epsilon}(\mathbf{n}) = -\tau(\mathbf{z}(\mathbf{n})) + (1 - \nabla_{\mathbf{z}} T(\mathbf{z}(\mathbf{n}))) \cdot \frac{\partial \mathbf{z}}{\partial \epsilon}(\mathbf{n})$ is surjective for each $\mathbf{n}$. This follows because for each value of $\tau(\mathbf{z}(\mathbf{n}))$, the mapping $\nabla_{\mathbf{z}} \tau(\mathbf{z}) \mapsto \frac{\partial \mathbf{z}}{\partial \epsilon}(\mathbf{n})$ is surjective because $\mathbf{H}^{-1}$ is invertible by Assumption 2 and $\mathbf{B}(\mathbf{n})$ is invertible because it's diagonal. Hence, for any target values of $\left(\frac{\partial c}{\partial \epsilon}(\mathbf{n}), \frac{\partial \mathbf{z}}{\partial \epsilon}(\mathbf{n})\right)$ given by $(v_1, \mathbf{v}_2)$ just choose $\tau(\mathbf{z}(\mathbf{n})) = -v_1 + (1 - \nabla_{\mathbf{z}} T(\mathbf{z}(\mathbf{n}))) \cdot \mathbf{v}_2$ and then choose $\nabla_{\mathbf{z}} \tau(\mathbf{z})$ s.t. $-\mathbf{H}^{-1}(\mathbf{n})[\mathbf{a}(\mathbf{n})\tau(\mathbf{z}) + \mathbf{B}(\mathbf{n}) \cdot \nabla_{\mathbf{z}} \tau(\mathbf{z})] = \mathbf{v}_2$. Thus, the mapping from tax schedules $T \in \mathcal{T}$ to $(c(\mathbf{n}), \mathbf{z}(\mathbf{n}))$ for each $\mathbf{n}$ has a (Frechet) derivative which is locally surjective (note

[29] We prove the technical result that the set (c, z) is open provided that $\mathcal{T}$ is open at the very end of this Appendix.

that we have used the assumption that $\mathcal{T}$ is itself open in the set of smooth functions to infer that $T + \epsilon\tau \in \mathcal{T}$ for sufficiently small ϵ). Thus, by the Constant Rank Theorem for Banach spaces (Blot, 1986), the map from tax schedules $T \in \mathcal{T}$ to $(c(\mathbf{n}), \mathbf{z}(\mathbf{n}))$ is locally open for each $T \in \mathcal{T}$ and hence open.

□

B.4 Proof of Proposition 2 Multidimensional

Proof. Suppose that one starts with the objective function $W(T) = -\int_{\mathbf{N}} M(c(\mathbf{n}), \mathbf{z}(\mathbf{n}), \mathbf{n}) f(\mathbf{n}) d\mathbf{n}$ with $M(c, \mathbf{z}, \mathbf{n}) = \int m(c, \mathbf{z}, \mathbf{n}) dc$ and $m(c, \mathbf{z}, \mathbf{n})$ an arbitrary smooth function. Consider taking the Gateaux derivative of W in the direction of τ starting at an initial tax schedule T:

$$DW_T[\tau] = \int_{\mathbf{N}} M_c(c(\mathbf{n}), \mathbf{z}(\mathbf{n}), \mathbf{n}) \tau(\mathbf{z}(\mathbf{n})) f(\mathbf{n}) d\mathbf{n} - \int_{\mathbf{N}} [M_c(1 - T'(\mathbf{z}(\mathbf{n})) + M_{\mathbf{z}}] \frac{d\mathbf{z}}{d\epsilon} f(\mathbf{n}) d\mathbf{n}$$

Next, note that:

$$\int_{\mathbf{N}} [M_c(1-T'(\mathbf{z}(\mathbf{n}))+M_{\mathbf{z}}] \frac{d\mathbf{z}}{d\epsilon} f(\mathbf{n}) d\mathbf{n} = \int_{\mathbf{N}} [M_c(1-T'(\mathbf{z}(\mathbf{n})))+M_{\mathbf{z}}][\vec{\eta}(\mathbf{n})\tau(\mathbf{z}(\mathbf{n}))+\mathbf{X}(\mathbf{n})\nabla_{\mathbf{z}}\tau(\mathbf{z}(\mathbf{n}))] f(\mathbf{n}) d\mathbf{n}$$

Multiplying by $\left(\frac{d\mathbf{z}}{d\mathbf{n}}\right)^{-1} \frac{d\mathbf{z}}{d\mathbf{n}}$ (where $\frac{d\mathbf{z}}{d\mathbf{n}}$ is a $J \times J$ matrix) and then applying multidimensional integration by parts (using the fact that $f(\mathbf{n}) \to 0$ as $\mathbf{n} \to \partial\mathbf{N}$) then yields:

$$\begin{aligned}
&\int_{\mathbf{N}} [M_c(1 - T'(\mathbf{z}(\mathbf{n}))) + M_{\mathbf{z}}] \mathbf{X}(\mathbf{n}) \nabla_{\mathbf{z}} \tau(\mathbf{z}(\mathbf{n})) f(\mathbf{n}) d\mathbf{n} \\
&= -\int_{\mathbf{N}} \nabla_{\mathbf{n}} \cdot \left[[M_c(1 - T'(\mathbf{z}(\mathbf{n}))) + M_{\mathbf{z}}] \mathbf{X}(\mathbf{n}) f(\mathbf{n}) \left(\frac{d\mathbf{z}}{d\mathbf{n}} \right)^{-1} \right] \tau(\mathbf{z}(\mathbf{n})) d\mathbf{n}
\end{aligned}$$

Thus:

$$\begin{aligned}
DW_T[\tau] = \int_{\mathbf{N}} \Bigg\{ & m(c(\mathbf{n}), \mathbf{z}(\mathbf{n}), \mathbf{n}) f(\mathbf{n}) - [M_c(1 - T'(\mathbf{z}(\mathbf{n}))) + m_{\mathbf{z}}] \vec{\eta}(\mathbf{n}) f(\mathbf{n}) \\
& + \nabla_{\mathbf{n}} \cdot \left[[M_c(1 - T'(\mathbf{z}(\mathbf{n}))) + M_{\mathbf{z}}] \mathbf{X}(\mathbf{n}) f(\mathbf{n}) \left(\frac{d\mathbf{z}}{d\mathbf{n}} \right)^{-1} \right] \Bigg\} \tau(\mathbf{z}(\mathbf{n})) d\mathbf{n}
\end{aligned}$$

Hence:

$$m(c(\mathbf{n}), \mathbf{z}(\mathbf{n}), \mathbf{n}) f(\mathbf{n}) - [M_c(1-T'(\mathbf{z}(\mathbf{n})))+m_{\mathbf{z}}] \vec{\eta}(\mathbf{n}) f(\mathbf{n}) + \nabla_{\mathbf{n}} \cdot \left[[M_c(1 - T'(\mathbf{z}(\mathbf{n}))) + M_{\mathbf{z}}] \mathbf{X}(\mathbf{n}) f(\mathbf{n}) \left(\frac{d\mathbf{z}}{d\mathbf{n}} \right)^{-1} \right]$$

are consistent generalized welfare weights by Theorem 1 because they form the Gateaux derivative of the functional $W(T)$.

□

B.5 Examples with Behavioral Agents

First, suppose that we begin with the following objective functional to evaluate tax policies, where $u(c, z, n)$ is the *decision* utility functional:

$$W(T) = \int_N \Psi(u(c(n), z(n), n), n) f(n) dn \tag{49}$$

Because the envelope theorem holds, we can calculate the Gateaux derivative of $W(T)$ as:

$$DW_T[\tau] = -\int_N u_c(c(n), z(n), n) \Psi_u(u(c(n), z(n), n), n) f(n) dn$$

Setting $\psi(u(c(n), z(n), n), n) = -\Psi_u(u(c(n), z(n), n), n)$, we get that:

$$\psi(u(c(n), z(n), n), n) u_c(c(n), z(n), n)$$

is a set of consistent generalized welfare weights in the form $g(c, z, n)$ by Theorem 1.

Alternatively, suppose that we begin with the following objective functional to evaluate tax reforms, where $\hat{u}(c, z, n)$ is the *experience* utility functional:

$$W(T) = \int_N \Psi(\hat{u}(c(n), z(n), n), n) f(n) dn \tag{50}$$

Suppose that Assumption 1 holds and that we take the Gateaux variation of Equation (50) in the direction of τ (i.e., from T to $T + \epsilon\tau$), yielding:

$$\int_N \left[\Psi_u(\hat{u}(c(n), z(n), n), n) \left[-\hat{u}_c \tau(z(n)) + (\hat{u}_c(1 - T'(z)) + \hat{u}_z) \frac{dz(n)}{d\epsilon} \right] \right] f(n) dn$$

where the envelope theorem does not hold (because individuals maximize u rather than $\hat{u}$). By Equation (6) we can rewrite this as:

$$\int_N \left[\Psi_u(\hat{u}(c(n), z(n), n), n) \left[-\hat{u}_c + (\hat{u}_c(1 - T'(z)) + \hat{u}_z) \left[\xi(n) \tau'(z(n)) + \eta(n) \tau(z(n)) \right] \right] \right] f(n) dn$$

Integrating by parts we get that the Gateaux derivative of Equation (50):

$$\begin{aligned}
&\int_N \left[\Psi_u(\hat{u}(c(n), z(n), n), n) \left[-\hat{u}_c + \left(\hat{u}_c(1 - T'(z)) + \hat{u}_z \right) \eta(n) \right] \right] \tau(z(n)) f(n) dn \\
&- \int_N \frac{d}{dn} \left[\Psi_u(\hat{u}(c(n), z(n), n), n) \left(\hat{u}_c(1 - T'(z)) + \hat{u}_z \right) \xi(n) f(n) \left(\frac{dz}{dn} \right)^{-1} \right] \tau(z(n)) dn \\
&= \int_N \left[\Psi_u(\hat{u}(c(n), z(n), n), n) \left[-\hat{u}_c + \frac{(\hat{u}_c(1 - T'(z)) + \hat{u}_z) \left[u_{cc}(1 - T'(z)) + u_{cz} \right]}{-u_c T''(z) + u_{cc}(1 - T'(z))^2 + 2(1 - T'(z)) u_{cz} + u_{zz}} \right] \right] \tau(z(n)) f(n) dn \\
&- \int_N \frac{d}{dn} \left[\Psi_u(\hat{u}(c(n), z(n), n), n) \left(\hat{u}_c \frac{-u_z}{u_c} + \hat{u}_z \right) \frac{u_c}{u_{cn} \frac{-u_z}{u_c} + u_{zn}} f(n) \right] \tau(z(n)) dn
\end{aligned} \tag{51}$$

where we have used the fact that, by the implicit function theorem applied to the FOC (under u not $\hat{u}$):

$$\frac{dz}{dn} = \frac{u_{cn}(1 - T'(z)) + u_{zn}}{-u_c T''(z) + u_{cc}(1 - T'(z))^2 + 2(1 - T'(z)) u_{cz} + u_{zz}} \tag{52}$$

and that by the FOC:

$$u_c(1-T'(z))+u_z=0$$

Hence, generalized welfare weights associated with the objective function (50), which depends on *experience* utility, are given by:

$$\left[\Psi_u(\hat{u}(c(n),z(n),n),n)\left[-\hat{u}_c+\frac{(\hat{u}_c(1-T'(z))+\hat{u}_z)\left[u_{cc}(1-T'(z))+u_{cz}\right]}{-u_cT''(z)+u_{cc}(1-T'(z))^2+2(1-T'(z))u_{cz}+u_{zz}}\right]\right]f(n)$$
$$-\frac{d}{dn}\left[\Psi_u(\hat{u}(c(n),z(n),n),n)\left(\hat{u}_c\frac{-u_z}{u_c}+\hat{u}_z\right)\frac{u_c}{u_{cn}\frac{-u_z}{u_c}+u_{zn}}f(n)\right] \tag{53}$$

which depend on $T''(z)$ via the first term as well as via the second term because this second term (after differentiating w.r.t. n) depends on $\frac{dz}{dn}$ which, in turn, depends on $T''(z)$ as can be seen in Equation (52). Hence, objective functions which are "utilitarian" in experience utility generate (consistent) generalized welfare weights of the form $g(c,z,n)$ whereas objective functions that are utilitarian in decision utility generate (consistent) generalized welfare weights which depend on $T''(z)$ and therefore cannot be expressed as $g(c,z,n)$.

B.6 Discontinuous Poverty Alleviation Weights

Next, we illustrate Proposition 2 General via an example by constructing consistent generalized welfare weights using "poverty alleviation" *mechanical welfare weights*. The key difference relative to Section 5.1 is that now the "mechanical" weights will be discontinuous (as in the example from Saez and Stantcheva (2016)) in contrast to the continuous mechanical weights illustrated in Section 5.1 (as in the example from Sher (2024)). This means that the smoothness conditions on $M(c,z,n)$ from Proposition 2 do *not* hold. For simplicity of exposition, we will continue to assume Assumption 1 holds and that $f(n)\to 0$ as $n\to\underline{n}$ and $n\to\overline{n}$.

Suppose that a policymaker evaluates the desirability of a mechanical tax perturbation (i.e., ignoring behavioral responses) via the following mechanical welfare weights:

$$-\int_N \mathbb{1}[c\leq\bar{c}]\tau(z(n))f(n)dn \tag{54}$$

where $\bar{c}$ is the poverty line. The mechanical weights $-\mathbb{1}[c\leq\bar{c}]$ capture a poverty alleviation motive insofar as a tax cut is valued only for individuals below the poverty line $\bar{c}$. As in Section 5.1, we adopt the quasi-linear utility specification so that individuals solve:

$$\max_z \quad z-T(z)-\frac{(z/n)^{1+k}}{1+k}$$

To construct consistent generalized welfare weights via Theorem 2 General, we set $M(c,z,n)=\int m(c,z,n)\,dc$ by integrating the mechanical weights with respect to c:

$$M(c,z,n)\equiv -c\mathbb{1}[c\leq\bar{c}]-\bar{c}\mathbb{1}[c>\bar{c}]+K(z,n)$$

for an arbitrary function $K(z,n)$. The next step is to calculate the Gateaux derivative of:

$$\begin{aligned}
&-\int_N M(c,z,n)f(n)dn \\
&=\int_N \left[c(n)\mathbb{1}[c(n)\leq \bar{c}]+\bar{c}\mathbb{1}[c(n)>\bar{c}]-K(z,n)\right]f(n)dn \\
&=\int_{\underline{n}}^{\hat{n}}\left[c(n)-K(z(n),n)\right]f(n)dn+\int_{\hat{n}}^{\overline{n}}\left[\bar{c}-K(z(n),n)\right]f(n)dn
\end{aligned} \tag{55}$$

where $\hat{n}$ is the type n for whom $c(\hat{n})=\bar{c}$ (noting that this type is unique by Assumption 1). Relative to Section 5.1, the key difference is again that $M(c,z,n)$ is not differentiable and so we cannot just differentiate under the integral.

Instead, consider a tax perturbation from T to $T+\epsilon\tau$, recognizing that $c(n)$ depends on the tax schedule via its dependence $z(n)$ and $\hat{n}$ also depends on the tax schedule because this type has first order condition:

$$u_c(\bar{c},z(\hat{n}),\hat{n})(1-T'(z(\hat{n})))+u_z(\bar{c},z(\hat{n}),\hat{n})=0$$

Differentiating Equation (55) w.r.t. ϵ noting that $z(n)$, $c(n)$, and $\hat{n}$ are all functions of ϵ:[30]

$$\begin{aligned}
&-\int_{\underline{n}}^{\hat{n}}\tau(z(n))f(n)dn+\int_{\underline{n}}^{\hat{n}}(1-T'(z(n))\frac{dz}{d\epsilon}f(n)dn-\int_N K_z(z(n),n)\frac{dz}{d\epsilon}f(n)dn \\
&+\left[c(\hat{n})-K(z(\hat{n}),\hat{n})-(\bar{c}-K(z(\hat{n}),\hat{n}))\right]\frac{d\hat{n}}{d\epsilon}
\end{aligned}$$

Because $c(\hat{n})=\bar{c}$ and $\frac{dz}{d\epsilon}$ is given by Equation (15) we can simplify the above expression to:

$$\begin{aligned}
&-\int_N \mathbb{1}[c\leq\bar{c}]\tau(z(n))f(n)dn+\int_{\underline{n}}^{\hat{n}}(1-T'(z(n))\left[\frac{-\tau'(z(n))}{T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}}\right]f(n)dn \\
&-\int_N K_z(z(n),n)\left[\frac{-\tau'(z(n))}{T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}}\right]f(n)dn \\
&=-\int_N \mathbb{1}[c\leq\bar{c}]\tau(z(n))f(n)dn+\int_{\underline{n}}^{\hat{n}}(1-T'(z(n))\left(\frac{dz}{dn}\right)^{-1}\left[\frac{-\tau'(z(n))\frac{dz}{dn}}{T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}}\right]f(n)dn \\
&-\int_N K_z(z(n),n)\left(\frac{dz}{dn}\right)^{-1}\left[\frac{-\tau'(z(n))\frac{dz}{dn}}{T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}}\right]f(n)dn \\
&=-\int_N \mathbb{1}[c\leq\bar{c}]\tau(z(n))f(n)dn+\int_{\underline{n}}^{\hat{n}}\frac{d}{dn}\left[\frac{nf(n)}{1+k}\right]\tau(z(n))dn-\frac{\hat{n}f(\hat{n})}{1+k}\tau(z(\hat{n})) \\
&-\int_N \frac{d}{dn}\left[\frac{K_z(z(n),n)nf(n)}{(1+k)(1-T'(z(n)))}\right]\tau(z(n))dn
\end{aligned} \tag{56}$$

[30] For simplicity, we will assume that $K(z,n)$ is chosen to be smooth.

where we used that $\frac{dz}{dn} = \frac{(1+k)z(n)^k/n^{k+2}}{T''(z(n))+\frac{kz(n)^{k-1}}{n^{k+1}}} = (1+k)(1-T'(z(n))/n$ by Equation (16) and the assumptions that $f(n) \to 0$ as $n \to \underline{n}$ and $n \to \overline{n}$. The key difference then relative to Section 5.1 is that there is a *discrete* weight on the type $\hat{n}$ equal to $\frac{\hat{n}f(\hat{n})}{1+k}$, which represents the impact that the tax schedule has on the number of people who have consumption levels below the poverty line. Thus, in the context of Proposition 2 General:

$$DM_{\hat{T}}[T_\epsilon^{(\epsilon)}]|_{\hat{T}=T=T^{(\epsilon)}} = \int_{\epsilon_1}^{\epsilon_2}\int_{\mathbf{N}}\left\{\mathbb{1}[c(n)\leq \bar{c}]\frac{d}{dn}\left[\frac{nf(n)}{1+k}\right] - \frac{d}{dn}\left[\frac{K_z(z(n),n)nf(n)}{(1+k)(1-T'(z(n))}\right]\right\}T_\epsilon^{(\epsilon)}(z(n))f(n)dnd\epsilon$$
$$-\int_{\epsilon_1}^{\epsilon_2}\frac{\hat{n}f(\hat{n})}{1+k}T_\epsilon^{(\epsilon)}(z(\hat{n}))d\epsilon \tag{57}$$

In other words, the consistent way to evaluate large policy changes in a way that is compatible with "mechanical weights" given by Equation (54) is via:

$$\int_{\epsilon_1}^{\epsilon_2}\int_{\mathbf{N}}\left\{\mathbb{1}[c(n)\leq \bar{c}]\left(-1+\frac{d}{dn}\left[\frac{nf(n)}{1+k}\right]\right) - \frac{d}{dn}\left[\frac{K_z(z(n),n)nf(n)}{(1+k)(1-T'(z(n))}\right]\right\}T_\epsilon^{(\epsilon)}(z(n))f(n)dnd\epsilon$$
$$-\int_{\epsilon_1}^{\epsilon_2}\frac{\hat{n}f(\hat{n})}{1+k}T_\epsilon^{(\epsilon)}(z(\hat{n}))d\epsilon \tag{58}$$

B.7 General Equilibrium Wage Effects

Next, we illustrate how to construct consistent generalized welfare weights in a labor demand/labor supply model where, contrary to much of the optimal taxation literature, the labor demand side of the market is *not* assumed to be infinitely elastic implying that wages are endogenous to the tax schedule. The setup of this model is similar to Sachs et al. (2020) and Bergstrom and Dodds (2026) and much of the analysis is very similar to Bergstrom and Dodds (2026). For simplicity of exposition, we will continue to assume Assumption 1 holds and that $f(n) \to 0$ as $n \to \underline{n}$ and $n \to \overline{n}$.

Consider a population of individuals indexed by a unidimensional type n who choose an income $z = wnl$, where l is labor supply and w is a wage paid on effective effort nl, to maximize a quasi-linear iso-elastic utility function:

$$\begin{aligned} U(n;T,w) &= \max_z \; c - \frac{[z/(nw)]^{1+k}}{1+k} \\ \text{s.t. } & c = z - T(z) + s(n)\pi(w) \end{aligned} \tag{59}$$

where c is again numeraire consumption, $\pi(w)$ represents firm profits, and $s(n)$ represents the share of profits owned by a given type n with $\int_N s(n)f(n)dn = 1$ (this share is exogenous). There is also a single firm that produces the consumption good c by hiring labor to maximize profits. Firm output depends on total hired effective effort, L. Thus,

firm profits are given by $\pi = Y(L) - wL$ where $Y(L)$ is the firm's production function. Market clearing requires:

$$L = \int_N nl(n)dF(n) \tag{60}$$

Next, suppose we are given a set of smooth mechanical weights $m(c, z, n)$. As in Proposition 2, we set $M(c, z, n) = \int m(c, z, n)\, dc$ (noting that $M(c, z, n)$ is only defined up to a constant function $K(z, n)$). Before calculating the Gateaux derivative of

$$-\int_N M(c, z, n)f(n)dn \tag{61}$$

let us first determine how the wage w varies with a tax perturbation; i.e., we want to express $\frac{\partial w}{\partial \epsilon}$ in terms of $\tau(z)$. Multiplying the market clearing condition, Equation 60, by w and implicitly differentiating with respect to ϵ (recognizing that labor demand does not react directly to a change in $\tau(z)$, only indirectly via the changing wage and that $wnl(n) = z(n)$):

$$\frac{\partial w}{\partial \epsilon}L + w\frac{\partial L}{\partial w}\frac{\partial w}{\partial \epsilon} - \int_{\mathbf{N}} \left(\frac{\partial z(n)}{\partial w}\bigg|_\epsilon \frac{\partial w}{\partial \epsilon} + \frac{\partial z(n)}{\partial \epsilon}\bigg|_w \right) dF(n) = 0 \tag{62}$$

We can recover $\frac{\partial z(n)}{\partial w}|_\epsilon$ by implicitly differentiating the individual first order condition with respect to w:

$$\frac{\partial z(n)}{\partial w}\bigg|_\epsilon = \frac{(1+k)\left(\frac{z(n)}{nw}\right)^k \frac{1}{nw^2}}{k\left(\frac{z(n)}{nw}\right)^{k-1}\frac{1}{n^2w^2} + T''(z(n))}$$

Similarly, by implicitly differentiating the individual first order condition with respect to ϵ, we have:

$$\frac{\partial z(n)}{\partial \epsilon}\bigg|_w = \frac{\tau'(z(n))}{-k\left(\frac{z(n)}{nw}\right)^{k-1}\frac{1}{n^2w^2} - T''(z(n))} \equiv \tau'(z(n))\xi(n)$$

The firm's first order condition is that $Y'(L) - w = 0$. Thus, $\frac{\partial L}{\partial w} = \frac{\partial Y'^{-1}(w)}{\partial w}$. Plugging in $\frac{\partial z(n)}{\partial \epsilon}\big|_w \equiv \tau'(z(n))\xi(n)$ to Equaton 62 we have that:

$$\frac{\partial w}{\partial \epsilon}\left[L + w\frac{\partial Y'^{-1}(w)}{\partial w} - \int_N \left(\frac{\partial z(n)}{\partial w}\bigg|_\epsilon\right) dF(n)\right] = \int_N \tau'(z(n))\xi(n)dF(n) \tag{63}$$

Multiplying and dividing by $\frac{dz}{dn}$ (which is given by Equation (16) after replacing n with nw) and applying integration by parts, we find that (denoting $\overline{\frac{\partial z}{\partial w}}\big|_\epsilon \equiv \int_N \left(\frac{\partial z(n)}{\partial w}\big|_\epsilon\right) dF(n)$):

$$\frac{\partial w}{\partial \epsilon} = \frac{\int_N \frac{d}{dn}\left[\frac{nf(n)}{(1+k)(1-T'(z(n)))}\right]\tau(z(n))dn - \frac{nf(n)}{(1+k)(1-T'(z(n)))}\tau(z(n))\big|_{\underline{n}}^{\overline{n}}}{L + w\frac{\partial Y'^{-1}(w)}{\partial w} - \overline{\frac{\partial z}{\partial w}}\big|_\epsilon} \tag{64}$$

Given that we assumed $f(n) \to 0$ as $n \to \underline{n}$ and $n \to \overline{n}$ then:

$$\frac{\partial w}{\partial \epsilon} = \frac{\int_N \frac{d}{dn}\left[\frac{nf(n)}{(1+k)(1-T'(z(n)))}\right]\tau(z(n))dn}{L + w\frac{\partial Y'^{-1}(w)}{\partial w} - \overline{\frac{\partial z}{\partial w}}\big|_\epsilon} \equiv \int_N p(z(n))\tau(z(n))dn \tag{65}$$

for $p(z(n)) = \frac{\frac{d}{dn}\left[\frac{nf(n)}{(1+k)(1-T'(z(n)))}\right]}{L + w\frac{\partial Y'^{-1}(w)}{\partial w} - \overline{\frac{\partial z}{\partial w}}\big|_\epsilon}$

Next, let us take the Gateaux variation of Equation 61 in the direction of $\tau(z)$:

$$\int_N \Bigg\{ \left(\tau(z(n)) - s(n)\pi'(w)\frac{\partial w}{\partial \epsilon} \right) M_c \\ - \left(M_c(1-T'(z(n))) + M_z\right) \left[\left.\frac{dz(n)}{d\epsilon}\right|_w + \left(\frac{z(n)}{nw}\right)^{1+k} \frac{1}{w}\frac{\partial w}{\partial \epsilon} \right] \Bigg\} f(n)dn \tag{66}$$

Substitution in that $\left.\frac{\partial z(n)}{\partial \epsilon}\right|_w = \tau'(z(n))\xi(n)$ and $\frac{\partial w}{\partial \epsilon} = \int_N p(z(n'))\tau(z(n'))dn'$ Equation (66) equals:

$$\int_N \Bigg\{ \left(\tau(z(n)) - s(n)\pi'(w)\int_N p(z(n'))\tau(z(n'))dn' \right) M_c \\ - \left(M_c(1-T'(z(n))) + M_z\right) \left[\tau'(z(n))\xi(n) + \left(\frac{z(n)}{nw}\right)^{1+k} \frac{1}{w}\int_N p(z(n'))\tau(z(n'))dn' \right] \Bigg\} f(n)dn \tag{67}$$

Performing integration parts on the term involving $\tau'(z(n))$ exactly as in the proof to Proposition 2, this equals:

$$\int_N \left\{ m(c,z,n)f(n) + \frac{d}{dn}\left[[M_c(1-T'(z(n))) + M_z]\xi(n)f(n)\left(\frac{dz}{dn}\right)^{-1} \right] \right\} \tau(z(n))dn - \\ \int_N p(z(n'))\tau(z(n'))dn' \int_N \left\{ s(n)\pi'(w)m(c,z,n) + \left(M_c(1-T'(z(n))) + M_z\right)\left(\frac{z(n)}{nw}\right)^{1+k}\frac{1}{w} \right\} f(n)dn \tag{68}$$

Letting $q(z(n)) \equiv p(z(n)) \times \int_N \left\{ s(n)\pi'(w)m(c,z,n) + \left(M_c(1-T'(z(n))) + M_z\right)\left(\frac{z(n)}{nw}\right)^{1+k}\frac{1}{w} \right\} f(n)dn$, we get that the following is a consistent set of generalized welfare weights, by Theorem 1, with general equilibrium wage effects:

$$m(c,z,n)f(n) + \frac{d}{dn}\left[[M_c(1-T'(z(n))) + M_z]\xi(n)f(n)\left(\frac{dz}{dn}\right)^{-1} \right] + q(z(n)) \tag{69}$$

Intuitively, $q(z(n))$ captures the impact of a tax perturbation at income $z(n)$ on wage, $p(z(n))$, multiplied by the consumption impact this generates for types, $\int_N s(n)\pi'(w)m(c,z,n)f(n)dn$ plus the behavioral impact that wage changes have on the implied objective function, $\int_N \left(M_c(1-T'(z(n))) + M_z\right)\left(\frac{z(n)}{nw}\right)^{1+k}\frac{1}{w}f(n)dn$. In the notation of Proposition 2 General then:

$$DM_{\hat{T}}[T_\epsilon^{(\epsilon)}] = \int_N \left\{ \frac{d}{dn}\left[[M_c(1-T'(z(n))) + M_z]\xi(n)f(n)\left(\frac{dz}{dn}\right)^{-1} \right] + q(z(n)) \right\} T_\epsilon^{(\epsilon)}(z(n))dn \tag{70}$$

B.8 Bunching and Jumping Example

Next, we illustrate how to apply Proposition 2 General when we relax three assumptions relative to Proposition 2: (1) the tax schedule is everywhere differentiable, (2) individuals have a unique optima under all $T \in \mathcal{T}$, and (3) type space $\mathbf{N}$ and the choice space $\mathbf{Z}$ are the same dimension. We will assume individuals differ in terms of two parameters $(n, v) \in [\underline{n}, \overline{n}] \times [\underline{v}, \overline{v}] = N \times V$ and have a utility function $u(c, z, n, v)$. We suppose that $u(c, z, n, v)$ satisfies the single crossing property in n so that $\frac{-u_z(c,z,n,v)}{u_c(c,z,n,v)}$ is decreasing in n for all c, z, v. This ensures that $z(n, n)$ is monotonic in n for a fixed value of v. Also, suppose that $f(n) \to 0$ as $n \to \underline{n}$ and $n \to \overline{n}$. Next, suppose that $\mathcal{T}$ is a set of tax schedules which are continuous, infinitely differentiable except at some finite set of points, only have a finite number of types n with multiple optima for every v, and are compactly supported.

Suppose we are given some set of mechanical weights $m(c, z, n)$. As in Proposition 2, we set $M(c, z, n) = \int m(c, z, n)\, dc$ (noting that $M(c, z, n)$ is only defined up to a constant function $K(z, n)$). We want to calculate the Gateaux derivative of:

$$W(T) = -\int_V \int_N M(c(n, v), z(n, v), n, v) f(n, v) dn dv \tag{71}$$

Let us first consider the impacts of a tax perturbation from $T(z)$ to $T(z) + \epsilon\tau(z)$ for individuals who have a unique optima $z(n, v)$ at which $T(z)$ is smooth. For these individuals, we have the individual first order condition:

$$u_1(z - T(z) - \epsilon\tau(z), z, n, v)\,(1 - T'(z) - \epsilon\tau'(z)) + u_2(z - T(z) - \epsilon\tau(z), z, n, v) = 0$$

Lemma 3 of Bergstrom and Dodds (2021) shows that, given the assumptions above on the tax schedules $T \in \mathcal{T}$, almost all individuals have second order conditions holding strictly, which means we can apply the implicit function theorem:

$$\begin{aligned} \frac{\partial z}{\partial \epsilon}(n, v) &= \frac{u_c \tau'(z) + [u_{cc}(1 - T'(z)) + u_{cz}]\,\tau(z)}{u_{cc}(1 - T'(z))^2 + 2(1 - T'(z))u_{cz} + u_{zz}} \\ &\equiv \xi(n, v)\tau'(z(n, v)) + \eta(n, v)\tau(z(n, v)) \end{aligned} \tag{72}$$

where $\xi(n, v) \equiv \frac{u_c}{u_{cc}(1-T'(z))^2+2(1-T'(z))u_{cz}+u_{zz}}$ and $\eta(n, v) \equiv \frac{[u_{cc}(1-T'(z))+u_{cz}]}{u_{cc}(1-T'(z))^2+2(1-T'(z))u_{cz}+u_{zz}}$. Next, for individuals with a unique optima where $T'(z)$ exists but higher derivatives don't exist, we will just "ignore" this set of people because they are (by assumption) measure zero and so don't impact the integral as long as they move continuously in response to tax perturbations (and they must because they have a unique optima, meaning that their new optima must be nearby as a result of standard continuity arguments). For individuals with a unique optima who bunch at kink points, these individuals do not change their income in response to small tax perturbations because they are at a corner solution to begin with so that they strictly prefer this income level to all others; hence,

$\frac{\partial M(c,z,n)}{\partial \epsilon} = -M_c(c,z,n)\tau(z(n))$ for these individuals.[31] For expositional purposes, let us just assume that there is a single z where $T(z)$ is not differentiable (the argument extends immediately for some finite set but the notation is more tedious) and denote the set of bunching types $[n_1, n_2]$.

Finally, we consider the behavioral responses of the types with multiple optima. Again, let's assume for the sake of exposition, that there is a single type $n_3(v)$ for each v with multiple optima. Let $z^-(v)$ and $z^+(v)$ denote the lower and upper optimal incomes for type $n_3(v)$. Dropping the v argument, $z^-(v)$ and $z^+(v)$ satisfy the following indifference condition:

$$u(z^+ - T(z^+) - \epsilon\tau(z^+), z^+, n_3, v) = u(z^- - T(z^-) - \epsilon\tau(z^-), z^-, n_3, v) \tag{73}$$

We can calculate how the indifferent individual (for each v) changes with the tax schedule by applying the implicit function theorem to Equation 73:[32]

$$\frac{\partial n_3}{\partial \epsilon} = \frac{u_c(z^+ - T(z^+), z^+, n_3, v)\tau(z^+) - u_c(z^- - T(z^-), z^-, n_3, v)\tau(z^-)}{u_n(z^+ - T(z^+), z^+, n_3, v) - u_n(z^- - T(z^-), z^-, n_3, v)} \equiv \frac{u_1^+\tau(z^+) - u_1^-\tau(z^-)}{u_n^+ - u_n^-} \tag{74}$$

Let us then calculate the Gateaux variation of Equation (71) after splitting up (where we assume WLOG that $n_3 > n_2$):

$$\begin{aligned} W(T) = &- \int_{\underline{n}}^{n_1(v)} M(c(n,v), z(n,v), n, v) f(n) dn - \int_{n_1(v)}^{n_2(v)} M(c(n,v), z(n,v), n, v) f(n) dn \\ &- \int_{n_2(v)}^{n_3(v)} M(c(n,v), z(n,v), n, v) f(n) dn - \int_{n_3(v)}^{\overline{n}} M(c(n,v), z(n,v), n, v) f(n) dn \end{aligned} \tag{75}$$

Taking the Gateaux variation of Equation (75) yields:

$$\begin{aligned} &\lim_{\epsilon \to 0} \frac{W(T + \epsilon\tau) - W(T)}{\epsilon} = \\ &\int_V \Bigg\{ \int_{\underline{n}}^{n_1(v)} \left[\tau(z(n,v)) M_c - \left(M_c(1 - T'(z(n)) + M_z \right) \frac{dz(n,v)}{d\epsilon} \right] f(n|v) dn \\ &+ \int_{n_1(v)}^{n_2(v)} \tau(z(n,v)) M_c f(n|v) dn \\ &+ \int_{n_2(v)}^{\overline{n}} \left[\tau(z(n,v)) M_c - \left(M_c(1 - T'(z(n)) + M_z \right) \frac{dz(n,v)}{d\epsilon} \right] f(n|v) dn \\ &- \left(M(c(z^-(v)), z^-(v), n, v) - M(c(z^+(v)), z^+(v), n, v) \right) f(n_3(v)|v) \frac{\partial n_3}{\partial \epsilon}(v) \Bigg\} f(v) dv \end{aligned} \tag{76}$$

The last term of Equation 76 results from applying the Leibniz integral rule (recognizing

[31] Footnote 21 of Bergstrom and Dodds (2021) discusses this point in more detail.

[32] Note, we have implicitly assumed that the individual first order condition holds for $n_3(v)$ at both $z^+(v)$ and $z^-(v)$ in deriving Equation (74). However, this assumption can be dropped without changing Equation (74); see Appendix A.6 of Bergstrom and Dodds (2021).

that this is the only such term arising from differentiating the limits of integration via the Leibniz integral rule because $z(n,v)$ is continuous in n at the other limits of integration). Plugging in the value of $\frac{\partial z(n,v)}{\partial \epsilon}$ from Equation 72:

$$\begin{aligned}
&\lim_{\epsilon \to 0} \frac{W(T+\epsilon\tau) - W(T)}{\epsilon} = \\
&\int_V \Bigg\{ \int_{\underline{n}}^{n_1(v)} \left[\tau(z(n,v))M_c - \left(M_c(1-T'(z(n)) + M_z\right)\left[\xi(n,v)\tau'(z(n,v)) + \eta(n,v)\tau(z(n,v))\right]\right] f(n|v)dn \\
&+ \int_{n_1(v)}^{n_2(v)} \tau(z(n,v))M_c f(n|v)dn \\
&+ \int_{n_2(v)}^{\overline{n}} \left[\tau(z(n,v))M_c - \left(M_c(1-T'(z(n)) + M_z\right)\left[\xi(n,v)\tau'(z(n,v)) + \eta(n,v)\tau(z(n,v))\right]\right] f(n|v)dn \\
&- \left(M(c(z^-(v)), z^-(v), n, v) - M(c(z^+(v)), z^+(v), n, v)\right) f(n_3(v)|v)\frac{\partial n_3}{\partial \epsilon}(v) \Bigg\} f(v)dv
\end{aligned} \tag{77}$$

Multiplying and dividing by $\frac{\partial z(n,v)}{\partial n}$ (which exists again by Lemma 3 of Bergstrom and Dodds (2021) for almost all individuals) and then integrating by parts, yields:

$$\begin{aligned}
&\lim_{\epsilon \to 0} \frac{W(T+\epsilon\tau) - W(T)}{\epsilon} = \\
&\int_V \Bigg\{ \int_{\underline{n}}^{n_1(v)} \left[M_c - \left(M_c(1-T'(z(n)) + M_z\right)\eta(n,v)\right]\tau(z(n,v))f(n|v)dn \\
&+ \int_{\underline{n}}^{n_1(v)} \frac{d}{dn}\left[\left(M_c(1-T'(z(n)) + M_z\right)\xi(n,v)\left(\frac{dz(n,v)}{dn}\right)^{-1} f(n|v)\right]\tau(z(n,v))dn \\
&- \left[\left(M_c(1-T'(z(n)) + M_z\right)\xi(n,v)\left(\frac{dz(n,v)}{dn}\right)^{-1} f(n|v)\tau(z(n,v))\right]\Bigg|_{n_1(v)} + \int_{n_1(v)}^{n_2(v)} M_c\tau(z(n,v))f(n|v)dn \\
&+ \int_{n_2(v)}^{\overline{n}} \left[M_c - \left(M_c(1-T'(z(n)) + M_z\right)\eta(n,v)\right]\tau(z(n,v))f(n|v)dn \\
&+ \int_{n_2(v)}^{\overline{n}} \frac{d}{dn}\left[\left(M_c(1-T'(z(n)) + M_z\right)\xi(n,v)\left(\frac{dz(n,v)}{dn}\right)^{-1} f(n|v)\right]\tau(z(n,v))dn \\
&+ \left[\left(M_c(1-T'(z(n)) + M_z\right)\xi(n,v)\left(\frac{dz(n,v)}{dn}\right)^{-1} f(n|v)\tau(z(n,v))\right]\Bigg|_{n_2(v)} \\
&- \left(M(c(z^-(v)), z^-(v), n, v) - M(c(z^+(v)), z^+(v), n, v)\right) f(n_3(v)|v)\frac{u_1^+\tau(z^+(v)) - u_1^-\tau(z^-(v))}{u_n^+ - u_n^-} \Bigg\} f(v)dv
\end{aligned} \tag{78}$$

Hence, from Equation (78), we can see that Proposition 2 General applies with:

$$
\begin{aligned}
&DM_{\hat{T}}[\tau] = \\
&\int_V \Bigg\{ \int_{\underline{n}}^{n_1(v)} \left[-\left(M_c(1 - T'(z(n)) + M_z \right) \eta(n,v) \right] \tau(z(n,v)) f(n|v) dn \\
&+ \int_{\underline{n}}^{n_1(v)} \frac{d}{dn} \left[\left(M_c(1 - T'(z(n)) + M_z \right) \xi(n,v) \left(\frac{dz(n,v)}{dn} \right)^{-1} f(n|v) \right] \tau(z(n,v)) dn \\
&- \left[\left(M_c(1 - T'(z(n)) + M_z \right) \xi(n,v) \left(\frac{dz(n,v)}{dn} \right)^{-1} f(n|v) \tau(z(n,v)) \right] \Bigg|_{n_1(v)} \\
&+ \int_{n_2(v)}^{\overline{n}} \left[-\left(M_c(1 - T'(z(n)) + M_z \right) \eta(n,v) \right] \tau(z(n,v)) f(n|v) dn \\
&+ \int_{n_2(v)}^{\overline{n}} \frac{d}{dn} \left[\left(M_c(1 - T'(z(n)) + M_z \right) \xi(n,v) \left(\frac{dz(n,v)}{dn} \right)^{-1} f(n|v) \right] \tau(z(n,v)) dn \\
&+ \left[\left(M_c(1 - T'(z(n)) + M_z \right) \xi(n,v) \left(\frac{dz(n,v)}{dn} \right)^{-1} f(n|v) \tau(z(n,v)) \right] \Bigg|_{n_2(v)} \\
&- \left(M(c(z^-(v)), z^-(v), n, v) - M(c(z^+(v)), z^+(v), n, v) \right) f(n_3(v)|v) \frac{u_1^+ \tau(z^+(v)) - u_1^- \tau(z^-(v))}{u_n^+ - u_n^-} \Bigg\} f(v) dv
\end{aligned}
\tag{79}
$$

where it should be clear that $DM_{\hat{T}[\tau]}$ is both linear and bounded (assuming that all of the terms in Equation (79) are bounded).

B.9 Proposition Sher Without Assumption 1

Proposition Sher illustrates that in sets $\mathcal{T}$ where tax schedules are "well-behaved", consistent weights of the form $g(c, z, n)$ must be structurally utilitarian. In other words, Proposition Sher only illustrates that $g(c, z, n)$ must be structurally utilitarian for (c, z) that are chosen under a tax schedule $T \in \mathcal{T}$ for some region $\mathcal{T}$ of well-behaved tax schedules. Hence, Proposition Sher does not restrict $g(c, z, n)$ at (c, z) that are chosen by individuals only under tax schedules that are not "well-behaved" as in Assumption 1. We will continue to suppose that $n \in N$ and $z \in Z$ are scalars and make the following assumption:

Assumption 3. *Suppose that $\mathcal{T}$ is the set of continuous tax schedules. Also suppose that $u(z(1 - T') + R, z, n)$ is globally concave for all scalars T', R and that the single crossing property holds in n so that $\frac{-u_z(c,z,n)}{u_c(c,z,n)}$ is decreasing in n for all c, z. Also suppose that $u_z(c, z, n)$ and $u_c(c, z, n)$ are finite for all $(c, z) \in \mathbb{R}^2_{++}$.*

We can prove a version of Proposition Sher replacing Assumption 1 with the (significantly) weaker Assumption 3:

Proposition 3. *Suppose Assumption 3 holds. Generalized welfare weights $g(c, z, n)$ are consistent (when comparing tax schedules in $\mathcal{T}$) only if they are equal to $g(c, z, n) =$*

$u_c(c, z, n) \cdot \psi(u(c, z, n), n)$ *for some function* ψ.

Proof. First, note that the mapping $n \mapsto z$ is bijective, other than potential bunching at $z = 0$, under all tax schedules $T(z) = z(1 - T') - R$; this is a standard fact, see, e.g., Lemma 1 from Bergstrom and Dodds (2021). Let us suppose that we have a set of generalized welfare weights $g(c, z, n)$. Starting from a tax schedule $T(z) = z(1-T') - R$, we will evaluate the desirability of a reform in the direction of γ via:

$$\begin{aligned} &\int_N g(c(z(n)), z(n), n)\, \gamma(z(n))\, f(n)\, dn \\ &= \int_{\underline{n}}^{\hat{n}} g(c(0), 0, n)\, \gamma(0)\, f(n)\, dn + \int_{\hat{n}}^{\overline{n}} g(c(z(n)), z(n), n)\, \gamma(z(n))\, f(n)\, dn \end{aligned} \tag{80}$$

where $z(n)$ and $c(z(n))$ denote the optimal choices of z, $c = z - (1 - T')z + R$ for a particular type n, and types $[\underline{n}, \hat{n}]$ locate at 0 whereas types $[\hat{n}, \overline{n}]$ do not. Now suppose that we consider how $\int_N g(c(z(n)), z(n), n)\, \gamma(z(n))\, f(n)\, dn$ varies if we perturb the tax schedule from $T(z)$ to $T(z) + \epsilon\tau(z)$, again noting that $z(n)$ and $c(z(n))$ are both functions of the tax schedule. For everyone who does not earn 0, we can apply the implicit function theorem to determine the impacts of a tax perturbation on each individual's optimal income and thereby use Equation (6).

Next, almost all individuals that locate at 0 under a given tax schedule $T(z) = z(1 - T') - R$ will continue to locate at 0 for any small tax (potentially non-linear) perturbation because they have a unique optima and are at a corner solution; see discussion in Section 3 of Bergstrom and Dodds (2021) for more details. Thus:

$$\frac{d}{d\epsilon} \int_{\underline{n}}^{\hat{n}} g(c(z(n)), z(n), n)\, \gamma(z(n))\, f(n)\, dn = -\int_{\underline{n}}^{\hat{n}} g(c(0), 0, n)\, \gamma(0)\tau(0)\, f(n)\, dn$$

Differentiating Equation (80) with respect to ϵ thereby yields:

$$\begin{aligned} &\frac{d}{d\epsilon} \int_N g(c(z(n)), z(n), n)\, \gamma(z(n))\, f(n)\, dn \\ &= \int_N \left[g_c\, c'(z)\, \frac{dz(n)}{d\epsilon}\, \gamma(z(n)) + g_z\, \frac{dz(n)}{d\epsilon}\, \gamma(z(n)) + g\, \gamma'(z(n))\, \frac{dz(n)}{d\epsilon} \right] f(n)\, dn \\ &- \int_{\hat{n}}^{\overline{n}} g_c\, \gamma(z(n))\tau(z(n))\, f(n)\, dn \\ &- \int_{\underline{n}}^{\hat{n}} g(c(0), 0, n)\, \gamma(0)\tau(0)\, f(n)\, dn \end{aligned} \tag{81}$$

The second two integrals of Equation (81) are symmetric in τ and γ, so we focus on

the first integral, substituting $\frac{dz(n)}{d\epsilon} = \xi(n)\,\tau'(z) + \eta(n)\,\tau(z)$ from Equation (6) yielding:

$$\begin{aligned}
&\int_N \left[g_c\, c'(z)\, \frac{dz(n)}{d\epsilon}\, \gamma(z(n)) + g_z\, \frac{dz(n)}{d\epsilon}\, \gamma(z(n)) + g\, \gamma'(z(n))\, \frac{dz(n)}{d\epsilon} \right] f(n)\, dn \\
&= \int_N \Big(\xi(n)\, \tau'(z(n)) + \eta(n)\, \tau(z(n)) \Big) \Big[g_c\, \gamma(z(n)) \\
&+ g_z\, \gamma(z(n)) + g\, \gamma'(z(n)) \Big] f(n)\, dn
\end{aligned} \tag{82}$$

From here, identical logic as in Section 3 illustrates that Equation 82 is symmetric in τ and γ if and only if:

$$u_c \Big[g_c(c,z,n)\, \frac{-u_z}{u_c} + g_z(c,z,n) \Big] - \left[u_{cc} \frac{-u_z}{u_c} + u_{cz} \right] g(c,z,n) = 0 \tag{83}$$

Hence, because the set $\mathcal{T}$ of continuous functions is convex, Equation (83) must be satisfied (by Theorem 1) at all (c,z) chosen under a linear tax schedule for a given n. Next, we will show that all (c,z) chosen by an arbitrary n under any arbitrary tax schedule are also chosen under *some* linear tax schedule. To see this, first note that if an individual optimally chooses a given (c,z) under an arbitrary (non-linear) $T(z)$ that is differentiable at (c,z), then they would choose that same (c,z) under a linearized schedule that goes through that (c,z) by the global concavity assumption in Assumption 3.[33] Finally, if $T(z)$ is not differentiable at the given $(c,z) \in \mathbb{R}^2_{++}$, then note that (c,z) is optimal for type n under the linearized tax schedule that goes through (c,z) with slope $-\frac{u_z(c,z,n)}{u_c(c,z,n)}$ (which exists by Assumption 3).

Thus, at *every* (c,z) chosen by n under *any* $T \in \mathcal{T}$, Equation (83) must hold. From here, we just apply the exact logic from Proposition Sher, noting that now we have shown $g(c,z,n) = u_c(c,z,n) \cdot \psi(u(c,z,n),n)$ for each n and for all (c,z) chosen under *any* non-linear tax schedule, rather than just (c,z) chosen by n under T in the set of well-behaved tax schedules.

□

[33] Note, in particular, this global concavity assumption rules out things like discrete costs of working as this implies that utility is not globally concave (and this can be seen from, for example, and indifference curve diagram).