Inequality as an Externality: Consequences for Tax Design*

Morten Nyborg Støstad† and Frank Cowell‡

Abstract

This paper introduces an income inequality externality into the classical optimal non-linear income taxation model, noting that such an externality arises if income inequality affects pertinent societal outcomes. There are two new mathematical terms in the optimal taxation formula, corresponding to a Pigouvian tax and an increased social taste for (in)equality. The two new terms vary more across the distribution than standard externality terms and always influence optimal rates in the same direction at the top of the distribution. As a result, optimal top tax rates are largely driven by the size of the inequality externality. The overall tax schedule is strongly reliant on the type of inequality externality (pre-tax income, post-tax income, utility) and on the externality structure. The findings indicate a new equality-relevant dimension to the optimal policy problem, provide a theoretical basis for previously unsupported tax policies, and show that the size of the inequality externality could be considered a crucial economic variable.
JEL Codes: H21, H23, D62, D63

*We thank Stéphane Gauthier, Marc Fleurbaey, Emmanuel Saez, Daniel Waldenström, Olof Johansson-Stenman, Fredrik Carlsson and Karine Nyborg for helpful comments and discussions. We have also benefited from suggestions from Etienne Lehmann, Marie Young Brun, Max Lobeck, Elif Cansu Akoğuz, Stefanie Stantcheva, Thomas Blanchet, Antoine Bozio, François Fontaine, Damián Vergara, Thomas Piketty, and seminar participants at the Paris School of Economics, UC Berkeley, the University of Oslo, the GT Économie de la Fiscalité, ECINEQ 2019, the 2021 EEA Congress, LAGV 2021, the 2021 IIPF Annual Congress, and the 2021 NTA Annual Conference on Taxation. Version: July 27, 2022.

†Paris School of Economics, 48 Boulevard Jourdan, 75014 Paris, France. Phone: +33766142152. Email: morten.stostad@psemail.eu (corresponding author).

‡London School of Economics, Houghton Street, London, W2CA 2AE, UK. Email: f.cowell@lse.ac.uk.

I Introduction

It is a common feature of tax design that the tax system has an effect on economic inequality. However, inequality may also play an important role as an *input* to the optimal design problem. Suppose that economic inequality causally affects at least one public good in a welfare-relevant way – such as the amount of political polarization in society, the crime rate, or the amount of innovation.[1] This paper explores two natural consequent questions, both related to the optimal taxation framework. First, how should such effects be taken into account in a utility-based framework? Second, when they are taken into account, what are the changes to optimal policy design – and in particular to the standard Mirrlees (1971) framework of optimal non-linear income taxation?

The standard way to introduce inequality concerns into welfare frameworks is through the social welfare function (SWF), so to change social preferences. We instead postulate that individual preferences or well-being may be sensitive to the level of inequality due to how inequality affects society. If so, inequality itself is an externality – and using a standard inequality-averse or Rawlsian social planner is generally not a sufficient way to model the problem. For instance, individuals will normally choose their own labor effort without taking into account how this effort impacts the global level of income inequality. If income inequality affects pertinent societal factors, there is an externality dimension to this choice that is not well-modeled by simply discounting individual utility. Instead, a more appropriate solution is the inclusion of an inequality metric in the individual's utility function.[2]

Notice that such an externality can exist even when individuals are fully rational and selfish, unlike models with traditional other-regarding preferences (ORP), the other primary reason to include inequality into the utility function. As an example, imagine a perfectly self-interested individual in a society where income inequality increases crime through an Becker (1968)-type opportunity cost framework.[3] Suppose income inequality and thus crime increases, and that the person's bike is stolen as a consequence. The individual experiences a negative shock and would undoubtedly, absent any other changes, prefer the prior (more equal) state of the world. Thus, if inequality leads to more crime, inequality should enter her utility function. A similar argument could be made with any other pertinent variable affected by inequality, such as political polarization or corruption – and is noticeably less problematic than the arguments which motivate ORP.[4]

To illustrate our framework we take the Mirrlees (1971) model and introduce various types of inequality terms in the individual's utility function. As a widely used model describing optimal

[1] There are potentially many such channels, and a large associated empirical literature. In Section V we create simple micro-foundations for how economic inequality could affect crime, trust, innovation, economic growth, political capture, social unrest, political polarization, and the supply of public goods.

[2] We also discuss the benefits of such a simplification as compared to modeling each potential externality channel individually.

[3] In effect, higher economic inequality leads more individuals to have relatively low market wages and relatively high potential economic gains from criminal activity (such as stealing a bike).

[4] It is not straightforward to assume that the social planner should account for personal sentiments in the optimal taxation problem – see Harsanyi (1977) or Goodin (1986), for example. Mathematically, though, these two concepts are similar, and models with an ORP term could alternatively be motivated through a concern for inequality's externality effects. We discuss further in Section II.

income taxation (OIT), the Mirrlees model represents both an important pillar of public economics and a compelling example of how standard economic models rely on the no-externality assumption. The standard model focuses instead on two equality-relevant factors; diminishing marginal utilities of income and social welfare weights. We explore several types of inequality externalities in this framework, focusing on a post-tax income (consumption) inequality externality. This specification corresponds to an unstudied type of consumption externality where each individual's effect on the externality depends both on her income and on her position in the income distribution. The main contrast to the externalities found in the ORP or the environmental economics literature is that the marginal externality of each individual's income depends on that individual's location in the income distribution.

Keeping to conventions in these literatures we introduce separability such that the introduction of the externality does not change the individual's maximization problem. The introduced modifications are thus to the social planner's incentives for taxation. Whereas the standard Mirrlees model only focuses on the revenue effects of taxation, we must also consider the *equality effects* of taxation. This leads to a "double dividend" of taxation when a tax increase has both a revenue and equality benefit, such as at the top under a negative inequality externality when the tax rate is below the revenue-maximizing tax rate.

These new equality effects are driven by two forces: the well-known behavioral and mechanical effects of a tax increase described in Saez (2001). The behavioral effect describes how agents at and potentially above the tax bracket change their work decisions when their tax burden increases. The mechanical effect describes the increase in tax revenue from any agent above the tax bracket. With an inequality externality, these two consequences change not only the tax revenue collected and redistributed, but also income inequality itself. This has a welfare effect that differs from the standard revenue case.

To explain this, consider first the standard case. From the social planner's point of view, the mechanical effect leads to more redistribution and is thus an incentive for higher marginal tax rates. The behavioral responses distorts individuals' choices and reduces their output, and is thus a disincentive for higher marginal tax rates. The two effects always oppose each other, which illustrates the classical trade-off.

In contrast, the two *equality* effects do not always oppose each other. The mechanical effect is simple, as it always decreases inequality. This is because a lump-sum redistribution away from all agents above any given tax bracket decreases inequality. The behavioral responses, however, increase or decrease income inequality depending on the location of the tax hike. At the bottom, a behavioral shift away from work effort increases income inequality. At the top, a behavioral shift away from work effort *reduces* income inequality. The behavioral responses to taxation can thus be welfare-positive in certain situations.

This creates a distributional asymmetry. The mechanical effect always decreases inequality, whereas the behavioral effect increases inequality at the bottom and decreases inequality at the top. The two equality effects thus oppose each other at the bottom of the distribution and harmonize

at the top. This leads to our main result; top marginal tax rates are particularly sensitive to the inequality externality. This is true in both the theoretical framework and in the numerical simulations, is a direct consequence of the rationale presented above, and holds regardless of whether inequality is a *positive* or *negative* externality. An intuitive way to explain this result follows. First, the location of the tax-payer is crucial when evaluating equality effects, and the social planner's incentive to change incomes is larger towards the ends of the distribution.[5] Second, given that only top income-earners can be specifically targeted with marginal tax rates, the social planner can more easily affect inequality through top income-earners than through bottom income-earners.

In order to conduct numerical simulations we estimate the inequality externality by three distinct methods, primarily through survey data from Carlsson et al. (2005), and find a negative median inequality externality. This median externality value results in the optimal top marginal tax rate increasing from 68% to 85%. Given standard parameter values and reasonable magnitudes of the externality we find a very wide range of possible optimal marginal top tax rates, ranging from negative (<0%) marginal top tax rates if inequality is a positive externality to extremely high (>90%) marginal top tax rates if inequality is a negative externality. Optimal lower- and middle-class marginal tax rates are less affected.

The range of optimal top tax rates is wider than what is supported by standard parameter values in the no-externality case, where optimal top marginal tax rates usually range between 50% and 80%. This narrow range in the classical case is partly because every standard SWF converges to the same optimal top tax rate in the Mirrlees model.[6] This has arguably decreased the focus on the "equality dimension" in optimal top tax rate analysis – which we show can be highly relevant as long as inequality itself affects the individual. The *individual* concerns that arise from an inequality externality thus differ from the *social* concerns modeled by an inequality-averse SWF. We also naturally find optimal top tax rates above the revenue-maximizing Laffer rate, as direct equality effects imply that the social planner might trade off some revenue for changed equality levels. Our results provide a theoretical basis for previously unsupported policy arguments, such as the high post-war top marginal tax rates in the United States and the United Kingdom (if inequality is a negative externality), or the low contemporary top marginal tax rates in many countries (if inequality is a positive externality and even if the social planner is Rawlsian).

The theoretical predictions also change significantly from the original literature, as expected from the introduction of a consumption externality. Our OIT model generates unambiguously more progressive (regressive) tax- and transfer systems under a negative (positive) inequality externality, where more progressive is defined as a lower level of consumption inequality. The classical U-shape found in the optimal income taxation literature is fragile to the inclusion of a negative (but not positive) inequality externality, which is a consequence of the new welfare-positive effects of behavioral responses. Generally, the overall structure of the optimal tax schedule depends strongly

[5]Note that this contrasts to revenue effects, where the location of the tax-payer is secondary to the magnitude of potential revenue – in the extreme case, the Rawlsian min-max, "one dollar is one dollar" as long as it is not taken from the very bottom of the distribution.

[6]This is because the benefit of additional income near the top approaches zero in most standard models, either due to income-decreasing social welfare weights or diminishing marginal utilities of income.

on the type and magnitude of the inequality externality. We solve the externality problem in both the mechanism design and small-perturbation frameworks, examine more types of externalities than is standard in the literature (in pre-tax income, post-tax income, and utility) and explore several different types of externality structures (different inequality metrics, in effect). We also show that various such inequality externality channels can be micro-founded and discuss how an inequality externality can have broad consequences for the many literatures that rely on the no-externality assumption.

Related Literature

Various types of income or consumption externalities have been explored in the Mirrlees model.[7] With regards to the externality literature, the main mathematical novelty in this work is to introduce a consumption externality into the Mirrlees model whose *marginal* effect depends on both the individual's income and on the full income distribution.[8] With a negative inequality externality, for example, the bottom-agent's income will cause a positive externality and the top-agent's income will cause a negative externality. This contrasts to the generally flat externality structure in most of the literature outlined below.

Kanbur and Tuomala (2013) and Aronsson and Johansson-Stenman (2020) are closest to our main analysis of a consumption inequality externality. Kanbur and Tuomala (2013) introduces relative income concerns into a mechanism design-based Mirrlees framework through a flat negative externality on average consumption, and find similar analytic results to our mechanism design case. We add to this analysis by examining an externality with a changing marginal effect across the distribution which leads to distributionally distinct externality-based incentives for taxation. This, combined with our more extensive exploration of different types of distributional externalities, leads to novel insights relating to the effect of inequality's societal effects on optimal income taxation (and particularly on optimal *top* taxation). We also examine inequality externalities in both the small perturbation framework and mechanism design framework, which leads to new insight about the origin of the two new terms in the optimal taxation formula found in Kanbur and Tuomala (2013) and their distributional effects. In short, these two terms correspond to the behavioral and mechanical effects of a tax increase and depend on, respectively, the marginal externality in the tax bracket and the average marginal externality above the tax bracket. Aronsson and Johansson-Stenman (2020) discusses various types of other-regarding preferences including classical Fehr and Schmidt (1999)-type inequality aversion in a three-agent OIT model. We further this analysis by using a broader set of inequality-related specifications in a full continuous Mirrlees-type model, which allows for more complete dynamics across the income distribution.

There is also a related literature on "dirty goods" from environmental economics, including

[7]The Mirrlees model is the standard starting point in the optimal income taxation literature. See for example Diamond and Mirrlees (1971); Atkinson and Stiglitz (1976); Mirrlees (1976); Diamond (1998); and Saez (2001). Non-analytical solutions to the standard problem are found in Blundell and Shephard (2011) and Aaberge and Colombino (2013).

[8]There are also natural motivational differences between this work and those discussed below, as they largely relate to various types of other-regarding preferences or environmental externalities.

Sandmo (1975), Bovenberg and van der Ploeg (1994), and Cremer et al. (1998), among others. This literature examines second-best solutions to externality problems and finds, in general, two new terms in the optimal taxation formula; as Cremer et al. (1998) describes them, a Pigouvian modification and an expression related to the optimal tax on non-externality goods. Our solutions follow a similar pattern with two new terms in the optimal taxation formula; one Pigouvian term and one term that modifies the existing social welfare weights. Each individuals' marginal effect on the externality is held constant in this literature, as one unit of pollution is always one unit of pollution regardless of who emits it. This contrasts to our work, where an additional unit of consumption affects the inequality externality differently depending on the distributional location of the individual who benefits from the additional unit.

Other related works in the ORP literature include Boskin and Sheshinski (1978) and Oswald (1983), both of which introduce relative income concerns in optimal income taxation frameworks as flat negative average consumption externalities in respectively linear and non-linear frameworks. Aronsson and Johansson-Stenman (2018) examines the effect of inequality aversion on income taxation, focusing on the first-best case and Pareto-optimal taxation. Further works on relative income concerns and optimal taxation include Persson (1995); Aronsson and Johansson-Stenman (2008); and Aronsson and Johansson-Stenman (2015). The potential for a direct focus on distributional concerns in the OIT model is also found in Kanbur et al. (1994) in terms of poverty concerns in the social welfare function and Prete et al. (2016), which employs a non-welfarist approach and piecewise taxation to minimize post-tax income inequality. We differ from these two approaches by introducing the distributional term directly into the agent's utility function.

A recent literature on rent-seeking is also conceptually related to this paper through the externality dimension. Piketty et al. (2014) introduces tax avoidance and compensation bargaining into the standard model and establishes the relevant elasticities in the case of such externality-inducing behavior, focusing particularly on top income taxation. Rothschild and Scheuer (2016) explores a model with a traditional sector and a rent-seeking sector, where the social planner must correct for the rent-seeking externalities without directly observing the sector difference. Lockwood et al. (2017) considers the allocation of talented individuals under the assumption that productivity externalities range from positive to negative from low-paying to high-paying jobs. Both of these latter papers include a dimension of imperfect targeting, unlike this work, which dampens the externality benefit of income taxation and changes the scope of the analysis. In general, our work differs from the rent-seeking literature by considering income differences *themselves* as a negative externality, regardless of origin. This naturally increases the role of income taxation in the optimal policy solution.

We are also, to the best of our knowledge, the first work to analytically solve first-order conditions for a utility function which includes a comprehensive and endogenous income inequality term in an optimal taxation model. Most inequality metrics are analytically intractable, and we believe this issue is partly why the mathematical analysis in this paper has not been previously been done. We achieve this by using a family of rank-specific income inequality metrics similar to

those concurrently developed in Simula and Trannoy (2022a,b). As in these papers we exploit the rank-invariancies between wage-earning ability, pre-tax income, and post-tax income when there is no bunching to simplify the analytical problem. We also illustrate generalized inequality metrics that function as analytically tractable versions of top-income share metrics.

Finally, there is a related literature to the concept of inequality as an externality. The idea which was first mentioned by Thurow (1971). Thurow's paper itself consists of a short discussion on how the First Welfare Theorem fails if the income distribution is a pure public good. Alesina and Giuliano (2011) briefly considers both ORP and how inequality might affect consumption and thus utility, and Kaplow (2010) mentions that the economic distribution could affect variables such as crime, which could imply optimal taxation effects. Continuing the analysis from these works, we discuss the potential utility impact of inequality through non-consumption channels, create micro-foundations for various ways in which inequality can affect pertinent societal factors, and delve deeper into the theoretical implications of the concept.

Other papers on topics related to inequality as an externality include Pauly (1973) and Ashworth et al. (2002), both of which model redistribution itself as a public good. We also note specific externality-related theoretical frameworks in Lindbeck (1985), which discusses the consequences of inequality on macroeconomic policies, Anbarci et al. (2009), which suggest an externality effect from rising inequality through an increase in traffic fatalities, and Rueda and Stegmueller (2016), which considers crime as a negative externality of inequality. Dimick et al. (2018) briefly notes that beliefs around inequality's effects on society can affect redistributive preferences. Lobeck and Støstad (2022) examines U.S. citizens' beliefs in various inequality externality channels and notes a causal effect of these beliefs on redistributive preferences. There are also large empirical literatures on inequality's potential effects on various societal outcomes, such as crime, corruption, social unrest, economic growth, innovation, individual health, political polarization, trust, and more. Examples of relevant reviews can be found in Rufrancos et al. (2013) for crime, Cingano (2014) for economic growth, and Bergh et al. (2016) for individual health. These literatures are too expansive to even briefly encompass here, however.

This paper is organized as follows. Section II examines the concept of inequality as an externality and how it differs from other ways in which distributional concerns are modeled in conventional OIT analysis. Section III incorporates an inequality externality in a standard OIT model and investigates the impact of the externality on optimal tax rates. Section IV conducts numerical simulations in the OIT model. Section V discusses the inequality externality concept generally, creating micro-foundations and discussing other potential mathemetical formulations. Section VI concludes.

II Inequality and Social Welfare: An Externality Approach

Suppose that economic inequality causally affect non-consumption goods individuals care about, the relevant of which we capture in an vector $\overrightarrow{\Psi}$. The most natural example of such goods are public goods (such as the amount of political polarization), but they might also be individual-specific (such

as individual health) – see Section V.A for a further discussion on various channels. Suppose further that economic inequality can affect individual consumption x_i (Alesina and Giuliano, 2011),[9] and that individuals may have other-regarding preferences over economic inequality $\bar{\theta}$ (Cooper and Kagel, 2016). The individual's utility can thus be written as,

$$U_i(x_i(\bar{\theta}), \bar{\theta}, \overrightarrow{\Psi}(\bar{\theta}), ...). \tag{1}$$

Detailed information on each component in the specification (1) is unlikely to be available; such complexity would also be unrealistically cumbersome for most models. We propose a simplification, noting that the separate contributions are less important than the overall impact of inequality in the utility function. The specification (1) could be written more compactly as the simplified form:

$$\tilde{U}_i(\tilde{x}_i, \bar{\theta}, ...) \tag{2}$$

where $\tilde{U}_i$ is the modified utility function, $\tilde{x}_i$ is the portion of individual income which is not determined by economic inequality, and the term $\bar{\theta}$ represents the total impact on the individual from the inequality externality.

As this is a simplification, it may seem like an imperfect way to analyze implications of inequality's externality effects. In short, the idea is similar to why one introduces consumption directly into individual utility. Often, individuals care about the indirect effects of higher consumption rather than (or in addition to) higher consumption in itself. And as is the case with consumption, modeling every way in which inequality could affect individual utility is generally not a practical solution.[10] We discuss this further in Appendix A.

The simpinlification we discuss in (1) and (2) does not rely on the existence of any of the three components we show in (1). The externality exists as long as one of the components is deemed policy-relevant. For instance, individuals could be wholly self-serving and still have a utility function that is strongly dependent on economic inequality if economic inequality affects some pertinent public good. Given the many philosophical problems with introducing ORP into the welfarist framework, this scenario may often be appropriate, and we focus on it for the remainder of the article. Before we continue, however, it is worth noting that as expressed in the form (2), the inequality externality as a whole is mathematically equivalent to an ORP term in the utility function. It follows that many of the results from the ORP literature can be applied to our framework.

We also note that the inequality externality could be heterogeneous. Various inequality externality channels could affect people in different ways, perhaps depending on their individual income or their position in the income distribution. In this work we focus on a homogenous inequality ex-

[9] Alesina and Giuliano (2011) discusses how income inequality could affect the income of individuals through three channels; externalities in education, crime and property rights, and incentive effects. One could also imagine that individual income is affected through some of the other channels we discuss in this work (political capture, innovation, social unrest, and so on).

[10] Even when it is, the solutions might have many of the same mathematical properties.

ternality for simplicity, although most of our theoretical work can be extended to a heterogeneous inequality externality without loss of generality.[11] In particular, the analysis conducted in Section III and IV is essentially unchanged under a heterogeneous externality (see Section III.B).

The model also needs the choice of an inequality metric. Such metrics are often analytically difficult. To simplify the optimal tax problem we use a family of absolute inequality metrics of the form,

$$\bar{\theta}\left(\boldsymbol{x}, F\right) = \int_{\underline{x}}^{\bar{x}} \kappa(x')x'dF(x'), \tag{3}$$

where x' is the individual's resource which corresponds to the type of inequality measured (post-tax income x in the main specification), $\kappa(x')$ is the weight of the agent with this amount of resources in the inequality metric (positive near the top and negative near the bottom of the distribution), and $F(x')$ is the cumulative distribution function of x'. Absolute inequality metrics are used to keep scale invariance. The weight $\kappa(x')$ depends only on the *rank* of the individual in the distribution. This is notable, as the rank in pre-tax income z_i and post-tax income x_i of individual i when second-order conditions hold (which we assume). In other words, we have that $F(z_i) = F(x_i)$ and $\kappa(z_i) = \kappa(x_i)$, which allows us to write the inequality metric as follows,

$$\bar{\theta}\left(\boldsymbol{x}, F\right) = \int_{\underline{z}}^{\bar{z}} \kappa(z)x(z)dF(z), \tag{4}$$

where z is pre-tax earnings. This trick combined with the rank-specific inequality metric solves an otherwise difficult endogeneity problem when solving for a consumption inequality externality.[12]

For the main discussion it is useful to have a specific inequality metric in mind; for this purpose we use a particular form of the (absolute) Gini coefficient in post-tax income, where the weight κ_G is,

$$\kappa_G(z) = 2F(z) - 1 \tag{5}$$

taken from Cowell (2000). Expressions 4 and 5 shows that the absolute Gini can be calculated as a sum of weighted incomes in the population, where the weight $\kappa_G(z)$ depends only on the rank of the agent in the pre-tax income distribution. In the numerical simulations we will also explore other post-tax income inequality metrics based on other types of rank-specific weights $\kappa(z)$ where $\int_0^\infty \kappa(z)dF(z) = 0$, such as those in the Lorenz (Aaberge, 2000) or S-Gini families (Donaldson and Weymark, 1980).

It is worth mentioning that the true inequality metric is likely to be a function of several different inequality metrics. To show an example of this, suppose that inequality's effect on crime is dependent on relative poverty and that inequality's effect on political capture is dependent on

[11] For simplicity we also avoid other concerns that, while nonetheless important, complicate a first approach to an inequality externality. These issues include questions related to perceived inequality, inequality in different regions, (non-)meritocratic inequality, and so on.

[12] The same trick with wage-earning ability n is used in the mechanism design solution.

the proliferation of top incomes. Both relative poverty θ_p and top income proliferation θ_t are distributional metrics, which we represent in our framework by the distributional weights κ_p and κ_t, the rest of the inequality metric being determined by Equation 4. Take then an example with separability and homogeneity in these externality effects, such as in the simple example of $U = x - \eta_p\theta_p - \eta_t\theta_t$ where η_p and η_t indicate externality magnitudes. The total externality effect is $-\eta_p\theta_p - \eta_t\theta_t = -(\eta_p + \eta_t)\int_{\underline{z}}^{\bar{z}}\left(\frac{\eta_p}{\eta_p+\eta_t}\kappa_p(z) + \frac{\eta_t}{\eta_p+\eta_t}\kappa_t(z)\right)x(z)dF(z)$. The modified inequality metric is thus $\theta_{true} = \int_{\underline{z}}^{\bar{z}}\left(\frac{\eta_p}{\eta_p+\eta_t}\kappa_p(z) + \frac{\eta_t}{\eta_p+\eta_t}\kappa_t(z)\right)x(z)dF(z)$, a weighted sum of the two inequality metrics. As such, the inequality metrics we use could be seen as a combination of potentially several externality-determining inequality metrics.

We will now make a short detour to discuss how the inequality externality fits into the general utilitarian framework. In such models the social planner maximizes a social welfare function consisting of some weighted sum of every individual's utility. In addition to the inequality externality, there are thus two other channels through which inequality-related concerns can enter into the formulation of social welfare comparisons. These are (i) the cumulative effect of diminishing marginal utilities of income, and (ii) social welfare weights. We posit that these three channels are mathematically and intuitively distinct. Except for special cases,[13] they cannot be mathematically interchanged, and each channel is caused by a particular theoretical mechanism. We summarize this framework in Table I, where the relevant channel is shown.

Table I

The Three Welfarist Consequences of Inequality

	Diminishing marginal utility of income	Social welfare weights	Inequality externality
Formulation	$\int_i g_i U_i(\underbrace{x_i}, \bar{\theta}, ...)di$	$\int_i \underbrace{g_i} U_i(x_i, \bar{\theta}, ...)di$	$\int_i g_i U_i(x_i, \underbrace{\bar{\theta}}, ...)di$
Causes	The decreased value of a dollar with increased income	Societal considerations of fairness, philosophical concerns	The societal effects of inequality, other-regarding preferences

Note: The three channels through which inequality could influence welfarist modeling. For each channel the key expression is highlighted by an underbrace. Individual consumption is denoted by x_i, resource inequality is denoted by $\bar{\theta}$, and the utility-based social welfare weight is denoted by g_i.

An economic inequality externality cannot be fully captured – or approximated – by a combination of utility functions with only standard individualist parameters (say, income and work effort) and utility-based social welfare weights. This follows the same logic as for why other externalities cannot be captured by such utility functions combined with social welfare weights; when an externality is present, individual labor decisions are socially sub-optimal. The sub-optimality of individual decisions cannot be approximated by suitable social welfare weights, as discounting

[13] We discuss this further in Section V.C.

utility is dissimilar from discounting *income*, and also cannot be approximated by modifications to the simple individualistic parameters as such modifications would have to depend on other agents' incomes. As a result, neglecting externality issues in individualistic frameworks leads to potentially misleading policy conclusions if inequality does in fact affect society. We show a simple proof in the case of social welfare weights and the inequality externality in Appendix B.I. We also discuss income-based social welfare weights in Section V.C.

We will now show the effect of introducing three types of inequality externalities – pre-tax income, post-tax income, and utility – into the Mirrlees (1971) framework. We discuss the concept of an inequality externality further in Section V.

III Optimal Income Taxation: Theory

We consider the second-best solution for a non-linear optimal income taxation schedule with a continuum of individuals in the presence of an inequality externality. The inequality externality is formalized as an inequality term $\bar{\theta}$ in the utility function $U(x_i, h_i, \bar{\theta})$. The main discussion will be for a post-tax income (consumption) inequality externality; extensions for pre-tax income inequality and utility inequality are in Section IV.B. We will solve the problem in both a mechanism design framework in the vein of Diamond (1998) and in a small-perturbations framework in the vein of Saez (2001).

III.A Mechanism design

We first solve the problem in a mechanism design framework, where we fully specify the utility function as,

$$U(x, h, \bar{\theta}) = u(x) - V(h) - \Gamma(\bar{\theta}), \tag{6}$$

where u is the utility of consumption (after-tax income) x, $V(h)$ is the disutility of work h, and Γ is disutility from post-tax consumption inequality $\bar{\theta}$ (a society-wide parameter, indicated by the overbar). The functions $u(x)$, $V(h)$ and $\Gamma(\bar{\theta})$ are continuous and second-order differentiable in their arguments. The function $u(x)$ is strictly concave in x, $V(h)$ is strictly convex in h, and $\Gamma(\bar{\theta})$ has no restriction. We also have that $u_x > 0$ and $V_h > 0$ where subscripts indicate partial derivatives. Equation (6) assumes that agents are homogeneous, with identical individual utility functions. There are a continuum of agents along the wage-earning ability n, with density $f(n)$ and a cumulative distribution function $F(n)$.

Agents do not take their own effect on income inequality into account when making labor decisions, as their effect on the inequality metric is negligible for their own optimization problem in a continuum of agents.[14] However, their actions have welfare-pertinent effects as the change in

[14] As we use a continuum of agents, this effect is indeed negligible in our model. Furthermore, the assumption is theoretically supported by Dufwenberg et al. (2011), which finds that individuals' demands are independent of other allocations given a separability condition that is satisfied here.

income inequality affects every other agent. Note that due to this assumption (and separability), the individual choice of (x, h) does not require that the individual is aware of or estimates the magnitude of the inequality externality.

1 The General Solution We present the full solution to this general framework in Appendix C. The resulting non-linear optimal tax rates are,

$$\frac{t}{1-t} = \frac{\zeta_n u_{x(n)}}{f(n)n} \int_n^\infty \left[\frac{1}{u_{x(p)}} - \frac{W_{U(p)}}{\lambda} \right] dF(p) + \frac{\gamma}{\lambda} \left[\kappa(n) + \frac{\zeta_n u_{x(n)}}{f(n)n} \int_n^\infty \frac{\kappa(p)}{u_{x(p)}} dF(p) \right], \tag{7}$$

where the solution is split into two halves; the standard result from Diamond (1998) and the two new externality terms.[15] Here u_x is the marginal utility of consumption, $W_{U(p)}$ is the derivative of the SWF, λ is the Lagrange multiplier on the government's budget constraint, $\kappa(n)$ is the weight of the individual at wage-earning ability n in the inequality metric,[16] $\zeta_n = \frac{V_{hh}h}{V_h} + 1$ is a term closely related to the inverse compensated elasticity of labor,[17] and γ is the shadow price of inequality such that

$$\gamma = \lambda \frac{\int_0^\infty \frac{\Gamma_{\bar{\theta}}}{u_x} f(n) dn}{1 - \int_0^\infty \frac{\Gamma_{\bar{\theta}}}{u_x} \kappa(n) f(n) dn}, \tag{9}$$

where $\Gamma_{\bar{\theta}}$ is the partial derivative of the inequality disutility function Γ. This parameter indicates the shadow price of the inequality constraint expressed in units of public funds. A negative inequality externality implies a positive $\Gamma_{\bar{\theta}}$, and thus a positive γ. To rephrase, this is the unsurprising result that equality itself has a cost in a world with a negative inequality externality. If $\Gamma(\bar{\theta}) = 0$, as in the standard case when the inequality externality does not exist, then $\gamma = 0$, and the solution in Equation 7 becomes identical to that in Diamond (1998). If we assume a linear inequality externality of the form $\Gamma(\theta) = \eta\theta$, then $\frac{\gamma}{\lambda} = \eta$.[18]

The externality introduces two new terms; (i) a Pigouvian term, $\frac{\gamma}{\lambda}\kappa(n)$, measuring both the size of the externality itself in terms of public funds ($\frac{\gamma}{\lambda}$) and the contribution of the individuals at the given tax bracket to the externality ($\kappa(n)$, which changes sign across the distribution), and (ii) $\frac{\gamma}{\lambda} \frac{\zeta_n u_{x(n)}}{f(n)n} \int_n^\infty \frac{\kappa(p)}{u_{x(p)}} dF(p)$, which indicates a change to the redistributive benefit of the tax – in effect modifying the social welfare weights. Beyond standard Mirrleesian parameters, this latter

[15] By denoting the part of the optimal tax function found in Diamond (1998) as $\frac{t_i}{1-t_i}$, we can isolate and evaluate the effect of the inequality externality as

$$\frac{t}{1-t} = \frac{t_i}{1-t_i} + \frac{\gamma}{\lambda} \left[\kappa(n) + \frac{\zeta}{f(n)n} \int_n^\infty \frac{u_{x(n)}}{u_{x(p)}} \kappa(p) dF(p) \right]. \tag{8}$$

[16] In the absolute Gini, $\kappa(z) = 2F(z) - 1$ and $\overline{\kappa}(z) = F(z)$.

[17] With quasi-linear preferences, $\zeta = \frac{1}{E_L} + 1$.

[18] With a squared inequality externality, which we discuss specifically in Appendix C.II, the term in the utility function is $\eta(\bar{\theta} - \theta_{opt})^2$ and the MRS becomes $2\eta(\bar{\theta} - \theta_{opt})$, which implies that the effect of the externality on the optimal tax schedule would be dependent on the distance from the optimal inequality level θ_{opt} .

term depends on both the size of the externality in terms of public funds $\frac{\gamma}{\lambda}$ and a measure similar to the total externality weight above the tax bracket, $\int_n^\infty \frac{\kappa(p)}{u_{x(p)}} dF(p)$.

This solution illustrates both similarities and differences between our approach and the standard Mirrlees externality literature. In Kanbur and Tuomala (2013), for example, where the externality is a flat negative consumption externality, there are also two new terms to the Diamond (1998) formula; a Pigouvian term and a social welfare weight modification. However, as the marginal externality effect in Kanbur and Tuomala (2013) is constant across the distribution, the analytical modification to the tax schedule is relatively independent of the location of the tax bracket. This is not true in our specification. Equation 7 illlustrates that, when consumption inequality is an externality – or when the consumption externality is dependent on the location of the individual in the income distribution, more generally – the modification to optimal marginal tax rates is also strongly dependent on the location of the tax bracket in the distribution. This location-dependence can be seen in both the marginal externality effect of the agent *in* the tax bracket (κ in the first term), and in the average marginal externality of all agents *above* the tax bracket ($\bar{\kappa}$ in the second term).

III.B Small-perturbations framework

To significantly simplify the intuition behind the above solution we now turn to the small-perturbation framework from Saez (2001). Instead of fully specifying a functional form of individual utility we suggest using the marginal rate of substitution between post-tax income inequality and individual income, $\eta_i = MRS_{x_i\bar{\theta}} = -\frac{dU_i/d\bar{\theta}}{dU_i/dx_i}$. This η_i measures how much consumption the individual would give up for or pay for one unit decrease in the relevant inequality metric. If $\eta_i = 0 \,\forall\, i$ we return to the standard case.

For the main specification we set η to be constant for all agents and income levels. This corresponds to a homogeneous inequality externality and assumes that the (absolute) inequality metric affects utility proportionately to the effect of consumption on utility. Our approach implicitly assumes separability in inequality and income such that individuals' work decisions are independent on the level of income inequality.

We note that, due to this separability, heterogeneous inequality externalities could easily be introduced. In this case the net social welfare weight of the externality, determined by $\bar{\eta} = \int_i g_i \eta_i di / \int_i g_i di$ where g_i is the social welfare weight, is the policy-determining variable.[19]

Although a fully specified utility function is not necessary in this framework, it may be useful to note that a special case of Equation 6 such that

$$U(x, h, \bar{\theta}) = log\left(x - v(h) - \eta\bar{\theta}\right) \tag{10}$$

fits these criteria for some disutility function of hours worked $v(h)$. Referring to the previous section, this specification implies that the now constant η corresponds to the shadow price of inequality $\frac{\gamma}{\lambda}$.

[19]This was also true in the previous section, where a heterogeneous inequality externality would only change the equation for $\frac{\gamma}{\lambda}$ to include individual-specific $\Gamma_{\bar{\theta}}$.

1 Intuition and Optimal Marginal Tax Schedules We can summarize the small perturbation method as follows. There are two main channels through which a tax increase affects incomes and thus social welfare; the behavioral and mechanical effects from the classical OIT literature. The behavioral effect captures how a small tax increase leads each agent located at that tax bracket (and potentially above, if there are income effects) to shift their work decision towards leisure. The mechanical effect captures how every agent above the tax bracket is taxed more in the absence of any behavioral response. These two channels both affect both revenue and post-tax income inequality.

In the literature, the key effect of each channel is that on tax revenue. More revenue allows more redistribution, which is generally positive due to either diminishing marginal utilities of income or decreasing social welfare weights in income. The behavioral response represents a tax revenue loss, while the mechanical effect represents a tax revenue gain. The two terms together represent a revenue collection trade-off, the sufficient statistics of which can be empirically estimated, and together form the basis for the calculation of the revenue-maximizing tax rate. We will discuss these consequences as *revenue effects*. In non-Rawlsian SWFs there is also a pertinent welfare loss from the agents above the tax bracket who have their individual incomes reduced – this effect dampens, but cannot cancel, the revenue-based benefit of the mechanical effect.[20]

These two channels (the behavioral and mechanical) also impact post-tax income inequality directly. This is not considered welfare-relevant in traditional models. The mechanical effect gathers income from those above a certain tax bracket and redistributes this income as a flat dividend to all individuals.[21] This always reduces income inequality, regardless of the tax bracket in question. The behavioral responses, meanwhile, reduce individuals' work effort and thus their income. This increases or decreases post-tax income inequality depending on the location of the tax increase. At the bottom, behavioral responses increase income inequality. At the top, behavioral responses decrease income inequality. The uneven impact of the behavioral response is a key difference between the traditional revenue effects and the new equality impacts and why our novel policy implications are disproportionately localized at the top. The summary of this discussion is in Table II.

Let us examine how this affects the optimal marginal tax rate. There are five pertinent welfare impacts of a small tax increase $d\tau$ in our framework. Three of these are well-known from the previous literature and discussed in Saez (2001); the effect of the behavioral responses on revenue (dB), the effect of the mechanical effect on revenue (dM), and the effect of the mechanical effect on the welfare of those above the tax bracket (dW). Two equality consequences are new in our work;

[20] The revenue benefit will always equal or exceed the welfare loss of these agents due to the assumption of social welfare weights that are non-increasing in income. Generally, this channel – represented as dW below – does not change the relevant intuition.

[21] Any change from such a flat dividend would be equivalent to changing the marginal tax schedule. Such flat transfers do not lead to any change in the *absolute* inequality metrics we use; thus we can focus on where post-tax income is reduced. If we were to use *non-absolute* inequality metrics (where flat income increases change the relevant statistic), the upcoming intuition would be largely the same. There would only be one minor difference; the behavioral channel would be somewhat less inequality-reducing due to the reduction in average income that follows from the behavioral responses. Overall, focusing on non-absolute inequality metrics is problematic due to a lack of scale invariance.

Table II

The Effects of a Small Tax Increase on Revenue R and Inequality $\bar{\theta}$

		Bottom incomes	Middle incomes	Top incomes
Behavioral response	Revenue effect†	← Decreases R →		
	Inequality impact†	Increases $\bar{\theta}$	Small / no change to $\bar{\theta}$	Decreases $\bar{\theta}$
Mechanical effect	Revenue effect‡	← Increases R →		
	Inequality impact⅂	← Decreases $\bar{\theta}$ →		

Note: The table describes the effect each channel exerts on inequality $\bar{\theta}$ and tax revenue R through a small marginal tax increase in the specified distributional location.

†: The behavioral response always decreases revenue, as individuals in the tax bracket shift away from work into leisure.

†: The behavioral response changes the work decision of the individuals in the tax bracket, which changes incomes. A tax increase on the bottom decreases the bottom agents' incomes, which increases inequality. A tax increase on the middle decreases the middle agents' incomes, with little to no inequality effect. A tax increase decreases the top agents' incomes, which decreases inequality.

‡: The mechanical effect always increases revenue, as individuals above the tax bracket have a higher average tax rate yet do not change their work decisions.

⅂: The mechanical effect always decreases inequality, as it redistributes a fixed amount of income from every individual above the bracket equally to every individual.

the effect of the behavioral responses on inequality ($d\bar{\theta}_B$) and the effect of the mechanical effect on inequality ($d\bar{\theta}_M$). At the optimum, the sum of the welfare effect of these five changes must equal to zero when making a small change around the optimum:

$$dM + dB + dW + d\bar{\theta}_M + d\bar{\theta}_B = 0 \tag{11}$$

The three standard welfare channels are simple. dM is always positive, as the tax increase leads to more revenue (and thus increased welfare from redistribution). This is dampened by the always negative dW, which signifies the income loss (and thus welfare loss) from agents above the tax bracket. The effect of the behavioral responses dB is also always negative, as the tax increase distorts work decisions and thus reduces tax revenue, which reduces the amount of redistribution to the bottom (and thus welfare) with no other benefits.

We now turn to the equality effects. Assume inequality is a negative externality for simplicity.[22] The mechanical effect always reduces inequality; as inequality is a negative externality, this implies a welfare benefit. The sign of $d\bar{\theta}_M$ is thus always positive, similar to the revenue case. The effect of the behavioral responses on inequality, however, changes across the distribution. Near the bottom, where inequality is increased, a negative inequality externality implies a negative sign of $d\bar{\theta}_B$. Around the middle, where inequality changes only marginally, the sign of $d\bar{\theta}_B$ goes from negative to positive. And near the top, where inequality decreases, the sign of $d\bar{\theta}_B$ is positive. In other words, the behavioral effect can be welfare-positive in our framework. This contrasts with the

[22]For a positive externality simply swap the signs on $d\bar{\theta}_M$ and $d\bar{\theta}_B$.

always negative dB and illustrates the trade-off between choosing equality levels and maximizing tax revenue.

We can now consider the sign change as compared to the standard optimal top marginal tax rates. At the bottom, where $d\bar{\theta}_M$ and $d\bar{\theta}_B$ are in opposition (regardless of whether the externality is positive or negative), the welfare effect of a tax increase through the externality dimension is ambiguous. The change to the optimal marginal tax rate due to the externality is thus also ambiguous. At the top, where the signs of $d\bar{\theta}_M$ and $d\bar{\theta}_B$ harmonize – both are positive (negative externality) or both are negative (positive externality) – the change to the optimal tax rates is unambiguous. Under a negative (positive) inequality externality there are unambiguously higher (lower) welfare benefits from increasing the marginal tax rate as compared to the standard case. Compared to the standard case, it follows that resulting top optimal rates are higher with a negative consumption inequality externality and lower with a positive consumption inequality externality. These are general results that do not depend on most of the assumptions we use for simplicity.[23]

We will now show the full optimal non-linear marginal tax rates $\tau(z)$ at earnings z, which are found by inserting known variables for the five terms in Equation 11 and solving for $\tau(z)$. The full derivation is presented in Appendix D, and the solution is presented below;

$$\tau(z) = \frac{1 + \Upsilon(z) - \bar{G}(z)}{1 + \Upsilon(z) + \alpha(z)\epsilon(z) - \bar{G}(z)}. \tag{12}$$

This differs from the standard Saez (2001) result by the term $\Upsilon(z)$. This new term is defined as $\Upsilon(z) = \eta\alpha(z)\epsilon(z)\kappa(z) + \eta\bar{\kappa}(z)$, and again consists of two parts (corresponding to those discussed in the preceding section – see Appendix D for derivation). The magnitude of the post-tax income inequality externality η is present in both. If η is large and positive, inequality is a significant negative externality (a public bad). If it is negative, inequality is a positive externality (a public good). We also use several of the standard parameters from the optimal taxation literature, both in $\Upsilon(z)$ and in the tax formula as a whole: the local Pareto parameter $\alpha(z) = \frac{zf(z)}{1-F(z)}$, the elasticity of earnings $\epsilon(z)$ (with respect to $1 - \tau(z)$), and the average social welfare weight above z denoted by $\bar{G}(z)$.[24] We will now further discuss the two new terms in Equation 12, assuming no income effects for simplicity.

The behavioral response: A Pigouvian tax The first term, $\eta\alpha(z)\epsilon(z)\kappa(z)$, comes from $d\bar{\theta}_B$ in Equation 11 and represents the behavioral responses of the individuals who are located at income z.[25] These agents work less due to the tax increase. The classical consequence is that tax revenue is reduced no matter the location of the tax increase. The equality impact, on the other hand, is conditional on the location of the individual. Unlike in the traditional case, this implies a

[23]This holds under any standard SWF and with income effects. Under the assumptions we use to find Equation 12 we can also note that optimal marginal tax rates above the median wage always increase (decrease) as compared to the standard case given a negative (positive) inequality externality.

[24]$\alpha(z) = \frac{zf(z)}{1-F(z)}$ is a distributional measure which becomes constant in a Pareto distribution. In the Rawlsian min-max framework, $\bar{G}(z) = 0$. See Saez (2001) for further discussion on these variables.

[25]Agents above z do not change their labor choice due to the assumption of no income effects. The intuition is similar with income effects.

potentially positive welfare consequence of the behavioral responses in many tax brackets. The new term incentivizes individuals who make socially suboptimal labor choices to substitute into leisure, keeping their utility relatively high.[26]

The term corresponds to a Pigouvian tax designed to correct the individual's socially suboptimal labor decision, and can be called a first-best motive for taxation. This suboptimality differs in magnitude and direction based on the position of the individual, and thus the optimal tax change from this term has different signs across the distribution. As an example, if we are examining an agent near the top in a negative inequality externality framework, their unbiased labor choice is skewed towards increasing individual income at a social cost. As $\kappa(z) > 0$ and $\eta > 0$, the optimal marginal tax rate on the agent is thus higher than in a no-externality framework; the new term makes the individual internalize part of the cost their high income places on society. Similarly, if we are in a positive externality framework such that $\eta < 0$, the agent will be subsidized to internalize the positive effect their increased income has on society.[27]

The term is affected by four parameters. First, how agent at income z affects inequality, represented by their weight in the inequality metric $\kappa(z)$. If the agent has a larger effect on the pertinent inequality metric, the optimal tax effect is likewise increased. Subsequently this term is large at the ends of the distribution (working in opposite directions at the top and bottom). Second, how inequality affects other agents, represented by the externality magnitude η. If other agents are significantly affected by inequality, the tax change will be larger. Third, the degree to which agents substitute away from work when taxed, represented by the elasticity $\epsilon(z)$. If agents substitute more to leisure, the equality impact of the tax increase is stronger. It follows that increase of optimal tax rates is largest when elasticities are *high*. Fourth, the total amount of agents at the tax bracket z, represented by the distributional term $\alpha(z)$. If there are more agents in the tax bracket, such that $\alpha(z)$ is large, there is a greater inequality impact and the optimal tax changes are larger.[28]

These last two factors imply that the standard intuition from the revenue channel – where a high elasticity and a high $\alpha(z)$ leads to a low tax rate – is partially reversed in our framework. In particular we draw attention to the role of the earnings elasticity. In the standard framework, high elasticities imply that the state should keep tax rates low to collect what little revenue they can. In our case, the state might instead prefer to place high tax rates (or subsidies) at the ends of the distribution to increase or decrease inequality as they see fit.

This Pigouvian term invalidates three classic results from the literature based on Mirrlees (1971) noted by Sadka (1976) and Seade (1977) – (i) that the optimal marginal tax rate at the top should be zero,[29] (ii) that the optimal marginal tax rate at the bottom should be zero, and (iii) that the

[26]This does not imply that the social planner wants to punish certain individuals. While the social marginal welfare of *income* can be negative, the social marginal welfare of *utility* is never negative, all else equal (upholding the Pareto principle).

[27]We note that this term exists specifically due to our choice of an *income* inequality externality. If the externality was in terms of utility, the behavioral response would not change the externality and the term would not exist.

[28]The local Pareto parameter $\alpha(z) = \frac{zf(z)}{1-F(z)}$ can be understood as a measure of the relative strength of the mechanical effect and behavioral response. The numerator amplifies the behavioral channel and the denominator amplifies the mechanical channel. Technically, part of the term comes from $d\bar{\theta}_B$ and part from $d\bar{\theta}_M$.

[29]Reducing the income of the top-earner has become a social cost or benefit in itself, and should be a subsidy or

optimal marginal tax rate is bounded between zero and one. These original results are fragile, generally no longer hold when consumption externalities are introduced, and change with many small modifications to the model – see Stiglitz (1982) and Saez (2001) for examples. As such, these changes are not very surprising. Still, the modifications to the classic OIT results are intuitively appealing given the simplicity of the inequality externality. One could see these previously controversial results as an intrinsic limitation of the Mirrlees (1971) model, where economic equality in itself has no value to individuals.

The mechanical effect: An increased taste for (in)equality The second term, $\eta\bar{\kappa}(z)$, is from the mechanical effect on the agents located above income z. As these agents' average tax rates increase, their post-tax incomes decrease. The classical consequence of this response is that tax revenue is increased, which is true (almost) no matter the location of the tax increase. The equality impact functions similarly and decreases absolute inequality by definition (almost) no matter where the tax increase occurs. The sole exceptions to these two statements are where no effective revenue is gathered; at the very top, where there are no agents above, and at the very bottom, where every agent is above. In every other case, the mechanical effect increases revenue and decreases inequality.

How much this affects optimal marginal tax rates depends on the average weight of the agents above the tax bracket in the inequality metric ($\bar{\kappa}(z)$) as well as how valuable or costly reductions to inequality are (η) and how many agents are above the bracket (which contributes to $\alpha(z)$).[30] Since the inequality impact from the mechanical effect functions similarly to the associated revenue effect, the new term is similar to the old, represented by the numerical constants in the numerator and denominator. The standard mechanical effect term is dampened or amplified by a multiplicative factor dependent on how inequality changes, $\bar{\kappa}(z)$, and whether or not this is welfare-enhancing, η.

As the individuals' work decision is unaffected by the mechanical effect, this term indicates the increased social willingness to change inequality levels absent any other changes. It has the same sign as the inequality externality η, as $\bar{\kappa}(z)$ is always positive for all inequality metrics with monotonically increasing weights – for the Gini, for example, $\bar{\kappa}(z) = F(z)$. Assuming a negative (positive) inequality externality, the full term unambiguously increases (decreases) the marginal rate in every tax bracket except at the very top and at the very bottom. The term exists whether or not the agent makes the socially optimal work decision.[31]

The externality thus introduces two new terms to the optimal tax formula. Both new terms always change after-tax income inequality in the direction of the externality. If we define progressivity as a lower after-tax Gini coefficient (Piketty and Saez, 2007), the resulting optimal tax rates with a negative (positive) inequality externality are unambiguously more progressive (regressive)

tax depending on the direction of the inequality externality. The optimal marginal tax rate at the top of a bounded distribution is the $\tau(z) = \frac{\eta\kappa(z)}{1+\eta\kappa(z)}$.

[30] $\alpha(z)$ contributes to both channels. See footnote 28.

[31] Unlike the behavioral effect, we note that this term can be "encapsulated" by a modified social welfare weight that only requires η as an empirical variable. In Equation 12, for example, such a modified weight would be $\bar{G}' = \bar{G} - \eta\bar{\kappa}(z)$. To do the same for the behavioral term, one would require $\alpha(z)$ and $\epsilon(z)$.

than the standard case. This shows an intuitive result; if inequality is considered a public bad, optimal income tax rates are more progressive than those previously found in the literature. If inequality is considered a public good, optimal income tax rates are more regressive than those previously found in the literature. In general, the new key parameters are the size of the inequality externality (represented by η) and the choice of the relevant inequality metric (represented by κ).

IV Optimal Income Taxation: Numerical Simulations

In this section we use numerical calculations to find optimal marginal tax rates in the presence of a post-tax income inequality externality. The main focus of the numerical simulations will be on how the inequality externality changes the results from the no-externality case.[32] We use the mechanism design solution from Section III.A throughout, which avoids the problem of an endogenous pre-tax income schedule.[33] We assume quasi-linear utility, a constant labor elasticity, and a linear homogenous inequality externality.[34]

Method In the traditional optimal tax literature, tax rates are largely determined by three factors; (i) the shape of the wage-earning ability distribution (Mankiw et al., 2009), (ii) the social welfare function, and (iii) labor or earnings elasticities..

The first factor is the shape of the wage-earning ability distribution $F(n)$. Our main specification uses empirical survey data for the 2018 U.S. wage distribution gathered from the Annual Social and Economic Supplement of the Current Population Survey.[35] The underlying density distribution $F(n)$ was extracted using a Kernel density estimation. Because survey data is incomplete towards the top, we also assume that the wage distribution approximates a Pareto distribution for wages above \$320/hour with a constant parameter estimated from the top. This is $\alpha(n) = 1.9$, very close to the $\alpha_{Top} = 2.0$ used in Saez (2001). Slightly more than 0.5% of all income-earners are affected. In addition to this empirical wage-earning ability distribution, we also present two standard theoretical distributions in Appendix E.I.

The second factor is the social welfare function. To span the range of non-increasing social welfare functions we use two extremes; (i) a fully Utilitarian SWF, and (ii) the Rawlsian minmax, which implies that the objective function of the government is to optimize the welfare of the worst-off member of society. In comparing to this most inequality-averse SWF we illustrate how the

[32] See the discussion in Saez (2001), among others, for a numerical exploration of the standard parameters.

[33] This does not significantly affect the results. Our no-externality results are nearly identical to those found in Saez (2001) and others.

[34] This implies a monotonic transformation of the following utility function:

$$U(x, h, \bar{\theta}) = log\left(x - \frac{h^{\left(1+\frac{1}{E_c}\right)}}{\left(1+\frac{1}{E_c}\right)} - \eta\bar{\theta}\right) \tag{13}$$

[35] Microdata were collected with IPUMS (Flood et al., 2018). Total wage income was divided by the average hours worked in a year to find the hourly wage distribution for individuals aged between 21 and 66 years. Individuals with no or negative wage income were excluded.

individual inequality concerns from an inequality externality are functionally distinct from the social inequality concerns from SWFs.

The third of these factors are the individuals' labor elasticities. We keep these homogenous for simplicity in our analysis, assuming that the elasticity of labor supply is constant at $E_L = 0.3$ for all income levels, a reasonable mid-range value from empirical estimates. This choice does not significantly affect the analysis.

These choices decide the shape of the no-externality optimal tax function. The externality necessitates two additional choices; the inequality metric and the magnitude (and direction) of the externality.

The two inequality metrics we show in the main text are the Gini, introduced in Equation 4, and a generalised Gini with weights of the following form,

$$\kappa_T(z) = (q+1)F(z)^q - 1, \tag{14}$$

which was designed to analytically approximate top income shares (which have a discrete jump and are thus analytically intractable). The Gini corresponds to $q = 1$ in this specification, while larger q approximates top income share inequality metrics. We use $q = 4$ in the main robustness test. The weights of the inequality metrics we use are plotted in Figure I. The figure shows the relative weight of the income of any agent when calculating the specified inequality metric. It also shows the weights used in the top 10% income share for comparison, which is discontinuous and thus not usable in an analytical setting. Other inequality metrics are examined in Appendix E.II.

Given the inequality metric we need to choose values for the inequality externality magnitude. The values of η depend on which inequality metric is chosen to be relevant for the externality, and we denote the values calculated for the Gini coefficient as η_G. As there are unavoidable empirical challenges in calibrating such a number,[36] we do not aim to strongly argue for any one parameter value. We instead use a range of realistic values to illustrate the potential tax policy consequences of various income inequality externalities.

To make a reasonable first-pass at an order of magnitude of η_G one could take the cross-country correlation between income inequality and externality dimensions – naïvely taking the correlation after controlling for observables as causal – and use willingness-to-pay estimates for each externality dimension to find the dimension's contribution to the total η. We do this for intentional homicides as an illustrative example. We use data from the World Bank for homicides, the World Inequality Database for the Gini coefficient, and Cohen et al. (2004) for the societal willingness to pay for fewer homicides.[37] The correlation between income inequality and intentional homicide is strongly positive, and through this very simple approach we find $\eta_{G,homicides} \approx 0.07$. As this only represents a single externality channel, the full η_G estimate would be found as $\eta_G = \sum_i \eta_{G,i}$. It seems

[36] Beyond specific empirical challenges relating to the existence and quality of the available data, it is very challenging – perhaps impossible – to find true exogenous variation in inequality. Further, revealed preferences is challenging as individuals' labor decisions have a negligible effect on societal inequality.

[37] Cohen et al. (2004) estimates the total social cost of a homicide as \$9.7 million, or \$12.8 million corrected for inflation to 2018.

Figure I: Weights of Inequality Metrics

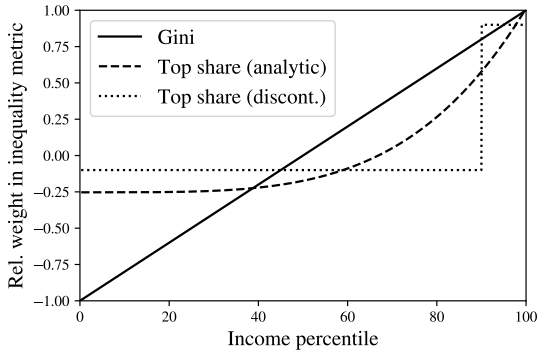
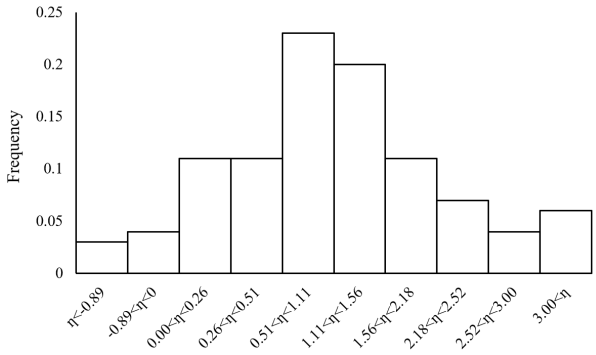


Figure II: Estimated η_G (Carlsson et al., 2005)



Notes: Figure I shows the relative weights of individuals' income in the inequality metrics we primarily use (the Gini and the analytic top share metric are used in Figures III and IV, respectively). This corresponds to $\kappa(z)$ in the general expression $\bar{\theta} = \int_{\underline{z}}^{\bar{z}} \kappa(z)x(z)dF(z)$. More inequality metrics are explored in Appendix E.II. Figure II shows the estimated magnitudes of the inequality externality magnitude η_G from the survey experiment in Carlsson et al. (2005). In the following numerical simulations we restrict η_G between -0.5 and 2.0 (and equivalent values for other inequality metrics).

reasonable, then, to believe that the full externality could have an order of magnitude roughly one order above this estimate.

To find a range of η_G that takes into account *all* externality dimensions we present estimates based on data from Carlsson et al. (2005). The work uses a survey design to estimate macroeconomic inequality aversion in Swedish university students.[38] The survey, which asks respondents to decide what income-inequality trade-off their hypothetical grandchildren would prefer, allows us to find individual preferences for η determined to an interval.[39]

The distribution is presented in Figure II. The median respondent in the survey has approximately $\eta_G = 1.00$. A majority of respondents have $0.26 < \eta_G < 2.18$.[40] A negative η_G – indicating a preference for inequality, or that inequality is a positive externality – is only observed in 7% of respondents. The equivalent externality magnitude values for top income shares, η_T, are calculated from the same experiment. As a general rule of thumb, $\eta_G \approx 2\eta_T$ when externality magnitudes are equal.

As these numbers are rather abstract, we present an alternative way of understanding the magnitudes through equivalent incomes. Answering the following question pins down either η: *What multiple of their current income should an average agent require to move from Denmark-like to United States-like inequality?*[41]

[38] Bergolo et al. (2021) finds comparable numbers for Uruguayan university students

[39] Using a survey experiment instead of a direct externality estimate means that we are relying on potentially biased beliefs to proxy for inequality's externality effects. There is also selection bias in the survey respondents and, because the only degree of freedom is being used to estimate the extent of inequality aversion, it is not possible to know how well our homogeneity assumption matches the respondents' perceived utility functions. All these reasons contribute to why we are using a *range* of η.

[40] Due to the design of the experiment, any one individual's inequality aversion is only pinned down to a range.

[41] Assuming the same leisure, that the mean income difference between the two countries is negligible, and that

Answering the question creates equivalent incomes for differing inequality levels. These equivalent incomes for Denmark and the United States, and their corresponding η when using the Gini, are shown in Table III. As an example, if we have an inequality externality of $\eta_G = 1.0$, the average individual in a society with Denmark's inequality level would require 13% more income to be indifferent if inequality increased to the U.S. level. If $\eta_G = 0$, the agent is indifferent without any change to their income.

Table III
The Magnitude of Inequality Externalities η_G

	$\eta = -0.5$	$\eta = 0.0$	$\eta = 0.5$	$\eta = 1.0$	$\eta = 2.0$	$\eta = 3.0$
U.S. Income Multiplier	0.94	1.00	1.06	1.13	1.25	1.38

Note: Which multiple of their current income would an average-income agent need to move from Denmark-like to U.S.-like inequality? Above are these equivalent incomes for various levels of the inequality externality η_G from the utility function in Equation 13.

Based on these two techniques we use the range $-0.5 \leq \eta_G \leq 2.0$ for the Gini-based externality and $-0.15 \leq \eta_G \leq 1.0$ for the top share-based externality in the main numerical simulations.

The numerical simulations were performed in Python through an iterative process.[42] For every result we check that the individual's second-order conditions hold using two different methods; first we ensure that earnings increases over ability (Lollivier and Rochet, 1983), and second we numerically ensure that the incentive compatibility constraint is satisfied for every agent.

Main Results: The Gini Externalities Our main specifications, using the Gini as the post-tax income inequality metric, are presented in Figure III. The introduction of even a small post-tax income inequality externality substantially changes the optimal tax structure. The effect is larger towards the top of the income distribution. Note that at the very top, the Utilitarian and Rawlsian results converge, [43] as in the classical literature – thus, all the negative externality Utilitarian SWFs we simulate have higher top rates than the no-externality Rawlsian case. This illustrates that a Rawlsian SWF, in itself, does not imply a maximum dislike of inequality.

The very top marginal tax rate increases from 68% to 90% when assuming a moderately large negative inequality externality, $\eta_G = 2.0$. For $\eta_G = 1.00$, the value closest to the empirical externality estimate taken from Carlsson et al. (2005), the optimal top marginal tax rate is 85%. With a small positive inequality externality ($\eta_G = -0.5$), the optimal top marginal tax rate is only 40%.

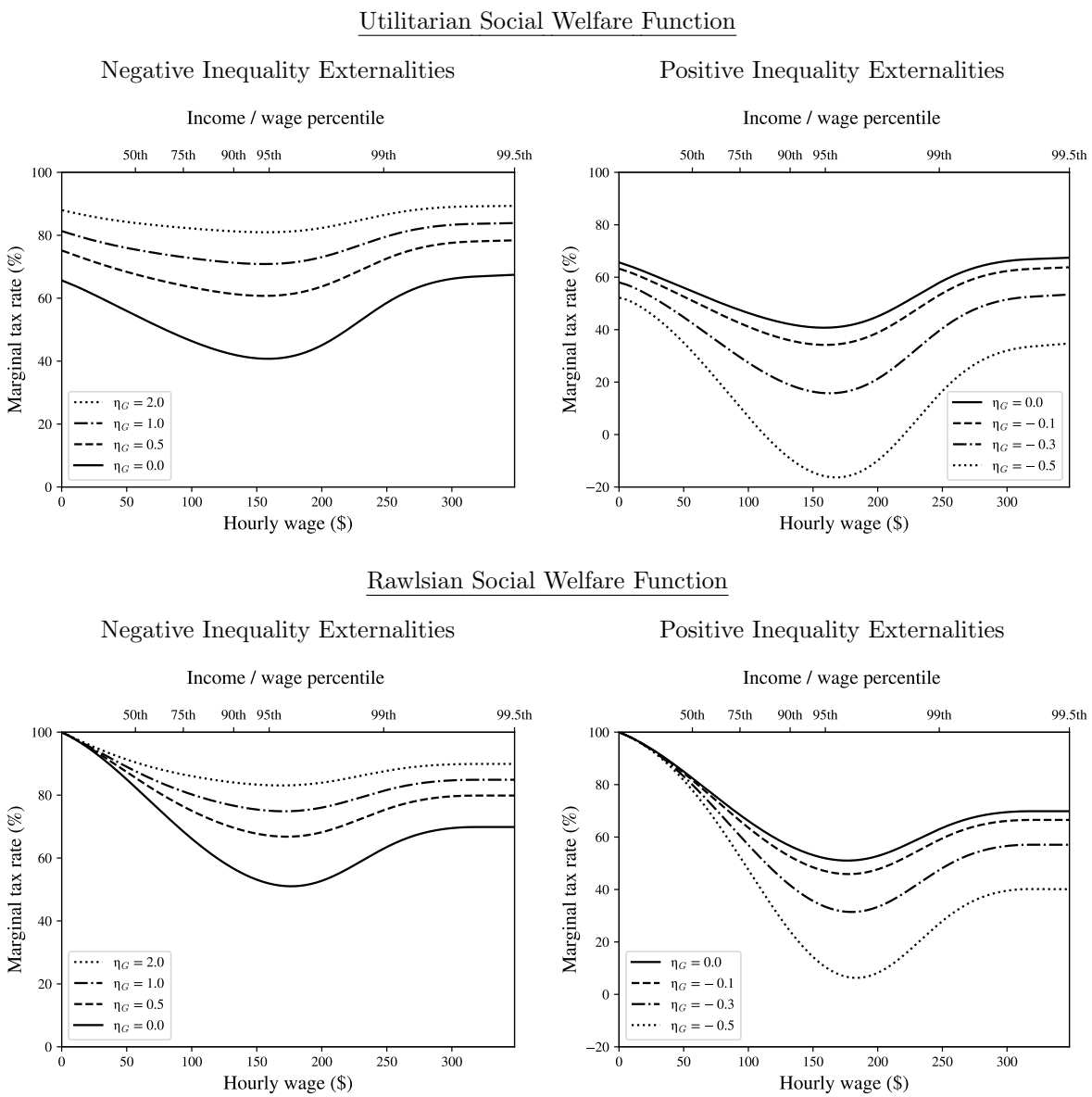
In the Utilitarian case, the marginal tax rates are shifted up or down from the standard case

relative position is irrelevant. According to the 2017 World Economic Outlook database GDP per capita is \$61,803 in Denmark, and \$59,707 in the United States. Calculations are based on Gini coefficients of 0.410 for the United States and 0.285 for Denmark.

[42] We assume an initial tax schedule, set agents' labor supply based on this tax schedule, and then calculate the resulting optimal tax rate. We iterate on this process until an optimum is found. This is further discussed in the Appendix of Mankiw et al. (2009).

[43] This is due to the assumptions of separability and a homogeneous inequality externality.

Figure III: Optimal Marginal Income Tax Schedules with Gini Inequality Externalities



Notes: Optimal marginal tax rates for various Gini-based inequality externalities with magnitudes η_G, where inequality is either a negative externality (left) or a positive externality (right). The social planner is Utilitarian (above) and Rawlsian (below). The two cases converge when moving towards the top. Empirical estimates indicate $\eta_G = 1.0$. The solid line, $\eta = 0$, is the standard no-externality case. Further explanation of η is in Table III. Note the different scales of the vertical axes between the negative and positive externalities.

over the entire distribution. This is due to the empirical strength of the mechanical effect (which increases/decreases optimal rates across the entire distribution for a negative/positive externality), which dominates that of the behavioral responses (which increases or decreases optimal rates differentially at the top and bottom) under our parameter choices.[44] The effects are larger near the top, which is particularly noticeable around the 95$^{\text{th}}$ percentile, where the optimal marginal tax rate is shifted from 42% in the no-externality case up to 81% under a negative externality (when $\eta_G = 2.0$) and down to -16% with a positive externality (when $\eta_G = -0.5$). The larger effects near the top of the distribution is due to the equality effects of the mechanical and behavioral channels working in the same direction in this region (as discussed in Section III.B). We observe negative optimal marginal tax rates for income earners between the 84$^{\text{th}}$ and 98$^{\text{th}}$ percentiles when $\eta_G = -0.5$. These negative optimal top rates come from the social planner's incentive to increase income inequality when inequality is a positive externality, even if this comes at a significant revenue cost – to the extent that a tax subsidy at the top can be optimal. We also note that all simulations have lower optimal tax rates around the 90$^{\text{th}}$–95$^{\text{th}}$ percentiles due to the well-known decrease of the local Pareto parameter around these values, which leads to the classical U-shape found in the literature (Diamond, 1998). We return to this shortly.

The Rawlsian externalities we introduce have only small impacts near the bottom of the distribution, where marginal tax rates are very high in the no-externality case. This is driven by a very high mechanical revenue benefit of taxation near the bottom (which is also found in the classical literature).[45] The effects of the inequality externality are mostly located above the 90$^{\text{th}}$ percentile for both negative and positive externalities. Under a positive externality, top marginal tax rates approach zero around the 97$^{\text{th}}$ percentile.

The extent of the classical U-shape varies across simulations. It is most striking in the positive externality and no-externality simulations, and is difficult to notice in the negative externality simulations. As the U-shape has been widely discussed as having potential implications for practical tax design it is relevant to ask why this occurs. The U-shape emerges from the empirically estimated wage-earning (or income) distribution, as the local Pareto parameter α is high around these wage (or income) percentiles. In short this implies a relative over-density of individuals *in* these tax brackets compared to those *above* these tax bracket, which in turn implies that the relative strength

[44]This result is not universal, and the effect of the externality at the bottom is usually smaller than in this case due to the counteracting behavioral response. Indeed, the Utilitarian case with no income effects has among the least top-heavy distributional optimal policy effects of any of our simulations. It is notable that the effects are largest at the top even in this case. Using certain skill distributions, such as the full Pareto distribution in Appendix E.I, a negative externality *decreases* optimal marginal tax rates at the bottom. We also find this result with any pre-tax income inequality externality (see Section IV.B).

[45]The high optimal rates at the bottom of all the Rawlsian simulations are due to the large positive mechanical revenue effects of increasing bottom marginal tax rates. When one only cares about the very bottom agent, as in the Rawlsian case, redistributing away from any other agent is a net positive absent changed labor choices. Since we do not consider income effects, these labor choices do not occur for anyone above the tax bracket in question. The mechanical revenue effect is thus very large at the bottom and leads to very high marginal tax rates in this region. The introduced equality effects are not large enough to change this substantially. In contrast, the Utilitarian simulations take into account the income losses from agents above the tax bracket, which discounts the mechanical benefits of tax increases near the bottom. Very high bottom marginal tax rates are thus less appealing, and the effects of the inequality externality are more visible.

of the behavioral channel is high in this bracket (as compared to the relatively low strength of the mechanical effect). In other words, optimal tax policy in these brackets is increasingly set by the welfare consequences of agents' behavioral responses. This decreases the no-externality optimal tax rates in the region. How does this change when one introduces an inequality externality? In the negative externality case, there is a welfare-positive dimension to the behavioral responses (namely decreased inequality). It follows that an increased importance of the behavioral responses does not necessarily imply a U-shape and lower optimal tax rates – as we can see in the simulations.[46] In the positive externality case, meanwhile, the shift towards a concern for behavioral responses is still highly relevant, as the behavioral responses remain entirely welfare-negative (through decreased revenue and decreased inequality). To summarize, the classical U-shape from the optimal taxation literature may depend on the absence of a negative income inequality externality.

The exact optimal tax structure depends heavily on the model specification, so the numerical simulations should be interpreted with caution.

Robustness: Top Income Share Externalities The choice of the inequality metric naturally influences our results. And while the Gini coefficient is analytically appealing, it is often considered to over-weight middle-income inequalities. To address this concern we present a robustness check of our main findings in Figure IV by using the top income share metric shown in Figure I as the relevant inequality measurement. This inequality metric is defined as $q = 4$ in the general top income share metric family $\kappa(z) = (q+1)F(n)^q - 1,\ q \in \mathbb{N}$.

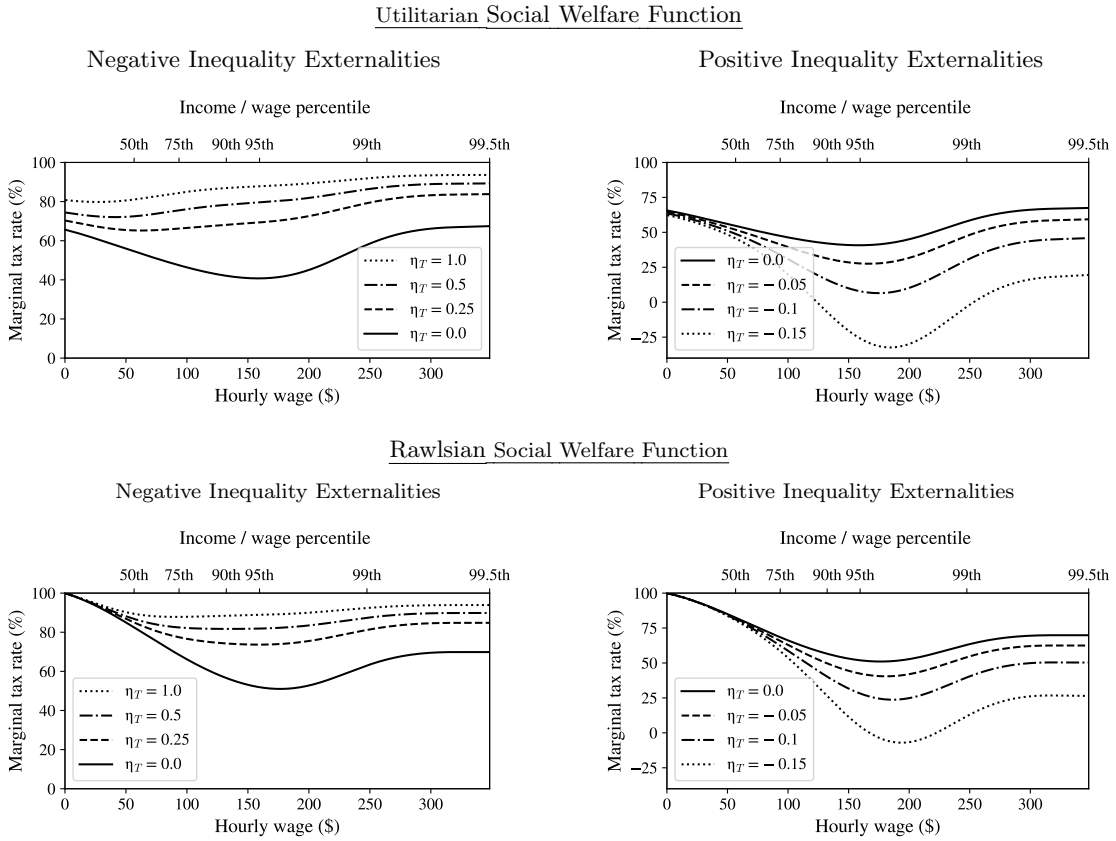
The externality effects are larger at the top and smaller at the bottom when using the top income share metric. With either a Utilitarian or Rawlsian SWF, the optimal top marginal income tax rate goes from 68% in the no-externality case to 90% when $\eta_T = 0.5$ (comparable to $\eta_G = 1.0$ in Figure III, the value closest to the empirical externality estimate taken from Carlsson et al. (2005)). For the largest negative externality, $\eta_T = 1.00$, the optimal top marginal tax rate is 94%. For the largest positive externality, $\eta_T = -0.15$, the optimal top marginal tax rate is only 26%.

In the Utilitarian case, the effects near the bottom are now relatively small. The negative externalities increase optimal marginal tax rates by around fifteen percentage points at most near the bottom, whereas the positive externalities have hardly any impact in the region. Around the top, the effects are now larger; the optimal marginal tax rates near the 97^{th} percentile change from 42% in the no-externality case up to 89% under a negative externality ($\eta_T = 1.00$) and down to -32% under a positive externality ($\eta_T = -0.15$). Negative optimal marginal rates are observed between the 87^{th} and 99^{th} percentiles when $\eta_T = -0.15$.

Similarly, the top Rawlsian tax rates can now be negative close to the top. If $\eta_T = -0.15$, optimal marginal tax rates begin at near a hundred percent and go below zero between the 96^{th} and the 99^{th} percentiles. Near the bottom, Rawlsian marginal tax rates remain similar to the Gini case.

[46] Optimal marginal tax rates can even increase in the region under different specifications. In Section F.I this occurs under a negative pre-tax income inequality externality.

Figure IV: Optimal Marginal Income Tax Schedules with Top Share Inequality Externalities



Notes: Optimal marginal tax rates for various top share-based inequality externalities with magnitudes η_T where inequality is either a negative externality (left) or a positive externality (right). The social planner is Utilitarian (above) and Rawlsian (below). The two cases converge when moving towards the top. Empirical estimates indicate $\eta_T = 0.5$. The solid line, $\eta = 0$, is the standard no-externality case. Note the different scales of the vertical axes between the negative and positive externalities.

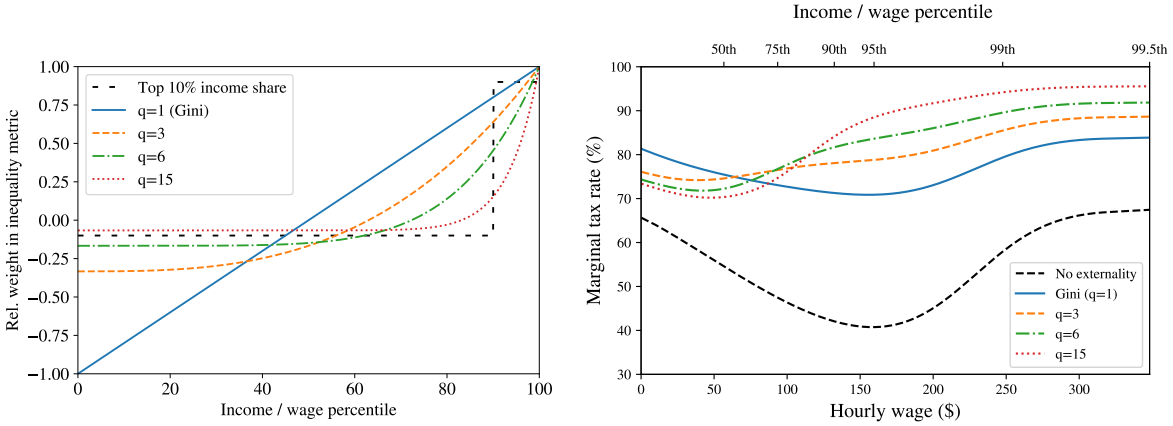
That the effects of the externality are increasingly concentrated towards the top of the distribution with increasing q, i.e. when we move away from the Gini towards a top income share, is a general result. We show this more clearly in Figure V. There we plot more inequality metrics from the top income share family alongside their resulting optimal tax rates. The externality is kept constant at the upper bound of the median inequality aversion range from Carlsson et al. (2005). Figure IV used $q = 4$; here we show the effect of moving from $q = 1$ (the Gini) to values increasingly focused on top incomes (up to $q = 15$). The no-externality case is shown as a reference in dotted black. The ability distribution is the empirical wage distribution, and the SWF is Utilitarian.

The optimal tax changes are larger near the top with increasing q, which should not be surprising given the increasing weight of top incomes in the inequality metric. It is also noticeable, however, that the effects near the bottom are reduced. This is not as obvious, as lower inequality metric weights near the bottom have opposite optimal tax effects through the behavioral channel (through which lower κ_{bottom} leads to higher τ) and the mechanical effect (through which lower κ_{bottom} generally leads to lower τ through a higher $\bar{\kappa}_{bottom}$).[47] In the numerical simulations, the mechanical

[47] In the case of the behavioral channel, the bottom-earner imposes less of an externality and the negative Pigouvian

Figure V

Varying the Inequality Metric with a Fixed Externality Magnitude



Note: Left: The income weights over the distribution of various inequality metrics in the family where $\kappa(z) = (q+1)F(n)^q - 1$, $q \in \mathbb{N}$. The top 10% income share is also plotted. Larger q indicates that top incomes are increasingly weighted. Right: Optimal marginal tax rates for these inequality metrics, keeping the magnitude of the inequality externality constant for all q at the upper bound of the median value from the empirical inequality aversion estimates in Carlsson et al. (2005). The social planner is Utilitarian. The productivity distribution is the empirical wage distribution. The black dotted line is the standard case of no inequality externality. The elasticity of labor E_L is 0.3.

effect is more powerful, indicating that the average marginal externality above is more impactful than marginal externality of the tax bracket itself.

Overall, using top income shares further concentrates the effect of the externality towards the top of the tax schedule. With other inequality metrics, such as those in the S-Gini family, results are overall similar. This is further discussed in Appendix E.II. In sum, the Gini is a conservative choice which dampens effects at the top in return for larger changes across the rest of the distribution. We will now discuss implications for top tax rates specifically.

IV.A Equality concerns: Top tax rates

Equality concerns – the consequence of the inequality externality – come in addition to the revenue concerns usually discussed in the OIT literature. Their policy importance differs according to income bracket. In particular, as we have discussed in the preceding sections, equality concerns have a large effect on the optimal top tax rate.

Revenue considerations, which in this context implies the direct individual effects from the redistribution of income, have few distributional biases. In a Rawlsian set-up, for instance, one tax dollar raised remains one tax dollar raised, regardless of which tax-payer pays it (if not taken from the very bottom). In other social welfare functions the welfare benefit from additional tax revenue is usually relatively stable in the top half of the distribution. Equality concerns are naturally different: *where* the income is taken from is of key importance. And, as we have seen, the tax policy effects

term is thus smaller. In the case of the mechanical effect, redistributing from everyone above is less impactful for inequality-reduction if everyone in the lower half is weighted relatively equally.

of these equality concerns generally increase as one approaches the top of the distribution.

It follows that some of the variation in international tax brackets, particularly at the top, could be due to policy setters' differing considerations of the inequality externality. Two Rawlsian governments might agree on the elasticity of earnings and revenue-maximizing tax rates and still strongly disagree on optimal top tax rates – *if* they disagree on how inequality changes society. In keeping with the logic of the inequality externality, this can be true even in the absence of jealousy and envy. Our numerical simulations strengthen this point.

Below we discuss specific findings related to these large impacts on optimal top income tax rates. First we show two practical implications of our model, justifying observed policy arguments that cannot be rationally explained under standard revenue considerations. Second we discuss the existence of optimal rates higher than the revenue-maximizing Laffer rate.

1 Large variation in top rates: A maximum income, or the Rawlsian Conservative? OIT models are generally considered more accurate towards the top of the distribution. Top marginal income tax rates often converge to around $60 - 70\%$, even in the Rawlsian case. Although these numbers depend heavily on parameter specifications, heterodox assumptions are required for optimal rates below 50% or above 80%.[48]

As we have shown in the preceding sections, varying the value of the inequality-sensitivity parameter η has a large effect on the top optimal income tax rates. This variation is large even when compared to the variation induced by changing standard parameter values. We examine this in Tables IV and V. These tables show how the optimal top tax rate varies with (1) combinations of Gini income inequality externality magnitudes η_G and the inverse local Pareto parameter $1/\alpha$ (Table IV) or (2) combinations of Gini income inequality externality magnitudes η_G and labor elasticity values E_L (Table V). The inequality externality induces changes that are generally larger than the effects from changing $1/\alpha$ or E_L. By changing η within reasonable bounds, the same Rawlsian social planner can find optimal top tax rates from close to zero to over 90%. Under stronger positive externalities the same social planner can even find negative optimal top rates. In other words, a wide range of top tax rates can be optimal depending on the magnitude of the inequality externality.

We use two real-world examples to illustrate the power of such a finding.

First, the idea of extremely high top tax rates (a "maximum income"). If one believes in a large negative inequality externality, here represented by $\eta = 3.0$, the negative effect of top income earners on the rest of society is sufficient to argue for top tax rates above 90%. These are similar to tax rates from the post-war period in the United Kingdom, Germany, and the United States. The disincentive for high earners at this stage begins to approach a maximum income.

Second, the idea of a Rawlsian government with low tax rates on the highest income-earners. If one believes in even a small positive inequality externality, here represented by $\eta = -0.5$, marginal rates at the top quickly fall below 50% and begin approaching zero. We call this the Rawlsian

[48] Piketty et al. (2014) finds revenue-maximizing rates varying from 57% to 83% with differing elasticity compositions, for instance.

Table IV

Optimal Top Tax Rates, Inequality Externalities and Distribution Parameters

		Inverse top Pareto parameter $1/\alpha$											
		0.25	0.27	0.29	0.31	0.33	0.36	0.40	0.44	0.50	0.57	0.67	0.80
	-0.50	4	7	11	14	18	22	27	32	37	42	49	55
	-0.25	36	38	40	43	45	48	51	54	58	62	66	70
	0.00	**52**	**54**	**55**	**57**	**59**	**61**	**63**	**66**	**68**	**71**	**74**	**78**
	0.25	62	63	64	66	67	69	71	73	75	77	79	82
Sensitivity	0.50	68	69	70	71	73	74	76	77	79	81	83	85
to	0.75	73	73	74	76	77	78	79	80	82	84	85	87
inequality	1.00	76	77	78	79	80	81	82	83	84	86	87	89
η	1.25	79	79	80	81	82	83	84	85	86	87	89	90
	1.50	81	81	82	83	84	84	85	86	87	88	90	91
	1.75	83	83	84	84	85	86	87	88	89	90	91	92
	2.00	84	85	85	86	86	87	88	89	89	90	91	93
	2.25	85	86	86	87	87	88	89	89	90	91	92	93
	2.50	86	87	87	88	88	89	90	90	91	92	93	94
	2.75	87	88	88	89	89	90	90	91	92	92	93	94
	3.00	88	88	89	89	90	90	91	91	92	93	94	94

Note: Top marginal tax rates from Equation 12 with varying values of an inequality externality and the inverse local Pareto parameter $1/\alpha$ at the top. The social planner is Rawlsian. The elasticity of labor E_L is 0.3. The inverse local Pareto parameter $1/\alpha$ is approximately 0.5 at the top in empirical data (and in the remainder of the paper). The standard no-externality case is in bold.

Table V

Optimal Top Tax Rates, Inequality Externalities and Labor Elasticities

		Elasticity of labor E_L									
		1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10
	-0.50	0	3	6	10	14	20	27	37	50	69
	-0.25	33	35	37	40	43	47	52	58	67	79
	0.00	**50**	**51**	**53**	**55**	**57**	**60**	**64**	**68**	**75**	**85**
	0.25	60	61	62	64	66	68	71	75	80	88
Sensitivity	0.50	67	68	69	70	71	73	76	79	83	90
to	0.75	71	72	73	74	76	77	79	82	86	91
inequality	1.00	75	76	76	77	79	80	82	84	88	92
η	1.25	78	78	79	80	81	82	84	86	89	93
	1.50	80	81	81	82	83	84	85	87	90	94
	1.75	82	82	83	84	84	85	87	89	91	94
	2.00	83	84	84	85	86	87	88	89	92	95
	2.25	85	85	86	86	87	88	89	90	92	95
	2.50	86	86	87	87	88	89	90	91	93	96
	2.75	87	87	87	88	89	89	90	92	93	96
	3.00	88	88	88	89	89	90	91	92	94	96

Note: Top marginal tax rates from Equation 12 with varying values of an inequality externality and elasticity of labor E_L. The social planner is Rawlsian. The inverse local Pareto parameter $1/\alpha$ is 0.5 in these calculations. The elasticity of labor E_L is 0.3 in the remainder of the paper. The standard no-externality case is in bold.

conservative; the argument that a low top tax rate will lead to the highest possible utility for the worst-off agent.

Both of these intuitive arguments have been proposed in political discourse. In standard OIT literature, however, they are unfounded. One strength of our model is that such arguments can be logically substantiated, and disagreements can be traced back to the variable η. Individual opinions on η could be related to (or even determinants of) political leanings and policy preferences.

2 The Laffer Curve The central idea of the Laffer curve is simple and true; above a certain tax threshold revenue drops with increased taxation. However, the Laffer curve is often also described as an upper bound on sensible taxation. Laffer (2004) describes this as the "prohibitive range" of taxation, and Manning (2015) argues that "one would not want a rate higher than the Laffer rate".

In the presence of an inequality externality the above statements could be either misleading or false. The externality negligibly changes agent behavior when there is a large number of agents, so the revenue-maximizing rate does not change. However, the welfare-maximizing rate can change, and is in fact often above the Laffer rate given the public benefit of distributional changes.

As an example, consider a society with ten agents, one vastly more wealthy than the other nine. Given the desirability of equality, the welfare-maximizing top marginal rate can be higher than the revenue-maximizing rate, which is zero at the top according to standard results. The Rawlsian numerical simulations in Section IV provide numerical examples.

The optimal income tax rate can be higher than the revenue-maximizing rate both at the top (given a negative externality), and at the bottom (given a positive externality). Specifically, the optimal marginal income tax rate is higher than the revenue-maximizing marginal income tax rate if, using the framework in Equation 12,[49]

$$\eta\alpha(z)\epsilon(z)\kappa(z) + \eta\bar{\kappa}(z) > \bar{G}(z),$$

that is, if the equality effects of taxation are larger than the welfare effects.[50] If the inequality externality does not exist, so that $\eta = 0$, the statement never holds unless social weights are negative – this is the standard result. In the case with an inequality externality, $\kappa(n)$ goes from negative to positive with higher incomes and η changes sign depending on the direction of the externality. Thus it can hold either at the bottom (with a positive externality, $\eta < 0$) or at the top (with a negative externality, $\eta > 0$).

In the Rawlsian case, the right-hand side of Equation 15 is zero above the very bottom earner. Thus, using the Gini values, the inequality simplifies to

[49] In the most general framework, see Appendix C, this is equal to,

$$\gamma\left[\kappa(n) + \frac{\zeta u_{x(n)}}{f(n)n}\int_n^\infty \left[\frac{\kappa(p)}{u_{x(p)}}\right] f(p)dp\right] > \frac{\zeta u_{x(n)}}{f(n)n}\int_n^\infty \left[W'(U(p))\right] f(p)dp, \tag{15}$$

which represents the same intuition; the equality effects of taxation must be larger than the welfare effects.

[50] This follows from comparing Equation 12 to the revenue-maximizing tax rate, which is the same equation when $\bar{G}(z) = 0$ and $\eta = 0$.

$$\frac{F(z)}{\alpha(z)\epsilon(z)} > 1 - 2F(z), \tag{16}$$

which is independent of η and holds for any income above the median.[51]

The Mirrlees literature occasionally uses the revenue-maximizing rate as a necessary upper bound for sensible tax rates. For example, Piketty et al. (2014) states that they "focused on the revenue-maximizing top tax rate, which provides an upper bound on top tax rates". This position would need to be modified in a model with societal effects of inequality.

IV.B Other types of inequality externalities

The preceding sections have discussed a *post-tax income* inequality externality. While such an externality could be reasonable for several reasons – some of which we outline in Section V.A – there is no *a priori* reason to exclude the possibility of other inequality externalities. Here we consider how the theoretical intuition changes with different types of inequality externalities in the optimal non-linear income taxation problem.

Pre-tax income inequality externality A pre-tax income inequality externality implies different equality impacts of the behavioral and mechanical effects. To start with the behavioral responses, note that any behavioral shift that follows from a tax increase would lead to a larger pre-tax income reduction than post-tax income reduction; pre-tax income being reduced by one unit reduces post-tax income by only $1 - \tau(z)$ units, which is generally between zero and one (excluding the extreme case of negative marginal rates). As such the effect of any behavioral response on pre-tax income inequality is generally larger than that on post-tax income inequality. Subsequently the pre-tax externality is more heavily affected by this channel than we saw in the post-tax case.

The mechanical effect, meanwhile, no longer has any impact on the externality. This follows from pre-tax income inequality being unchanged by the mechanical (post-tax) redistribution of income from those above the perturbation.

The optimal income tax rates in this case are

$$\tau(z) = \frac{1 + \eta_{pre} \cdot \kappa(z)\alpha(z)\epsilon(z) - \bar{G}(z)}{1 + \alpha(z)\epsilon(z) - \bar{G}(z)},$$

where η_{pre} is the pre-tax income inequality externality magnitude.[52] The full derivation is in Appendix F.I.

[51] This is intuitive; the Rawlsian rate is the revenue-maximizing rate, and the incentive for equality increases tax rates at least above the median agent.

[52] There is a subtle point to be made here about the magnitude of η_{pre}. Generally, pre-tax income inequality is higher than post-tax income inequality, which could influence this shadow price of inequality under different specifications (see Equation 9). The simple derivation we present here assumes a constant marginal rate of substitution between income and pre-tax income inequality, and as such only the marginal effect of the tax increase matters.

This result implies that a pre-tax income inequality externality could lead to a progressive modification of the standard Mirrlees tax rates (where we mean progressive in the traditional sense; marginal tax rates which increase with income). We see this in Figure VI, which shows negative pre-tax inequality externalities in the Utilitarian framework with the same specifications as in our main specification. Bottom tax rates are lower and top tax rates are higher than in the no-externality case, which is a general finding under separability. The marginal tax rates increase from 57% at the bottom to 81% at the top when $\eta_G = 2.0$.

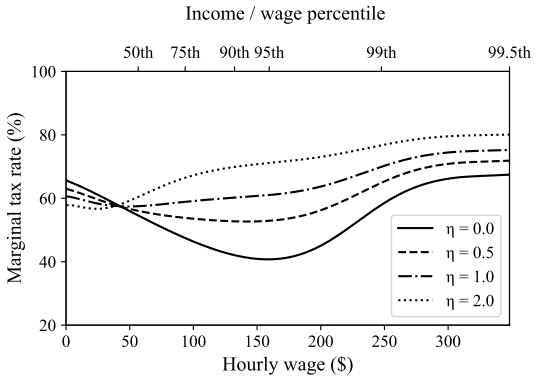


Figure VI: Optimal income tax rates with a pre-tax income inequality externality. The social planner is Utilitarian, and the remaining specification is identical to Figure III.

Interestingly, the inequality externality removes the well-known U-shape of optimal marginal tax rates from the classical literature in favor of a tax schedule where the marginal tax rate generally increases in income. Compared to the classical literature (or the case of a post-tax income inequality externality), this new optimal marginal income tax schedule is closer to that observed in most developed countries. One might wonder whether governments have, to some extent, considered pre-tax inequality as an ill in itself when designing tax schedules – if so, this could explain some of the differences between the numerical simulations from optimal tax theory and real-world tax schedules.

Utility inequality externality When considering a utility inequality externality, the behavioral channel no longer has an inequality impact. This follows from a miniscule tax perturbation from the optimum only leading to second-order utility effects. Thus, utility inequality would stay approximately the same through the behavioral responses. The mechanical effect would function similarly as in the post-tax income inequality case, as increasing the marginal tax rate reduces utility inequality by lowering the utility of those above the tax bracket.[53]

The optimal income tax rates in such a case are

$$\tau(z) = \frac{1 + \eta_U \cdot \bar{\kappa}(z) - \bar{G}(z)}{1 + \alpha(z)\epsilon(z) + \eta_U \cdot \bar{\kappa}(z) - \bar{G}(z)},$$

where η_U is the utility inequality externality magnitude. The full derivation is in Appendix F.II. Assuming that negative weights are acceptable, using the modified social welfare weights $\bar{G}'(z) = \bar{G}(z) - \eta_U \cdot \bar{\kappa}(z)$ allows this result to be simplified to the standard Mirrlees case without the need for empirical variables in the modified income-based welfare weights.[54] Further, this result can be approximated in the mechanism design case through utility-based social welfare weights, unlike both the pre-tax and post-tax externality results.

[53] This is more complicated outside the simple quasi-linear case, see Appendix F.II.

[54] To the extent that η_U is not an empirical variable, of course. A similar modification can be made to the income-based welfare weights in the post-tax income inequality case. However, there $\bar{G}''(z) = \eta\alpha(z)\epsilon(z)\kappa(z) + \eta\bar{\kappa}(z) - \bar{G}(z)$, indicating that the modified welfare weights are dependent on $\alpha(z)$ and $\epsilon(z)$.

Simply put, a utility inequality externality brings the problem closer to the standard no-externality case. Specifically, the utility problem can often be approximated by changing the inequality aversion of the SWF in the traditional Atkinson (1970) sense.[55] This is because the net effect of the utility inequality externality is to change the social benefit of each individuals' utility, which can be achieved through simply changing the standard social welfare weights.

There is a notable complication to this problem, namely that utility has to be carefully defined. Standard inequality metrics, such as those discussed in the post-tax income case, would not remain the same through monotonic transformations of utility. This complicates the problem both philosophically and analytically. The natural simplification we have used above is a quasi-linear utility function, in which case income changes have a one-to-one relationship with utility changes.

Table VI summarizes these results.

Table VI

Optimal Income Taxation Effects of Various Inequality Externalities

	Mechanical effect	Behavioral responses	Optimal tax rates $\tau(z)$
Post-tax income inequality	✓	✓	$\frac{1+\eta\alpha(z)\epsilon(z)\kappa(z)+\eta\bar{\kappa}(z)-\bar{G}(z)}{1+\eta\alpha(z)\epsilon(z)\kappa(z)+\eta\bar{\kappa}(z)+\alpha(z)\epsilon(z)-\bar{G}(z)}$
Pre-tax income inequality	-	✓ (stronger)	$\frac{1+\eta_{pre}\kappa(z)\alpha(z)\epsilon(z)-\bar{G}(z)}{1+\alpha(z)\epsilon(z)-\bar{G}(z)}$
Utility inequality	✓	-	$\frac{1+\eta_U\cdot\bar{\kappa}(z)-\bar{G}(z)}{1+\alpha(z)\epsilon(z)+\eta_U\cdot\bar{\kappa}(z)-\bar{G}(z)}$

Note: The table describes how each type of inequality externality functions in the optimal income taxation framework.

V Further Theoretical Discussion

We now turn to the more general implications of an inequality externality. The reframing of inequality as an externality leads to several intriguing realizations:

- Equality itself becomes policy-relevant and has an associated shadow price.[56] The trade-off between income maximization at the bottom and the preferred inequality level becomes relevant.
- A Rawlsian min-max is not the most inequality-averse modeling exercise. Similarly, a Utilitarian SWF is not the least inequality-averse modeling exercise if one restricts oneself to non-increasing social welfare weights.
- A change in marginal tax rates can lead to a "double dividend" of both more revenue *and* an inequality level closer to what is considered optimal, both of which are welfare-relevant.

[55] The exception is when separability does not hold such that individuals' behavior is directly affected by the externality.

[56] This shadow price corresponds to γ in Equation 7 and η in Equation 12. In general the shadow price is endogenous to the system solution; in the simplified small-perturbation solution it is constant.

- The marginal social welfare of income at the top can be negative (Carlsson et al., 2005). In a utilitarian framework with homogeneous agents and a negative inequality externality, the total welfare effect of additional income at the top is:

$$\frac{d\sum_j g_j U(x_j, \bar{\theta})}{dx_i} = g_i \frac{\partial U(x_i, \bar{\theta})}{\partial x_i} + \sum_j g_j \frac{\partial U(x_j, \bar{\theta})}{\partial \bar{\theta}} \frac{\partial \bar{\theta}}{\partial x_i}$$

The second term on the right-hand side comes from the inequality externality and can have significant magnitudes, as the results in Section III showed. It is negative if inequality increases ($\frac{\partial \theta}{\partial x_i} > 0$)[57] in a society with a negative inequality externality ($\frac{\partial U(x_j,\theta)}{\partial \theta} < 0$). It can be larger than the first term (the individual benefit from the consumption increase), indicating that additional income at the top can be detrimental if inequality is sufficiently socially disruptive.[58] The total effect depends on the relative importance of equality and consumption, a version of the familiar equity-efficiency trade-off.

This point may seem controversial. In the context of jealousy effects (ORP), Piketty and Saez (2013) argue that "hurting somebody with higher taxes for the sole satisfaction of envy seems morally wrong". In the context of inequality externality effects, however, the interpretation is perhaps more intuitive. Imagine, for instance, an extremely high-income agent who has a resource-determined control over the political process. If this political control hurts lower-income agents, taxation of the high-income agent designed to offset the political effects is intuitive and can be optimal in our framework. The same argument holds for other inequality externality effects.

This result is particularly important in the context of concentrated income gains. Extremely concentrated income gains – which are potentially becoming more prevalent with globalization and technical progress – are unambiguously good in standard models. The few agents receiving the additional income increase their utility, while every other agent's utility remains the same. If increased income inequality changes society, however, the other agents may be affected, positively or negatively, despite constant income levels. This is captured by an inequality externality, which illustrates a potential ambiguity in such cases. See Appendix B for further discussion.

V.A Micro-foundations

Generally, very few assumptions are needed for an inequality externality to exist. Several different channels can be directly created from simple and mechanical microfoundations that do not rely on agents' emotional reactions, as we show in the following three simplistic examples:[59]

- Political polarization: Assume that political opinions O_i are a linearly increasing function of individual income x_i and no other factors (for simplicity). Political polarization, denoted as

[57] Inequality measures generally have $\frac{\partial \theta}{\partial x_i} \neq 0$ for virtually all agents. The absolute Gini coefficient, for instance, can be written as $I_{\text{Gini}}(\mathbf{x}) = \sum_{i=1}^{n} \kappa(x_i) x_i$,where the indexing of i has been chosen in increasing order of x_i, such that $\kappa(x_i) := \frac{1}{n}\left[2\frac{i}{n} - \frac{1}{n} - 1\right]$. Evidently $\frac{\partial I_{\text{Gini}}}{\partial x_i} = \kappa(x_i)$.

[58] Even though the individual's marginal effect on the inequality metric is small (of the order $\frac{1}{n}$), it being summed over n agents creates a non-negligible welfare effect on the same order of magnitude as marginal changes in consumption.

[59] An overbar indicates a society-wide variable. Bold indicates a population-sized vector.

$\bar{P} = \varphi(\boldsymbol{O})$, is defined as an increasing function of a distributional metric of all opinions in the population $\boldsymbol{O}$. We assume that $\bar{P}$ enters into the individual's utility function $U_i(x_i, \bar{P}, ...)$. If income inequality $\bar{\theta} = I(\boldsymbol{x})$ increases, differences of opinion within the population mechanically increase as well, generally increasing $\bar{P}$ and affecting $U_i(...)$. Thus, inequality leads to more pronounced political polarization and subsequent individual utility impacts.[60]

- Innovation / Economic growth: Assume that agents view high inequality as an incentive to work such that h_i and thus x_i are increasing functions of income inequality $\bar{\theta} = I(\boldsymbol{x})$. If so, utility can be written as $U_i(x_i(\bar{\theta}), h_i(\bar{\theta}), ...)$ and inequality is an externality. Further, assume that there exists some societal variable which is positively increasing in total labor supply, such as economic growth rates $\bar{g}$ or innovation levels $\bar{L}$. If this variable has an independent effect on either individual utility $U_i(...)$ or productivity n_i, then the labor choice change has an additional welfare-relevant externality effect through $\bar{g}$ and/or $\bar{L}$.

- Income-sensitive taste for public goods: Consider the funding required for a public good project to be undertaken as $\widetilde{Q}_j$. Individual utility is defined as $U_i(x_i, \sum_j p_{i,j} q_{i,j}, ...)$, where the individual-specific quantity of public good j is $q_{i,j}$. Assume further that either the quantity $q_{i,j}$ or the taste parameter $p_{i,j}$ varies with income levels x_i. As an example, a new youth center may be most beneficial for low-income earners, whereas an expensive opera house could be preferred by high-income earners. If income inequality $\bar{\theta}$ increases, there is less agreement on which public goods to fund and fewer projects reach $\widetilde{Q}_j$. Larger income differences in this context leads to fewer completed public projects and lower individual utility in more unequal societies.

The above examples illustrate that inequality externality channels can be mechanical in nature and can exist under only mild assumptions.[61] We also create micro-foundations for inequality effects on trust, crime, and political capture in Appendix G. Before we move on, we note that these channels may imply cascading effects. For instance, increasing political polarization could increase crime rates and hamper economic activity. We present one specific case of such secondary effects;

- Social unrest: Assume that one of the channels discussed above decreases the utility of a subset of individuals. These individuals might then prefer a high fixed cost of social unrest to living in a society with high economic inequality. If these events affect the utility of all individuals, inequality can lead to individual utility losses even for agents who were not initially negatively affected by the inequality externality.

[60] The same argument also holds for diversity of opinions more generally. A different perspective is that increased income inequality could lead to a broader diversity of opinions, carrying a positive utility impact.

[61] Three qualifications should be noted here. First, it is not self-evident which types of inequality (income, wealth, status...) and which domains (neighborhood, country, global...) are relevant, nor which effects are likely to be large on which agents. For this paper we do not go beyond some illustrative calculations in fairly simple cases. Second, the transmission of some inequality effects are clear, such as the effect of inequality on the provision of public goods, while others are dependent on social context or perceived inequality. This implies that inequality effects can differ across societies that are equally unequal. Third, some effects are time-dependent: although not well-captured in single-period models, the basic argument remains the same.

This last point illustrates that the impacts of inequality externality effects can be starkly discontinuous. In such events the externality itself would have complex optimal policy consequences as a low-probability, high-impact catastrophe in the vein of Weitzman (2009).

V.B Consequences in the literature

In the above we have shown how the classical OIT model is affected by various types of inequality externalities. The modifications to the classical model are relatively large. Given that the inequality externality is harder to ignore than many other externalities, a natural question is how other optimal policy models would be affected by the inclusion of an inequality externality. While this is too large of a question to fully answer in this paper, we present a few thoughts below.

First, our results question the external validity of models which rely on utility functions that only take into account individuals' income and work hours in large-scale settings. This is particularly true for numerical solutions in models focusing on inequality-related issues. As a recent example of how policy discussion can be modified through the introduction of an inequality externality we examine the model in Heathcote et al. (2020), the 2019 *EEA Presidential Address* titled "*How should tax progressivity respond to rising income inequality?*". The work analyzes an optimal taxation model in a general equilibrium framework where the main benefit of higher progressivity is as insurance for idiosyncratic shocks. The authors find that tax progressivity should remain approximately unchanged given rising U.S. inequality levels, a result which is robust in both a Rawlsian and Utilitarian framework. Introducing an inequality externality would likely affect these results. Following our results (which admittedly come from a simpler model), a negative (positive) inequality externality would likely yield a more progressive (regressive) optimal tax rate. The methodology in Heathcote et al. (2020) is relatively standard, and similar models are common in the economic literature. In general, we believe it would be prudent to check such results for robustness in the face of various inequality externalities or mention the no-externality assumption explicitly.

Second, theoretical models focusing on the trade-offs between different forms of taxation such as Guvenen et al. (2019) and Jacquet and Lehmann (2021) could also be affected by an inequality externality. With an inequality externality the social planner has an added incentive to set the inequality level itself, which may be easier with one instrument or the other. Take the example of wealth taxation versus capital income taxation in Guvenen et al. (2019), where one instrument taxes a stock and the other a flow – if the externality itself is more dependent on either the stock or the flow, the relevant trade-off could be modified.

V.C Other potential mathematical formulations

It is a natural question to ask whether another type of mathematical structure can keep individualist utility functions while modeling resource inequality's societal effects. Here we consider several other ideas and detail where they succeed or fail to capture the complexity of a resource inequality externality.

Social welfare weights In general, utility-based social welfare weights cannot approximate the effects of an economic inequality externality. The optimal marginal tax rates in the mechanism design case, shown in Equation 7, illustrates one case where even best-designed social welfare weights would fail to approximate the inequality externality.

There are two main reasons for why social welfare weights poorly approximate a resource inequality externality. First, such weights discount *utility*, not resources, which implies that the individual's private labor decision is socially optimal. This is not true under an inequality externality. Second, unlike an inequality externality, social welfare weights cannot change individual behavior. In addition to these two points, approximating inequality's societal effects – real-world phenomena – through social welfare weights would imply a break with welfarist traditions in that the social weights would no longer be a purely philosophical concept.

In view of its prevalence in the literature, this conventional OIT approach is further discussed in Appendix B.

Generalized social welfare weights (Saez and Stantcheva, 2016) The generalized social welfare weights method – or income-based social welfare weights more broadly – make few predictions for individual behavior. As such, appropriately chosen "modified welfare weights", adjusted to include inequality externality concerns, can approximate the mathematical solutions from a resource inequality externality. This is visible in Equation 12, where the modified welfare weight $\bar{G}'(z) = \bar{G}(z) - \Upsilon(z)$ would equal our solution. However, there are two problems with this approach.

First, the weights become dependent on empirically estimated values such as individual labor elasticities or the local Pareto parameter. The intuition behind the elasticity case is simplest to explain. As the individual contribution to the inequality externality depends on the individual's income, the modified weight – which now takes into account the societal effects of income inequality – needs to account for any changes in the individual's labor decision. This is mathematically done through introducing the labor elasticity into the modified weight, which implies an unintuitive addition of empirical parameters into an otherwise philosophical concept.

Second, the modified weights can turn negative and thus implicitly break the Pareto principle. This happens when the marginal social welfare of income is negative. This explicitly breaks with the assumptions made in Saez and Stantcheva (2016).

Still, this approach might be useful in some cases. If modified in such a way, the modeler should be aware that the resulting welfare weights would be different in interpretation from the standard approach, as the modified weights would measure both philosophical issues and externality concerns put together, and could include negative values.

An additive resource inequality in the social welfare function If we move away from strict social welfare weights one could imagine a SWF of the form $\int_i g_i U_i(x_i, h_i, ...)di - \Gamma(\bar{\theta})$, as in Sen (1976), among others. Here U_i is a standard individualist utility function and $\Gamma(\bar{\theta})$ is some function of resource inequality. This can mathematically approximate the solutions from a resource inequality externality if and only if the externality does not affect individual behavior. In other words, this

is an accurate mathematical representation of the problem if the externality is fully separable and the number of agents is large. We mention this specifically as the mathematical set-up we use in Section III makes these assumptions for simplicity. In general, however, any inequality externality that affects individual behavior cannot be captured through such a modified social welfare function. Such a formulation assumes away most consumption-based inequality externality effects, for example. Intuitively it is also less clear to us whether the social planner should care about inequality effects if these effects do not affect individuals themselves.

VI Conclusion

This paper has introduced the concept of an *inequality externality* and has particularly focused on a *post-tax income* inequality externality.

Most standard models of welfarist policy design implicitly assume that income inequality has no societal effects. But as we have shown with microfounded examples, such effects likely exist and could be both numerous and important. They are often independent from individuals' personal feelings; if inequality increases crime, for example, even a selfish individual would prefer equality in the absence of other changes. Including such effects into simple welfarist models with only a combination of diminishing marginal utilities of income and social welfare weights is generally not possible. The inequality externality is thus intended as a simple and generalizable way to model these side-effects of economic inequality without having to specify the potentially numerous causal channels independently. The inequality externality concept itself is tractable and does not assume a direction to the externality, can include other-regarding preferences but does not require them, and can easily be extended to other dimensions such as wealth inequality or heterogeneous utility functions.

Introducing an income inequality externality to the welfarist framework leads income (in)equality itself to become a policy goal. Individual labor decisions become socially suboptimal, and the marginal social welfare of individual income can become negative. Frameworks known for only being self-selection problems – including the optimal taxation problem – take on a new externality dimension.

In the Mirrlees (1971) optimal income taxation model, the optimal non-linear tax structure becomes unambiguously more inequality-reducing with the introduction of a negative inequality externality. The two new terms in the optimal taxation formula correspond to the well-known mechanical effect and behavioral responses respectively, and represent (i) society's increased willingness to pay for redistribution, and (ii) the internalization of the individual externality on income.

We present three new insights to the optimal income taxation literature, all of which are relevant for tax design.

First: Optimal top marginal tax rates are largely determined by the magnitude of the inequality externality. We observe both very high top marginal tax rates (above 90%) when inequality is a significant social bad and very low optimal top tax rates (<30%) when inequality is a social good. Our median estimate is an 85% optimal top marginal tax rate. We thus find theoretical

support for several policy arguments previously unsupported by economic theory, including a near-maximum income (with a large negative externality) or low top tax rates under a Rawlsian social planner (with a large positive externality). The findings also imply that different beliefs about the magnitude of the inequality externality could be a potential source of political disagreement. An intuitive explanation of this finding is that individuals at the ends of the distribution naturally affect inequality the most, but only those at the top can be specifically targeted by marginal tax rate changes. Mathematically, it follows from the top of the distribution being the only place where the two consequences of a tax increase, the mechanical effect and the behavioral responses, both affect inequality in the same direction (a reduction). The direction of the optimal tax rate change at the top is thus unambiguous given the direction of the inequality externality. At the bottom, where the two channels have opposite equality impacts, the optimal tax change depends on model specification.

Second: The externality creates a trade-off between income inequality levels and tax revenue, which implies that the equality dimension of the optimal tax problem is more complex than simply choosing an appropriate social welfare function. The social planner must balance the benefit of higher tax revenue against the equality-related benefits from individuals shifting into leisure (from socially suboptimal high labor choices). This trade-off is particularly noticeable when inequality can change substantially with minimal revenue losses. Given that policy makers believe that inequality itself is concerning, the analysis presented here recommends more progressive taxes than those previously suggested by Saez (2001), Piketty et al. (2014), and others.

Third: The theoretical implications of the model change substantially. We find welfare *benefits* from high income elasticities due to the welfare gains from individuals shifting away from suboptimal labor choices, and as a consequence the U-shape of optimal marginal tax rates found in the classical literature is fragile to the inclusion of a negative inequality externality. The two new terms in the optimal taxation formula are more distributionally dependent than standard externality terms with constant marginal effects across the distribution; the Pigouvian term depends on the marginal externality in the tax bracket, whereas the social welfare weight-adjusting term depends on the average marginal externality above the tax bracket. Generally, many externality-related results take on a new significance.

Finally, we briefly discuss how our results could have policy implications beyond optimal income taxation. Given that many economic models rely on the assumption of no externalities, the idea of considering inequality's societal effects as an externality that cannot be captured by standard SWFs could have widespread implications. We encourage further work on the topic.

[FOR ONLINE PUBLICATION]

Appendix

A Discussion on Equations 1 and 2

Equations 1 and 2 show the following simplification:

$$U_i(x_i(\bar{\theta}), \bar{\theta}, \overrightarrow{\Psi}(\bar{\theta}), ...) \rightarrow \tilde{U}_i(\tilde{x_i}, \bar{\theta}, ...). \tag{17}$$

A skeptical reader may argue that we should rather explore each externality channel individually. For example, if we assume that income inequality increases the amount of crime, one might say that one should strengthen the prevention of crime rather than reduce income inequality, or explore the crime-channel more in depth instead of focusing on inequality itself as an externality. To this we note two points.

First, a very similar simplification is usually made implicitly when including x_i in the utility function. Individuals may care about consumption *per se*, but they might also care about what consumption brings them – such as improved health. As in our case, there are many such potential channels that can be captured in the vector $\overrightarrow{\Psi}$ which are usually not explicitly modeled. In effect, the following simplification is implicitly made,

$$U_i(x_i, \overrightarrow{\Psi}(x_i), ...) \rightarrow \hat{U}_i(x_i, ...), \tag{18}$$

where $\hat{U}_i$ is the modified utility function – the utility function that is largely used in practice. In effect, a consumption-dependent utility function is a useful shorthand for what is in reality a much more complicated concept. The concept we introduce in this paper simply employs the same method with the *distribution* of individual income.[62]

Second, channel-specific policy solutions would also carry an associated cost which should be modeled in the general framework. Indeed, the main argument in this work is that different levels of (in)equality carry a shadow price that need to be taken into account when choosing between tax schedules – and not that the income tax is the best policy-specific solution to every inequality externality channel. What does this mean in a practical example? We return to the example of inequality and crime. Suppose that a tax increase leads to a higher level of income inequality, and that a preventative system for crime is a more effective policy against this increase than a direct change to the income tax structure. Importantly, such a preventative system would carry a cost. This cost would be a consequence of inequality and would thus imply that inequality has a shadow price. In other words, there is an additional cost that arrives as a direct consequence of the less redistributive tax system which should thus be taken into account when designing the optimal tax

[62] One may also argue that the average income matters for the individual in a similar way. We note that the mathematical analysis in Kanbur and Tuomala (2013) addresses this point.

system.

B Varying welfare weights

Another approach to introducing a dislike of inequality, common in the optimal income taxation literature, is varying the utility-based social welfare weights. The weights vary with utility such that the derivative of the SWF, $W'(U(n))$, is non-constant. The most widely used case is that of *decreasing* social welfare weights, such that the welfare of the wealthy is weighted less. Such weights are often presented as social inequality aversion, as it implies that the social planner values utility equality in itself.

There are three significant differences between this approach and the individual inequality externality we use in this paper. The first of these points holds only when discussing a *resource* inequality externality, as we do in most of this paper. The second and third hold under a utility inequality externality as well.

First: Using only social weights and absent other distortions, there is no difference between the optimality of the private and social labor supply choice. *Utility* is discounted, not *income* (or *resources* more broadly). Agents make the socially correct work decision, which they do not in our model.

Second: Under only social weights, individual behavior is not affected by any other individual. If inequality is an externality, the resources or utility of other individuals can affect individual utility and thus behavior. A natural example would be an agent who increases their work effort to avoid a high inequality externality imposed on low-income agents in a heterogeneous inequality externality framework. This changes the implications of the exercise dramatically, from a pure self-selection problem (the standard problem, Stiglitz (1982)) to an externality and self-selection problem (our problem). In the standard case, any model consequences must come through the social planner's actions.

This point can also be framed in the following way. In the standard framework, a reduction of inequality is not felt by other individuals. This means that equality is not beneficial *per se*; it is only beneficial if income is actually redistributed.

Third: If the model attempts to capture inequality's societal effects through social welfare weights, the choice of a social planner is no longer a purely philosophical concept. This is problematic as it conflicts with standard welfarist traditions in several ways. The most obvious case is when large inequality effects lead to negative social welfare weights, which breaks the Pareto principle. Another counter-intuitive example is how a truly Utilitarian social planner would need to use non-Utilitarian social welfare weights to take account of inequality's societal effects. Moving to an inequality externality allows us to return to standard considerations when setting up the social welfare function at the same time as we allow for any inherent effects of inequality.

These points emphasize our larger argument, which is that there are three distinct ways to model the consequences of inequality in a welfarist framework; the cumulative effect of diminishing marginal utility, social welfare weights, and an inequality externality. These are distinct, occur

through different mechanisms, and have different policy implications.

We now present a simple example to illustrate how a resource inequality externality can add nuance that cannot be found when only using social weights and the diminishing marginal utility of income. Imagine a world where one agent has seized the vast majority of income and uses this inequality of income to enjoy disproportionate (and socially damaging) political power. All other agents are equally poor. Now, imagine reducing the income of the oppressive ruler slightly, all else equal. We evaluate this change in the presence of *only* (i) risk aversion (diminishing marginal utility), (ii) a weighted social welfare function with non-negative weights, and (iii) an inequality externality.[63]

(i) Social welfare is unambiguously reduced, as the top individual's income decreases.

(ii) Social welfare is either reduced or kept constant – the top individual's income decreases, but they might have a zero social weight.

(iii) The effect on social welfare is ambiguous. On one hand, the income of the top individual is reduced, reducing their utility and thus social welfare (if their weight is non-zero). On the other, income inequality is reduced, increasing every other agent's utility. The total effect on social welfare depends on the size of the inequality externality. In extreme cases, such as in this example, overall social welfare might *increase*.[64]

More generally, diminishing marginal utility of income and social welfare weights present no intrinsic externality issues. As such, concentrated income gains lead to unambiguously non-negative welfare changes in standard models. Considering the current academic and social focus on inequality, this could be a troubling feature.

We note that the social weights discussed here are in terms of utility. Income-based weights, which share some of these problems, are discussed further in Section V.C.

Below we present a proof below to show that appropriate utility-based social welfare weights cannot supplant an inequality externality.

B.I. Proof: The inequality externality cannot be approximated by social weights

The social planner aims to maximize:

$$W = \int_i g_i U(x_i, h_i, \theta(\boldsymbol{x}))di$$

Assume that g_i can have variation (social weights), and that $\frac{\partial U}{\partial \theta} \neq 0$ and $\frac{\partial \theta(\boldsymbol{x})}{\partial x_i} \neq 0$ (an inequality externality exists). x_i is income, h_i is hours worked, and $\theta(\boldsymbol{x})$ is inequality as a function of all incomes $\boldsymbol{x}$.

[63] The 'standard' case here is no risk aversion, a utilitarian welfare function, and no externality. For example, the first case will consider reducing the income of the top earner in a model with risk aversion, a utilitarian social welfare function and no externality.

[64] Further, other individuals might change their labor market behaviors following the change, leading to secondary welfare consequences.

It follows from the social planner's first-order conditions for x_i and h_i that for all $g_i \neq 0$:

$$\frac{\partial U(x_i, h_i, \theta(\boldsymbol{x}))}{\partial h_i} = \frac{\partial U(x_i, h_i, \theta(\boldsymbol{x}))}{\partial x_i} + \frac{1}{g_i} \int_j g_j \frac{\partial U(x_j, h_j, \theta(\boldsymbol{x}))}{\partial \theta(\boldsymbol{x})} \frac{\partial \theta(\boldsymbol{x})}{\partial x_i} dj \tag{19}$$

We proceed with a proof by contradiction. Say we want to approximate the effect of the inequality externality with new social weights $\hat{g}_i$ without explicitly including θ in the utility function, otherwise keeping the utility function the same. Denote this new utility function $\hat{U}$. If so, $\frac{\partial \hat{U}(x_j, h_j)}{\partial \theta(\boldsymbol{x})} = 0$ and the second term on the right-hand side of Equation 19 is zero. The solution to the social planner's problem would thus involve $\frac{\partial \hat{U}(x_i, h_i)}{\partial x_i} = \frac{\partial \hat{U}(x_i, h_i)}{\partial h_i} \forall \hat{g}_i \neq 0$, which is equivalent to $\frac{\partial U(x_i, h_i, \theta(\boldsymbol{x}))}{\partial x_i} = \frac{\partial U(x_i, h_i, \theta(\boldsymbol{x}))}{\partial h_i} \forall \hat{g}_i \neq 0$. However, in the correct solution we are trying to approximate, $\frac{\partial U(x_i, h_i, \theta(\boldsymbol{x}))}{\partial x_i} \neq \frac{\partial U(x_i, h_i, \theta(\boldsymbol{x}))}{\partial h_i} \forall\, g_i \neq 0$. This implies that $g_i \neq 0 \rightarrow \hat{g}_i = 0$, which cannot be the case. Thus there is a contradiction. This follows from the externality creating a difference between the optimal individual and social work decisions, which cannot be introduced through discounting utility with social weights.

An extension shows that the externality cannot be approximated by the individual parameters in the utility function. If x_j is changed, Equation 19 implies that it will affect the FOC for i. In the modified solution with $\hat{U}$, it has no effect. To correctly specify $\hat{U}(x_i, h_i)$, one would need x_j or h_j. This would amount to including a distributional parameter $\theta(\boldsymbol{x})$ in the individual utility function, again a contradiction.

C Analytical Solution of the OIT Problem

We write individual utility as;

$$U(x, h, \bar{\theta}) = u(x) - V(h) - \Gamma(\bar{\theta}) \tag{20}$$

where u is the utility of consumption (after-tax income), V is the disutility of work and Γ is the disutility of inequality. Equation (20) assumes that agents are homogeneous, with identical individual utility functions.

At the heart of the model is n, the exogenous wage-earning ability, unobservable to the social planner. There is a continuum of individuals with n varying according to a an exogenous density function $f(n)$, with a cumulative distribution function $F(n)$. Pre-tax earnings are defined as nh, and total consumption is $x = nh - T(nh)$, where $T(\cdot)$ is the tax schedule. The individual maximizes utility by choosing hours worked h given n and $T(\cdot)$. The utility-maximising values of consumption and hours worked are written as

$$x(n), h(n). \tag{21}$$

Given the individual's choice, the social planner chooses the tax schedule to maximize the social welfare function. We assume this to be an additively separable function of individual utility.

Accordingly the problem is,

$$\max_{T(\cdot)} \int_{\underline{n}}^{\overline{n}} W(U(x(n), h(n), \bar{\theta}))dF(n). \tag{22}$$

Notice that formulating individual utility as (20) avoids the complication of potentially heterogeneous effects of inequality if the social planner is strictly Utilitarian (Benthamite) – in this case only the average inequality externality has an effect. Similarly, a Rawlsian social planner will only take into account the inequality externality on the lowest-utility agent.

The problem (22) is subject to three conditions, the first two of which are standard constraints. First, there is the *revenue constraint* for any required amount R of non-redistributive public goods:

$$R \leq \int_{\underline{n}}^{\overline{n}} T(nh)f(n)dn. \tag{23}$$

For simplicity we assume that $R = 0$.

Second, we have the *incentive-compatibility constraint* from the possibility that an agent with (unobservable) wage-earning ability n could masquerade as an agent with $\hat{n}$. For any person with wage-earning ability n it must be true that:

$$u(x(n)) - V(h(n)) \geq u(x(\hat{n})) - V(h(\hat{n})) \tag{24}$$

where $x(\hat{n})$ and $h(\hat{n})$ are, respectively, the consumption and hours worked if the agent masquerades as someone with ability $\hat{n}$, possibly different from n. The IC constraint (24) ensures that the agent self-selects into the appropriate tax bracket.

Third, we need to introduce the role of inequality into the model. Individuals experience an amount $\bar{\theta}$ of after-tax inequality. This inequality is partly determined by F, the distribution of innate talent, and partly by the choices made by individuals, captured in (21). But it is also partly the result of decisions by the social planner, captured in the tax function T and therefore embedded in (21). We can represent this relationship as the following *inequality condition:*

$$\bar{\theta} = I(\boldsymbol{x}, F) \tag{25}$$

where $I(\cdot, \cdot)$ is an inequality measure, $\boldsymbol{x}(\cdot)$ is the full set of consumption choices from (21) and $F(\cdot)$ is the distribution function for n.

To complete the model we need an inequality metric $I(\cdot, \cdot)$. We use a specific form of the (absolute) Gini coefficient in after-tax income:

$$I_{\text{Gini}}(\boldsymbol{x}, F) = \int_{\underline{n}}^{\overline{n}} \kappa(n)x(n)dF(n), \tag{26}$$

where x is after-tax income (consumption), n is the exogenous productivity level, and

$$\kappa(n) = 2F(n) - 1 \tag{27}$$

is an expression for the weight of the agent in the Gini.[65] Expression (26) shows that the absolute Gini can be calculated as a sum of weighted incomes in the population, where the weight $\kappa(n)$ depends only on the *rank* of the agent in the wage-earning ability distribution, which is constant and exogenous by assumption. Using (26), condition (25) becomes

$$\bar{\theta} = \int_{\underline{n}}^{\overline{n}} [2F(n) - 1]\, x(n) dF(n).$$

One can also use other inequality metrics based on rank-specific weights, such as those in the Lorenz (Aaberge, 2000) or S-Gini families (Donaldson and Weymark, 1980).

With the inequality externality and inequality metric specified, we note that if the inequality externality $\Gamma(\bar{\theta})$ is linear and we are in a Utilitarian framework, the objective function amounts to the SWF derived in Sen (1976) with an additional labor disutility term. This Sen (1976) SWF is also a cumulation of Fehr-Schmidt preferences over the population (Schmidt and Wichardt, 2019), creating another link to the inequality aversion literature.

To solve the analytical problem we first re-write the incentive compatibility constraint. We note that consumption x, i.e. after-tax income, is a function of wage times hours worked: $x = c(nh)$. The individual maximization implies,

$$\frac{dU}{dh} = 0 = u'c'n - V', \tag{28}$$

and from the IC constraint we have (using either the Mirrlees (1971) trick or the envelope condition):

$$\frac{dU}{dn} = u'c'h \tag{29}$$

Taken together these two imply :

$$\frac{dU}{dn} = \frac{V'h}{n} =: g\,(n) \tag{30}$$

We can write $T = nh - x$, where x is after-tax consumption.[66] From this and the IC constraint, we observe that the tax schedule implicitly defines both work hours and total individual utility.

[65] This is a slight modification of Equation 27 in Cowell (2000) for the standard (relative) Gini $\int \frac{\kappa'(x(n))x(n)}{\mu(x)} dG(x)$, where $\kappa'(x) = 2G(x) - 1$ is the weight of the agent and $G(x)$ is the CDF of x. If individuals' post-tax income increases with wage-earning ability, the rank-dependent variable $\kappa(n) = \kappa'(x)$. In other words, if there is rank-equivalency between income and ability, we can use the ability ranking to calculate the individual weights in the income inequality metric. Simula and Trannoy (2022b), developed simultaneously with this paper, also exploits this rank-invariancy in ability and income. It is a novel method and vastly simplifies the analytical problem.

As we show in Appendix C.I, this assumption is equivalent to assuming that the individuals' second-order conditions hold. For all the numerical simulations we confirm that they in fact do.

[66] The model is a one-period model and does not contain savings.

Instead of setting the tax schedule T, then, we can say that the social planner chooses work hour schedules $h(n)$, utility schedules $U(n)$, and the inequality level $\bar{\theta}$.

The Lagrangian of the full problem classified in Equations 22–25 is,

$$\begin{aligned} L = \int_{\underline{n}}^{\overline{n}} W(U(n)) f(n) dn + \lambda \Big(\int_{\underline{n}}^{\overline{n}} \left[nh(n) - x \right] f(n) dn \Big) \\ + \int_{\underline{n}}^{\overline{n}} \alpha(n) \left[\frac{dU}{dn} - g(n) \right] dn + \gamma \left[\bar{\theta} - I_{Gini} \right] \end{aligned} \tag{31}$$

We note that the incentive compatibility constraint can be simplified using integration by parts, and we assume n goes from zero to infinity without loss of generality. After taking these factors into account and combining the rest of the integrals, we have:

$$\begin{aligned} L = \int_0^{\infty} \left[W(U(n)) + \lambda(nh(n) - x) \right] f(n) - \alpha(n) g(n) - \alpha'(n) U(n) dn \\ + \alpha(\infty) U(\infty) - \alpha(0) U(0) + \gamma \left[\bar{\theta} - I_{Gini} \right] \end{aligned} \tag{32}$$

We introduce the Gini coefficient in the form,

$$I_{Gini} = \int_0^{\infty} \left[2F(n) - 1 \right] x f(n) dn = \int_0^{\infty} \kappa(n) x f(n) dn \tag{33}$$

Where $f(n)$ and $F(n)$ are the PDF and CDF of n, respectively, and $\kappa(n) = 2F(n) - 1$ is the weight of the agent in the absolute Gini.

The Lagrangian becomes:

$$\begin{aligned} L = \int_0^{\infty} \Big[\left(W(U(n)) + \lambda \left[nh(n) - x \right] - \gamma \kappa(n) x \right) f(n) - \alpha(n) g(n) \\ - \alpha'(n) U(n) \Big] dn + \alpha(\infty) U(\infty) - \alpha(0) U(0) + \gamma \bar{\theta} \end{aligned} \tag{34}$$

From this we can find the first-order conditions with respect to $h(n)$, $U(n)$, and $\bar{\theta}$, as these variables together will implicitly set the tax schedule.[67] Before we begin, note that we can rewrite $x = y(h, U, \bar{\theta}) = u^{-1}(U + V(h) + \Gamma(\bar{\theta}))$, and find expressions for the derivatives y_h, y_U, and $y_{\bar{\theta}}$.[68] The first order conditions are the following:

$$U: \quad 0 = \left[W_{U(n)} - \lambda y_U \right] f(n) - \alpha'(n) - \gamma \kappa(n) f(n) y_U \tag{35}$$

$$h: \quad 0 = \lambda(n - y_h) f(n) - \alpha(n) \frac{V_{hh} h + V_h}{n} - \gamma \kappa(n) f(n) y_h \tag{36}$$

[67] We could use the derivative of $x(n)$ instead, but the methods are mathematically equivalent and this procedure is somewhat more straightforward.

[68] Using the rules for derivatives of inverse functions, these expressions are $y_h = \frac{V_h}{u_x}$, $y_{\bar{\theta}} = \frac{\Gamma_{\bar{\theta}}}{u_x}$, and $y_U = \frac{1}{u_x}$.

$$\bar{\theta}: \quad 0 = \gamma - \int_0^\infty \gamma\kappa(n)f(n)y_{\bar{\theta}}dn - \int_0^\infty \lambda y_{\bar{\theta}} f(n)dn \tag{37}$$

In the FOC for h we have used that $g = \frac{V_h h}{n}$ from Equation (30), and that $\frac{dg}{dh} = \frac{V_{hh}h+V_h}{n}$. Equation 37 implies,

$$\frac{\gamma}{\lambda} = \frac{\int_0^\infty \frac{\Gamma_{\bar{\theta}}}{u_x} f(n)dn}{1 - \int_0^\infty \frac{\Gamma_{\bar{\theta}}}{u_x}\kappa(n)f(n)dn} \tag{38}$$

Here $\frac{\gamma}{\lambda}$ is the shadow price of the inequality constraint expressed in units of public funds, and $\Gamma_{\bar{\theta}}$ and u_x are derivatives. If $\Gamma(\bar{\theta}) = 0$, as in the standard case when the inequality externality does not exist, then $\gamma = 0$. A negative inequality externality implies a positive $\Gamma_{\bar{\theta}}$, and thus a positive $\frac{\gamma}{\lambda}$. To rephrase, this is the unsurprising result that equality itself has a cost in a world with a negative inequality externality.

Now we move to finding an expression for $\alpha(n)$, the shadow price of the incentive compatibility constraint. We integrate the first order condition for U, Equation 35:[69]

$$\alpha(n) = \int_n^\infty \left[\frac{\lambda + \gamma\kappa(p)}{u_{x(p)}} - W_{U(p)}\right] f(p)dp \tag{39}$$

And substitute this into Equation (36):

$$0 = \lambda(n - y_h)f(n) - \gamma\kappa(n)f(n)y_h - \frac{V_{hh}h + V_h}{n}\int_n^\infty \left[\frac{\lambda + \gamma\kappa(p)}{u_{x(p)}} - W_{U(p)}\right] f(p)dp \tag{40}$$

$$\frac{(n - y_h)}{y_h} = \frac{\gamma}{\lambda}\kappa(n) + \frac{u_{x(n)}(V_{hh}h + V_h)}{\lambda f(n) n V_h}\int_n^\infty \left[\frac{\lambda + \gamma\kappa(p)}{u_{x(p)}} - W_{U(p)}\right] f(p)dp \tag{41}$$

We have that $\frac{n-y_h}{y_h} = \frac{nu_{x(n)}}{V_h} - 1 = \frac{1}{1-t} - 1 = \frac{t}{1-t}$, so we quickly have the expression for optimal marginal tax rates shown in Equation 7:

$$\frac{t}{1-t} = \frac{\zeta_n u_{x(n)}}{f(n)n}\int_n^\infty \left[\frac{1}{u_{x(p)}} - \frac{W_{U(p)}}{\lambda}\right] dF(p) + \frac{\gamma}{\lambda}\left[\kappa(n) + \frac{\zeta_n u_{x(n)}}{f(n)n}\int_n^\infty \frac{\kappa(p)}{u_{x(p)}} dF(p)\right], \tag{42}$$

Here $\frac{\gamma}{\lambda}$ is the price of inequality in terms of public funds (see Equation 38). If inequality is a negative externality (a public bad), γ will generally be large and positive.[70] The agent's weight in the Gini coefficient, $\kappa(n)$, is negative at the bottom and positive at the top. $\zeta_n = \frac{V_{hh}h}{V_h} + 1$ is a term closely related to the inverse compensated elasticity of labor,[71] and u_x is the marginal utility

[69] From the transversality conditions $\frac{dL}{dU(0)} = \alpha(\infty) = 0$. We use the new symbol p to denote the productivity n inside the integral.

[70] If we assume a linear inequality externality of the form $\Gamma(\theta) = \eta\theta$ then $\frac{\gamma}{\lambda} = \eta$ (see Equation 38).

[71] With quasi-linear preferences, $\zeta = \frac{1}{E_L} + 1$.

of consumption.

Two of these terms are equivalent to traditional OIT terms. By denoting the part of the optimal tax function found in Diamond (1998) as $\frac{t_i}{1-t_i}$, we can isolate and evaluate the effect of the inequality externality.

$$\frac{t}{1-t} = \frac{\gamma}{\lambda}\left[\kappa(n) + \frac{\zeta}{f(n)n}\int_n^\infty \frac{u_{x(n)}}{u_{x(p)}}\kappa(p)dF(p)\right] + \frac{t_i}{1-t_i} \tag{43}$$

For clarity let us assume a linear homogeneous inequality externality ($\Gamma(\bar{\theta}) = \eta\bar{\theta}$) and quasi-linearity in consumption.[72] The optimal tax rate condition simplifies to:

$$\frac{t}{1-t} = \eta\kappa(n) + \eta\left(1 + \frac{1}{E_L}\right)\Pi(n)F(n) + \frac{t_i}{1-t_i}, \tag{44}$$

where we denote the distributional thinness measure $\frac{1-F(n)}{f(n)n}$ as $\Pi(n)$.[73] This formula is functionally equivalent to the form we find with the small perturbations method (Equation 12), but uses the exogenous wage n instead of the endogenous earnings z.

C.I. Equivalence of income rankings

In using the modified Gini in Equation 4, we have assumed that the weight of the agent in the ability ranking is the same as the ranking of the agent in the post-tax income ranking. We asserted that this is equivalent to the second-order condition holding, or that $z'(n) > 0$ where $z(n)$ is pretax income (Lollivier and Rochet (1983)). This is not necessarily obvious. Recall that we have a monotonically increasing n: if we have that $x'(n) > 0$, then, we also have the desired equivalence in ability and post-tax rankings. The more standard assumption in the literature is the SOC $z'(n) > 0$. Here we show that $x'(n) > 0$ is equivalent to $z'(n) > 0$.

Assume quasi-linearity for simplicity and define $\Omega(n) = x(n) - V(\frac{z(n)}{n})$. Here $\Omega(n) \geq \Omega(\hat{n})\,\forall\, n, \hat{n}$ is equivalent to the IC constraint. The problem becomes

$$max_{V,y} \int \left[\Omega(n) - \Gamma(\bar{\theta})\right] dG(n)$$

$$s.t. \int \left[\Omega(n) + V(\frac{z(n)}{n}) - z(n)\right] dF(n) \leq 0,$$

[72] The resulting utility function is

$$U(x, h, \bar{\theta}) = x - \frac{h^{\left(1+\frac{1}{E_c}\right)}}{\left(1+\frac{1}{E_c}\right)} - \eta\bar{\theta}$$

Note that with quasi-linearity, $\int_n^\infty \kappa(p)dF(p)$ in (43) simplifies as $\int_n^\infty (2F(n)-1)\,dF(n) = F(n) - F(n)^2$.

[73] This is the inverse of the local Pareto parameter $\alpha(n)$, which becomes constant in a Pareto distribution. It is also the inverse elasticity of $P(n) = 1 - F(n)$ with regards to n; $\varepsilon_{P,n} = \frac{n}{1-F(n)}\frac{d(1-F)}{dn} = -\frac{nf(n)}{1-F(n)}$.

$$\Omega'(n) = \frac{z(n)}{n^2} V'(\frac{z(n)}{n}),$$

$$\bar{\theta} = I_{Gini},$$

where the second constraint is the individual's FOC. Then we note that:

$$x'(n) = \Omega'(n) + \left(\frac{nz'(n) - z(n)}{n^2}\right) V' = \left(\frac{z(n) + nz'(n) - z(n)}{n^2}\right) V' = \frac{z'(n)}{n} V'$$

And we have the sought-after equivalence; n and $V'(\frac{z(n)}{n})$ are positive, so $z'(n) > 0$ implies $x'(n) > 0$.

Finally, a word of caution: $\frac{t}{1-t}$ can fall below -1 at the bottom of the distribution given a sufficiently large negative externality if everyone works.[74] This is in reality not a solution, as the second-order conditions are violated and the assumption behind the ability-income rank equivalence fails. This example illustrates why our analytical specifications must be taken with caution; in certain settings, and particularly with large externalities, additional constraints should be added. A similar edge case can occur at the top with a large positive externality.

C.II. A squared inequality externality function

Our framework is sufficiently general for other functional forms of the MRS, or equivalently $\Gamma(\bar{\theta})$, the inequality function from the utility function (see Appendix C). Let us use $\Gamma(\bar{\theta}) = \eta(\bar{\theta} - \bar{\theta}_{opt})^2$, such that:

$$U(x, h, \bar{\theta}) = x - \frac{h^{\left(1+\frac{1}{E_c}\right)}}{\left(1+\frac{1}{E_c}\right)} - \eta(\bar{\theta} - \bar{\theta}_{opt})^2 \tag{45}$$

The resulting analytical optimal tax rates are:

$$\frac{t}{1-t} = 2\eta(\bar{\theta} - \bar{\theta}_{opt}) \left[\kappa(n) + \frac{\zeta}{f(n)n} \int_n^\infty \kappa(p) f(p) dp\right] + \frac{t_{orig}}{1 - t_{orig}} \tag{46}$$

Comparing these tax rates to Equation 43, we see that the effect of the inequality externality is attenuated by a factor of $2(\bar{\theta} - \bar{\theta}_{opt})$. The policy effect of the inequality externality will be larger in societies with high after-tax inequality. We find this intuitive; tax systems responding to inequality will respond more when initial inequality is high. The result is the same when using the small perturbations method.

Also note that this solution is endogenous, as $\bar{\theta}$ depends on the tax schedule. We thus need numerical methods to solve for the optimal tax schedule. This is not a unique feature of this formulation, and also occurs when the social weights are endogenous as in the non-Rawlsian solutions.

[74] The numerical simulations always have an atom of non-working individuals at the bottom to prevent this.

We do not perform numerical simulations in this case, primarily because of the complicated nature of estimating a suitable η when we have another unknown variable in $\bar{\theta}_{opt}$.

D Small Perturbation Solution to the OIT Problem

The core part of this approach follows Saez (2001) and Saez and Stantcheva (2016).

We introduce a small tax reform $d\tau_z$ where the marginal income tax is increased by $d\tau$ in a small band from z to $z+dz$. The reform mechanically increases average tax rates on everyone above this band. This is the mechanical effect of taxation, and collects $dz\partial\tau$ from $1-F(z)$ agents above z under the assumption of no income effects. Thus it collects $[1-F(z)]\,dz\partial\tau$ revenue. For each $dz\partial\tau$ collected, however, inequality also changes. The magnitude of this change per agent above differs based on which agent is considered. Noting that income rank $\kappa(z)$ does not change, each decrease in one unit of post-tax income at z changes absolute post-tax income inequality by $\kappa(z)f(z)$ (from Equation 4).[75] The mechanical effect thus has a differing equality effect of $-\kappa(z_j)f(z_j)dz\partial\tau$ at each point j above z, where z_j is the income of the agent and $f(z_j)$ is the number of agents at this point, and $\kappa(z_j)$ is that agent's weight in the inequality metric. As the income change of each agent above z is equal, we can define the average inequality weight above as $\overline{\kappa}(z)\,[1-F(z)] = \int_{\{j:z_j>z\}}\kappa(z)f(z)dj$ and write that the mechanical effect changes income inequality by $d\bar{\theta}_M = -\overline{\kappa}(z)\,[1-F(z)]\,dz\partial\tau$.[76]

Those who are located in the small band between z to $z+dz$ have a behavioral response to the tax change. They work less, and reduce their pre-tax earnings by an amount $\partial z = -\epsilon(z)z\partial\tau/\left(1-\tau(z)\right)$. $\epsilon(z)$ is the elasticity of earnings z with respect to $1-\tau(z)$. There are $f(z)dz$ individuals in the tax bracket who were taxed at $\tau(z)$ before the perturbation, so total revenue decreases by $-dz\partial\tau\cdot\epsilon(z)zf(z)\tau(z)/\left(1-\tau(z)\right)$. This change in total earnings is moderated by an effect $(1-\tau)/\tau$ for the inequality effect, as we are interested in the post-tax income decrease and not the tax revenue decrease.[77] Additionally we must multiply by the agents' weight in the inequality metric $\kappa(z)$. The behavioral response thus has an effect on the post-tax income inequality metric as $d\bar{\theta}_B = -\kappa(z)\cdot dz\partial\tau\cdot\epsilon(z)zf(z)$.

The total revenue effects are:

$$dR = dz\partial\tau\left(1-F(z)-\epsilon(z)zf(z)\tau(z)/\left(1-\tau(z)\right)\right)$$

The direct welfare effect through the individual income channels is $\int_j g_j dRdj$ for $z_j \leq z$ and $-\int_j g_j(\partial\tau dz - dR)dj$ for $z_j > z$. Thus the net individual income-based welfare effect is $dM+dB+dW = dR\cdot\int_j g_j dj - dz\partial\tau\int_{\{j:z_j\geq z\}} g_j dj$.

The total equality effect is $d\bar{\theta} = d\bar{\theta}_M + d\bar{\theta}_B$:

$$\partial\bar{\theta} = dz\partial\tau\left(-\overline{\kappa}(z)\,[1-F(z)] - \kappa(z)\epsilon(z)zf(z)\right)$$

[75] As $\kappa(z)$ is negative at low income values, this can be negative.

[76] In the absolute Gini, $\overline{\kappa}(z) = F(z)$.

[77] For the mechanical effect, the tax revenue increase and the individual post-tax income decreases are identical.

In terms of utility, this affects every individual as $\int_j g_j \frac{\partial U_j}{\partial \bar{\theta}} \cdot \partial\bar{\theta} \cdot dj$. As we assume an homogenous inequality externality and quasi-linearity in consumption such that $\eta = MRS_{x\bar{\theta}} = -\frac{\partial U/\partial \bar{\theta}}{\partial U/\partial x} = -\frac{\partial U}{\partial \bar{\theta}}$, the total welfare effect of the inequality change is $dI = \int_j g_j \cdot (-\eta) \cdot \partial\bar{\theta} \cdot dj = -\eta \cdot \partial\bar{\theta} \cdot \int_j g_j dj$.

The total welfare change, including all channels, is equal to zero at the optimum:

$$dM + dB + dW + dI = 0.$$

Thus, using the expressions for dR and dI, and the expression $\bar{G}(z)\left(1 - F(z)\right) = \int_{\{j:z_j \geq z\}} g_j dj \big/ \int_j g_j dj$, we have:

$$dz\partial\tau \int_j g_j dj \left[1 - F(z) - f(z)\epsilon(z)z\frac{\tau(z)}{1-\tau(z)}\right] - dz\partial\tau\bar{G}(z)\left(1 - F(z)\right)\int_j g_j dj$$
$$+\eta \cdot \int_j g_j dj \cdot \left[dz\partial\tau\left(\overline{\kappa}(z)\left[1 - F(z)\right] + \kappa(z)\epsilon(z)zf(z)\right)\right] = 0$$

Dividing by $zf(z)\epsilon(z)\int_j g_j dj \cdot dz\partial\tau$ and re-arranging, we find:

$$\frac{\tau(z)}{1-\tau(z)} = \eta \cdot \kappa(z) + \frac{1 - F(z)}{z \cdot f(z)}\frac{\left(1 - \bar{G}(z) + \eta\overline{\kappa}(z)\right)}{\epsilon(z)}$$

We use the local Pareto parameter $\alpha(z) = \frac{z \cdot f(z)}{1-F(z)}$ and write $\Upsilon(z) = \eta\alpha(z)\epsilon(z)\kappa(z) + \eta\bar{\kappa}(z)$ and find the optimal marginal income tax rates as specified in Equation 12.

E Additional notes for Section III

E.I. Theoretical ability distributions

We present Rawlsian optimal marginal income tax rates from two theoretical skill distributions in Figure VII, using the Gini as the inequality metric. The first is is a Pareto distribution with $\alpha(n) = 2.0$, which becomes nearly identical to the empirical case at the top of the distribution.[78] The second is a lognormal distribution with $\mu = 2.757$ and $\sigma = 0.5611$, using the values from Mankiw et al. (2009) based on the 2007 U.S. wage distribution.

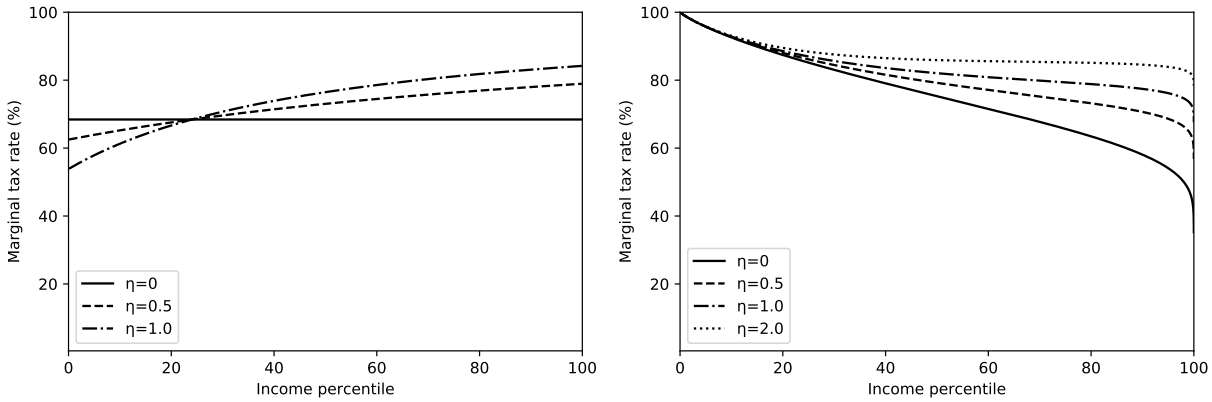
The Pareto case in Figure VIIa) illustrates the potentially positive effect of behavioral responses at the bottom. It is socially beneficial for low-income individuals to increase their incomes – so that inequality is reduced – which leads to a small income subsidy at the bottom as compared to the no-externality case. The goal of this tax subsidy is to make individuals internalize that their increased labor supply leads to positive societal outcomes.

The lognormal case further illustrates the localized effects at the top of the distribution. The standard top marginal tax rate in the lognormal case is 0%. With an inequality externality of $\eta = 2.0$ that increases to 67%. This illustrates the Pigouvian correction at the top, and is salient

[78] Here $\alpha(n) = 2.0$; in the numerical wage distribution $\alpha(n) = 1.9$. Under this Pareto distribution, second-order conditions fail at the bottom for $\eta = 2.0$. This is therefore not plotted.

Figure VII

Optimal Taxation with Inequality Externalities: Theoretical Ability Distributions



Note: Optimal marginal tax rates for various negative inequality externality magnitudes η. The social planner is Rawlsian and the productivity distribution is (a) a Pareto distribution with $\alpha(n) = 2.0$, (b) a lognormal distribution with $\sigma = 0.5611$ and $\mu = 2.757$. Inequality aversion estimates indicate $\eta = 1.0$. The solid line, $\eta = 0$, is the standard case of no inequality externality. See Table III for further explanation of the inequality externality magnitudes; $\eta = 2.0$ implies that a representative agent with mean income in a society with Denmark-like income inequality would be indifferent to increasing her income by 25% at the same time as income inequality increased to the United States' current income inequality level. The $\eta = 2.0$ case is excluded from the Pareto simulation because second-order conditions fail at the bottom. The elasticity of labor E_L is 0.3.

given the local "zero tax at the top"-result of standard models. This local result is not visible in the graph, but is borne out out in the simulations. At the 99^{th} percentile the marginal tax rate increases from 39% in the standard case to 79% when $\eta = 2.0$.

E.II. Varying inequality metrics

In the main specification we used the absolute Gini coefficient for our measure of inequality. Here we explore two different families of inequality metrics. The first is the top income shares also shown in the main text. The second is the S-Gini, which approximates the Gini with a larger focus on either end of the distribution. The distributional weights implied by both families are plotted in Figure VIII.[79]

1 Approximating top income shares The first family of inequality metrics, also used in the main robustness test, has some of the properties of top income shares. It is,

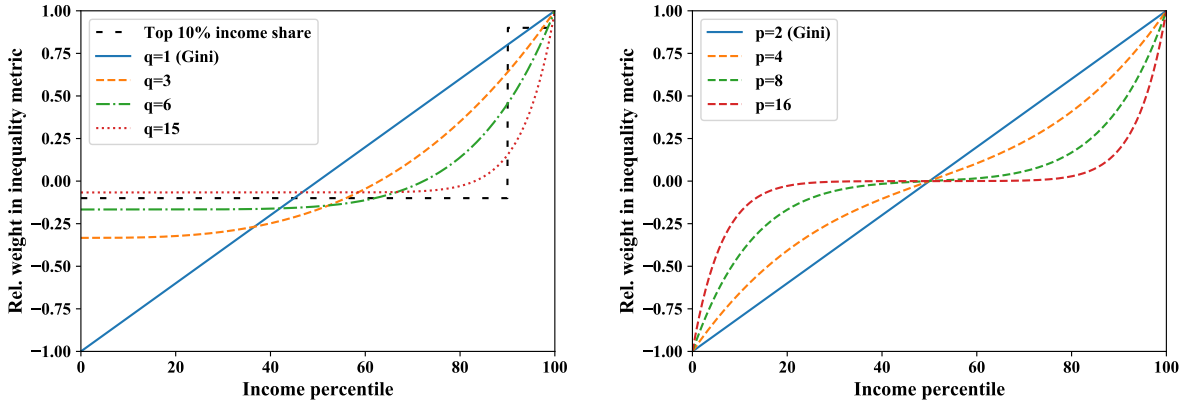
$$\bar{\theta} = \int_0^\infty \left[(q+1)F(n)^q - 1\right] x(n) dF(n),\ q \in \mathbb{N}. \tag{47}$$

When $q = 1$, this becomes the absolute Gini coefficient. In all cases, perfect equality implies $\bar{\theta} = 0$ and perfect inequality implies $\bar{\theta} = \mu$ (or $\bar{\theta} = 1$ in the non-absolute family). For increasing q, this

[79]The weights in Figure VIII are normalized such that the top weight is always 1.00. This normalization has no impact on our results due to our re-calculation of η before simulations.

Figure VIII

Weights for Families of Inequality Metrics



Note: Consumption weights for inequality metrics used in Appendix E.II. For each individual, their impact on the inequality metric is their proportional weight multiplied by their income. In both figures, the Gini is plotted in solid blue. (a) A family of inequality metrics similar to top income shares, as in Equation 47. The top 10% income share is plotted in dotted black for reference. (b) The S-Gini family from Equation 49.

indicates an increased focus on the very top of the distribution. The negative externality at the top becomes increasingly concentrated at the very top with increasing q, while the positive externality at the bottom becomes approximately constant for an increasing fraction of the population. In effect, increasing q leads to a metric closer to top income shares, but without the discontinuities that make the analytical problem intractable.

The resulting analytical optimal tax rates with the utility function in 13 become,

$$\frac{t}{1-t} = \eta_q \left[((q+1)F(n)^q - 1) + (1 + \frac{1}{E_c}) \frac{1}{f(n)n} \left[1 - F(n)^q\right] F(n) \right] + \frac{t_{orig}}{1 - t_{orig}}. \tag{48}$$

Here η_q is the magnitude of the inequality externality, which is dependent on q when fitting to empirical data. We ensure that values of η_q are comparable over simulations by re-calculating the parameter from experimental data for each q.[80]

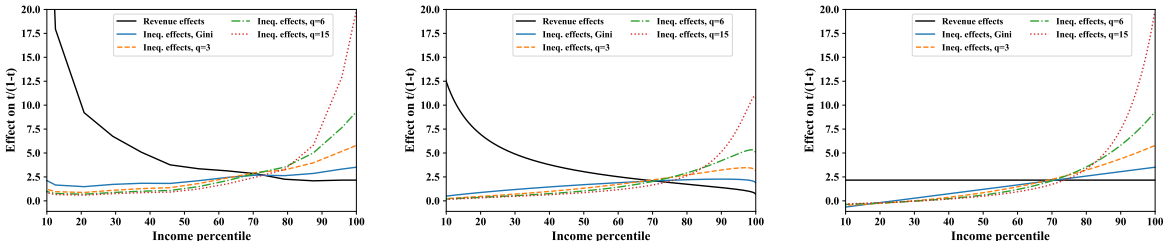
To further illustrate this point, we show the effect of both standard revenue considerations and the new equality considerations on $\frac{t}{1-t}$ with varying inequality metrics in Figure IX. We present this figure for several different underlying ability distributions. The interaction of equality and revenue considerations can make it difficult to interpret values of t, so this graph illustrates the more intuitive impact on $\frac{t}{1-t}$. All social planners are Rawlsian.[81]

[80] We estimated η with data from Carlsson et al. (2005) in the main text. To remain consistent, we have calculated for each inequality metric q comparable η_q from the experimental values in Carlsson et al. (2005) for all following simulations. This means that, while the value of η_q changes, the underlying estimation comes from the same data. This is true for all metrics.

[81] Equality considerations would not change with any other SWF due to the homogeneous nature of the externality. Revenue effects would decrease at the bottom and converge to the same at the top.

Figure IX

Effects on $\frac{t}{1-t}$: Top Income Share Externalities



Note: Effects on $\frac{t}{1-t}$ for various negative inequality metrics $\int_0^\infty \left[(q+1)F(n)^q - 1\right] x(n)dF(n),\ q \in \mathbb{N}$. The social planner is Rawlsian. The magnitude of the inequality externality is in each case calculated as the median value from the empirical inequality aversion estimates in Carlsson et al. (2005). This is done for comparability across inequality metrics. The productivity distribution is (a) the empirical wage distribution from the main text, (b) a log-normal distribution with $\sigma = 0.39$ and $\mu_{log} = -1$, and (c) a Pareto distribution with $a = 2$. See Figure VIII for an explanation of the inequality metrics. In particular, larger q indicates that top incomes are increasingly weighted. The elasticity of labor E_L is 0.3.

Several points are worth noting. First, as expected, increasing q leads to a more pronounced effect at the top of the distribution in all cases. Second, below the top the effects of changing the metric are small and generally dampen the effect of the externality. Third, equality considerations are relatively constant over different skill distributions; the major factor changing resulting tax rates over skill distributions are revenue considerations. Fourth, equality considerations are proportionally more important than revenue considerations towards the top of the distribution in all three cases. While by nature dependent on the ability distribution and social welfare function, this last point seems likely to hold in many specifications.

2 The S-Gini The second family of inequality metrics we use is the S-Gini family, which increases the weight of top- and bottom-incomes symmetrically.

$$\bar{\theta} = \int_0^\infty \left[F(n)^p - (1 - F(n))^p\right] x(n)dF(n),\ p \geq 2. \tag{49}$$

When $p = 2$, this becomes the absolute Gini coefficient. This family also retains the beneficial properties discussed above; perfect equality implies $\bar{\theta} = 0$ and perfect inequality implies $\bar{\theta} = \mu$. For increasing p, the top and bottom is increasingly weighted at the cost of middle incomes. Unlike the previous family, these metrics will always increase if an individual above the median increases their income, as well as decrease if an individual below the median increases their income. The resulting optimal tax rates with the utility function in 13 are,

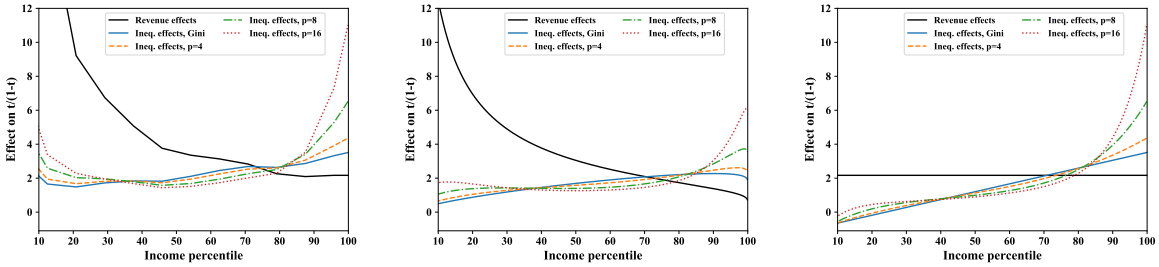
$$\frac{t}{1-t} = \eta_p \left[(F(n)^p - (1 - F(n))^p) + (1 + \frac{1}{E_c})\frac{1}{f(n)n}\nu\right] + \frac{t_{orig}}{1 - t_{orig}}, \tag{50}$$

where $\nu = \frac{1}{p+1}\left[1 - \left[F(n)^{p+1} + (1 - F(n))^{p+1}\right]\right]$.

In Figure X we show the effect of changing p on $\frac{t}{1-t}$ with the same methodology as in Figure IX. Increasing p again leads to larger effects towards the top of the distribution and relatively small changes at the bottom. It is notable that the effects at the bottom remain small despite the increased magnitude of the positive externality on these individuals' income. This is driven by the opposition of the mechanical and behavioral channels discussed in the main text. Both equality effects – the internalization of the externality and the increased want for equality – move in the same direction at the top, but work against each other near the bottom.

Figure X

Effects on $\frac{t}{1-t}$: The S-Gini Family



Note: Effects on $\frac{t}{1-t}$ for various S-Ginis. The social planner is Rawlsian. The magnitude of the inequality externality is held constant for all p at the upper bound of the median value from the empirical inequality aversion estimates in Carlsson et al. (2005). The productivity distribution is (a) the empirical wage distribution from the main text, (b) a log-normal distribution with $\sigma = 0.39$ and $\mu_{log} = -1$, and (c) a Pareto distribution with $a = 2$. See Figure VIII for an explanation of the inequality metrics. In particular, larger p indicates that top and bottom income variation is weighted more than middle-income variation. The elasticity of labor E_L is 0.3.

The majority of the new insight noted in the previous subsection also hold for the S-Gini. Unlike in the top income shares, however, the benefits of taxing near the bottom increase with increasing p. This is a somewhat surprising result. It is due to the mechanical effect being more potent when bottom externalities are very large; in effect, the average inequality metric weight above increases rapidly near the bottom. This leads to the generally large equality benefits from the mechanical effect being even larger than the increased benefits of subsidizing the poor to work more. We caution that this is a particularly model-driven result.

A last caveat; throughout the paper we use a family of *absolute* inequality metrics. This is done to keep scale independence in the additive utility function. However, as this means that the inequality metric can increase without bounds, caution is required when working with large externality values. A further exploration of other functional forms would be beneficial to understand how this changes the optimal tax problem.

F Different inequality externalities

In this section we calculate the optimal non-linear income tax rates in the presence of other types of inequality externalities, namely (i) pre-tax income inequality externalities, and (ii) utility inequality

externalities.

F.I. Pre-tax income inequality externality

A pre-tax income inequality externality problem is a simpler version of the post-tax income inequality externality problem. We solve it here in the small perturbation framework under no income effects. The majority of the solution is similar. In terms of inequality impacts, the mechanical channel falls away while the behavioral channel becomes stronger. The revenue and direct welfare portions are standard.

We introduce a small tax reform $d\tau_z$ where the marginal income tax is increased by $d\tau$ in a small band from z to $z + dz$. The reform mechanically increases average tax rates on everyone above this band. These agents do not change their work decisions or pre-tax income, so the effect of these individuals on *pre-tax* income inequality does not change.

The behavioral response is driven by agents changing their pre-tax income. The inequality impact of the behavioral responses is thus preserved, and in fact increased. Those who are located in the small band between z to $z + dz$ work less, and reduce their pre-tax income by an amount $\partial z = -\epsilon(z) z \partial\tau / (1 - \tau(z))$.[82] The behavioral response thus has an effect on the post-tax income inequality metric as $d\bar{\theta}_B = -\kappa(z) \cdot dz \partial\tau \cdot \epsilon(z) z f(z) / (1 - \tau(z))$. This differs from the post-tax inequality impact by a factor of $1 / (1 - \tau(z))$.

The total equality effect is only driven by these behavioral responses and thus $d\bar{\theta} = d\bar{\theta}_B$. In terms of utility, this affects every individual as $\int_j g_j \frac{\partial U_j}{\partial \bar{\theta}} \cdot \partial\bar{\theta} \cdot dj$. As we assume an homogenous inequality externality and quasi-linearity in consumption such that $\eta = MRS_{x\bar{\theta}} = -\frac{\partial U / \partial \bar{\theta}}{\partial U / \partial x} = -\frac{\partial U}{\partial \bar{\theta}}$, the total welfare effect of the inequality change is $dI = \int_j g_j \cdot (-\eta) \cdot \partial\bar{\theta} \cdot dj = -\eta \cdot \partial\bar{\theta} \cdot \int_j g_j dj$.

The total welfare change, including all channels, is equal to zero at the optimum:

$$dM + dB + dW + dI = 0.$$

Thus, using the expressions for dM, dB, dW and other variables from Appendix D, we have:

$$dz\partial\tau \int_j g_j dj \left[1 - F(z) - f(z)\epsilon(z) z \frac{\tau(z)}{1 - \tau(z)}\right] - dz\partial\tau \bar{G}(z) (1 - F(z)) \int_j g_j dj$$
$$+\eta \cdot \int_j g_j dj \cdot dz\partial\tau \cdot \frac{[\kappa(z)\epsilon(z) z f(z)]}{1 - \tau(z)} = 0$$

Dividing by $z f(z) \epsilon(z) \int_j g_j dj \cdot dz\partial\tau$ and re-arranging, we find:

$$\frac{\tau(z) - \eta \cdot \kappa(z)}{1 - \tau(z)} = \frac{1 - F(z)}{z \cdot f(z)} \frac{\left(1 - \bar{G}(z)\right)}{\epsilon(z)}$$

Which implies, after substituting $\alpha(z) = z f(z) / (1 - F(z))$,

[82] Unlike in the post-tax case, this is already in the relevant metric (pre-tax income) and therefore does not have to be multipled by $1 - \tau(z)$.

$$\tau(z)\left(1+\frac{\left(1-\bar{G}(z)\right)}{\alpha(z)\epsilon(z)}\right)=\frac{\left(1-\bar{G}(z)\right)}{\alpha(z)\epsilon(z)}+\eta\cdot\kappa(z)$$

And finally,

$$\tau(z)=\frac{1+\eta\cdot\kappa(z)\alpha(z)\epsilon(z)-\bar{G}(z)}{1+\alpha(z)\epsilon(z)-\bar{G}(z)}.$$

The effect of the mechanical inequality channel on the final result has fallen away. The behavioral channel is also stronger, as it is only present in the numerator. The result cannot be approximated by social welfare weights, whether in utility or income.

F.II. Utility inequality externality

We solve the utility inequality externality problem here in the small perturbation framework with an additive utility inequality externality such that the inequality metric is,

$$\bar{\theta}_{\mathrm{U}}\left(\boldsymbol{z},F\right)=\int_{\underline{U}}^{\bar{U}}\kappa_U(U(z))U(z)dF(U(z)),\tag{51}$$

where $U(z)$ is total individual utility, z is total individual earnings, and $\kappa_U(U(z))$ is some weight in the inequality metric such that $\int_{\underline{U}}^{\bar{U}}\kappa_U(U)dF(U)=0$. For simplicity we will refer to a utility function of the form:

$$U(x,h,\bar{\theta}_U)=x-v(h)-\eta_U\bar{\theta}_U.\tag{52}$$

The majority of the solution is similar. The revenue and direct welfare effects are standard. We will now focus on the (utility) inequality impacts.

We introduce a small tax reform $d\tau_z$ where the marginal income tax is increased by $d\tau$ in a small band from z' to $z'+dz$. We note that the utility of the agents making behavioral responses only changes on a second-order basis. We can thus focus on the mechanical effect.

For each $dz\partial\tau$ of revenue collected from those above the bracket, utility inequality changes. To explore the mechanical effect it is useful to first simplify the utility inequality term we need for this specific channel. We can first safely ignore the impact of the mechanical effect on the labor term in the utility function, as the mechanical channel is unrelated to any change in labor choice and the utility function is additive. Further, as $\int_{\underline{U}}^{\bar{U}}\kappa_U(U)dF(U)=0$ by assumption and any change in the inequality metric is flatly applied to everyone by the homogeneous externality assumption, we can also ignore the impact the mechanical effect has on utility through the externality term itself. We are using a quasi-linear utility function, and thus the remaining relevant part of utility is simply $x(z)$. Finally, we note that $\kappa_U(U)=\kappa(z)$ as ranks in post-tax income and utility are identical by assumption. We thus use a simplified inequality metric $\bar{\theta}_{U,mech}$ for the mechanical effect calculation,

$$\bar{\theta}_{\mathrm{U},mech}\left(\boldsymbol{z},F\right)=\int_{\underline{z}}^{\bar{z}}\kappa(z)x(z)dF(z),\tag{53}$$

which is identical to the post-tax absolute income inequality metrics used in the main text.

With this simplification the derivation of the remainder of the problem becomes nearly identical to that in Appendix D. To summarize, the behavioral response channel does not exist in the utility inequality case and the mechanical effect channel simplifies to that of a post-tax income inequality externality. Following the solution in Appendix D to its conclusion (excluding the behavioral response channel) we find:

$$\tau(z) = \frac{1 + \eta_U \cdot \bar{\kappa}(z) - \bar{G}(z)}{1 + \alpha(z)\epsilon(z) + \eta_U \cdot \bar{\kappa}(z) - \bar{G}(z)}.$$

Which is identical to the standard case after removing the behavioral response terms. Note that by using the modified social welfare weights $\bar{G}'(z) = \bar{G}(z) - \eta_U \cdot \bar{\kappa}(z)$ this can be simplified to the no-externality case without the need for $\alpha(z)$ or $\epsilon(z)$ in the modified social welfare weights.

1 Removing quasi-linearity Without quasi-linearity in the utility function such that $U(x, h, \bar{\theta}) = u(x) - v(h) - \eta\bar{\theta}$, the relevant inequality metric is:

$$\bar{\theta}'_{\mathrm{U},mech}(\boldsymbol{z}, F) = \int_{\underline{z}}^{\bar{z}} \kappa(z) u(x(z)) dF(z). \tag{54}$$

Here there are two significant effects on this absolute inequality metric from the mechanical effect. The first is the reduction of post-tax income (and thus utility) of everyone above the tax bracket. The second is the flat increase in post-tax income from the redistributed revenue.

We begin with the first of these. Each decrease in one unit of post-tax income changes absolute utility inequality by $-\kappa(z)u_x(x(z))f(z)$ (from Equation 53). The total decrease is thus $\int_{z'}^{\bar{z}} -u_x(x(z))\kappa(z)dz\partial\tau dF(z)$. This is as far as we can go in the general case as the sum of $u_x(x(z_j))\kappa(z_j)$ above z' is not easily simplified.

The flat increase in post-tax income changes utility inequality in a similar fashion, where if total revenue gathered per agent is dR', the total effect becomes $\int_{\underline{z}}^{\bar{z}} u_x(x(z))\kappa(z)dR'dF(z)$. This is again difficult to simplify.

When assuming a quasi-linear utility function the problem simplifies, as the reduction in post-tax income above z' leads to an inequality change of $\int_{z'}^{\bar{z}} -u_x(x(z))\kappa(z)dz\partial\tau dF(z) = -\overline{\kappa}(z)\left[1 - F(z)\right]dz\partial\tau$, and the flat increase in income has no effect as $\int_{\underline{z}}^{\bar{z}} u_x(x(z))\kappa(z)dR'dF(z) = dR'\int_{\underline{z}}^{\bar{z}} \kappa(z)dF(z) = 0$. We can thus write that the total utility inequality change from the perturbation is $d\bar{\theta}_U = -\overline{\kappa}(z)\left[1 - F(z)\right]dz\partial\tau$, which is equal to the mechanical effect from the standard externality case.

G Further externality micro-foundations

Below we show micro-foundations for three more inequality externality channels; trust, crime, and political capture.

- Trust: Assume that individuals have higher trust $t_{i,j}$ in other individuals who share a set of similar characteristics, where the set of relevant characteristics is denoted as the vector $\overrightarrow{T}$.

If income x is part of $\overrightarrow{T}$, or causes changes in individual parameters that are, a change in income inequality $\bar{\theta}$ would decrease individual i's general trust levels $T_i = \sum_i t_{i,j}$. If T_i enters into individual utility $U(x_i, T_i, ...)$, income inequality has an indirect utility effect.

- Crime: Assume that criminal activity gains a fraction α of another agent's income x_j, subtracting a fixed risk cost, where agent j is randomly chosen from some high-income subset. Further assume that the opportunity cost of crime is a wage-paying job with a salary proportional to the agent's income x_i, and that agents will commit crime if it is profitable. We define the Gini coefficient as $\overline{\theta}_G = \sum_i \sum_j (x_i - x_j)$. If $\overline{\theta}_G$ increases, the relative benefit of crime also generally increases, and criminal activity increases with subsequent society-wide utility effects from both victims and perpetrators. As richer individuals are able to spend more income to protect their assets, this effect might be moderated or even overturned.[83]

- Political capture: Assume that the political process is affected by a voting procedure between discrete options $\left\{\overline{V}_1, ..., \overline{V}_m\right\}$ where each agent has a number of votes $v_i(x_i)$ corresponding to an increasing function of their income x_i. Assume further that individual utility $U_i(x_i, \overline{V}_k, ...)$ is dependent on the outcome of this political process, with varying individual preferences. Changing income inequality $\bar{\theta}$ will mechanically change voting outcomes by giving higher-income agents a larger vote share. As the vote outcome affects the individual utility of every agent – positively or negatively – inequality indirectly affects individual utility.

[83] As with all these examples, this is a very simple illustration of a complex topic with several other potential causal strains. See Kelly (2000) for a broader discussion.

References

Aaberge, R. (2000). Characterizations of Lorenz Curves and Income Distributions. *Social Choice and Welfare 17*(4), 639–653.

Aaberge, R. and U. Colombino (2013). Using a Microeconometric Model of Household Labour Supply to Design Optimal Income Taxes. *The Scandinavian Journal of Economics 115*(2), 449–475.

Alesina, A. and P. Giuliano (2011). Preferences for Redistribution. In *Handbook of Social Economics*, Volume 1, pp. 93–131. Elsevier.

Anbarci, N., M. Escaleras, and C. A. Register (2009). Traffic Fatalities: Does Income Inequality Create an Externality? *Canadian Journal of Economics/Revue canadienne d'économique 42*(1), 244–266.

Aronsson, T. and O. Johansson-Stenman (2008). When the Joneses' Consumption Hurts: Optimal Public Good Provision and Nonlinear Income Taxation. *Journal of Public Economics 92*(5-6), 986–997.

Aronsson, T. and O. Johansson-Stenman (2015). Keeping up with the Joneses, the Smiths and the Tanakas: on International Tax Coordination and Social Comparisons. *Journal of Public Economics 131*, 71–86.

Aronsson, T. and O. Johansson-Stenman (2018). Inequality Aversion and Marginal Income Taxation. *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association 111*, 1–32.

Aronsson, T. and O. Johansson-Stenman (2020). Optimal Second-Best Taxation When Individuals Have Social Preferences. *Umeå Economic Studies* (973).

Ashworth, J., B. Heyndels, and C. Smolders (2002). Redistribution as a Local Public Good: An Empirical Test for Flemish Municipalities. *Kyklos 55*(1), 27–56.

Atkinson, A. B. (1970). On the Measurement of Inequality. *Journal of Economic Theory 2*(3), 244–263.

Atkinson, A. B. and J. E. Stiglitz (1976). The Design of Tax Structure: Direct versus Indirect Taxation. *Journal of Public Economics 6*(1-2), 55–75.

Becker, G. S. (1968). Crime and Punishment: An Economic Approach. In *The economic dimensions of crime*, pp. 13–68. Springer.

Bergh, A., T. Nilsson, and D. Waldenström (2016). *Sick of Inequality?: An Introduction to the Relationship Between Inequality and Health*. Edward Elgar Publishing.

Bergolo, M., G. Burdin, S. Burone, M. De Rosa, M. Giaccobasso, and M. Leites (2021). Dissecting Inequality-Averse Preferences. Technical report, IZA Discussion Papers.

Blundell, R. and A. Shephard (2011). Employment, Hours of Work and the Optimal Taxation of Low-Income Families. *The Review of Economic Studies 79*(2), 481–510.

Boskin, M. J. and E. Sheshinski (1978). Optimal Redistributive Taxation when Individual Welfare Depends upon Relative Income. *The Quarterly Journal of Economics*, 589–601.

Bovenberg, A. L. and F. van der Ploeg (1994). Environmental Policy, Public Finance and the Labour Market in a Second-Best World. *Journal of Public Economics 55*(3), 349–390.

Carlsson, F., D. Daruvala, and O. Johansson-Stenman (2005). Are People Inequality-Averse, or Just Risk-Averse? *Economica 72*(287), 375–396.

Cingano, F. (2014). Trends in Income Inequality and its Impact on Economic Growth. *OECD Social, Employment and Migration Working Papers* (163).

Cohen, M. A., R. T. Rust, S. Steen, and S. T. Tidd (2004). Willingness-To-Pay for Crime Control Programs. *Criminology 42*(1), 89–110.

Cooper, D. J. and J. Kagel (2016). Other-Regarding Preferences. *The Handbook of Experimental Economics 2*, 217.

Cowell, F. A. (2000). Measurement of Inequality. *Handbook of Income Distribution 1*, 87–166.

Cremer, H., F. Gahvari, and N. Ladoux (1998). Externalities and Optimal Taxation. *Journal of Public Economics 70*(3), 343–364.

Diamond, P. A. (1998). Optimal Income Taxation: An Example With a U-shaped Pattern of Optimal Marginal Tax Rates. *American Economic Review*, 83–95.

Diamond, P. A. and J. A. Mirrlees (1971). Optimal Taxation and Public Production II: Tax Rules. *The American Economic Review 61*(3), 261–278.

Dimick, M., D. Rueda, and D. Stegmueller (2018). Models of Other-Regarding Preferences, Inequality, and Redistribution. *Annual Review of Political Science 21*, 441–460.

Donaldson, D. and J. A. Weymark (1980). A Single-Parameter Generalization of the Gini Indices of Inequality. *Journal of Economic Theory 22*(1), 67–86.

Dufwenberg, M., P. Heidhues, G. Kirchsteiger, F. Riedel, and J. Sobel (2011). Other-Regarding Preferences in General Equilibrium. *The Review of Economic Studies 78*(2), 613–639.

Fehr, E. and K. M. Schmidt (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics 114*(3), 817–868.

Flood, S., M. King, R. Rodgers, S. Ruggles, and J. R. Warren (2018). Integrated Public Use Microdata Series, Current Population Survey: Version 6.0 [dataset]. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D030.V6.0.

Goodin, R. E. (1986). Laundering Preferences. *Foundations of Social Choice Theory 75*, 81–86.

Guvenen, F., G. Kambourov, B. Kuruscu, S. Ocampo-Diaz, and D. Chen (2019). Use It or Lose It: Efficiency Gains from Wealth Taxation. Working Paper 26284, National Bureau of Economic Research.

Harsanyi, J. C. (1977). Morality and the Theory of Rational Behavior. *Social Research 44*(4), 623–656.

Heathcote, J., K. Storesletten, and G. L. Violante (2020). Presidential address 2019: How should tax progressivity respond to rising income inequality? *Journal of the European Economic Association 18*(6), 2715–2754.

Jacquet, L. and E. Lehmann (2021). How to Tax Different Incomes? *CEPR Discussion Paper Series no. 16571, IZA Institute of Labor Economics*.

Kanbur, R., M. Keen, and M. Tuomala (1994). Optimal Non-Linear Income Taxation for the Alleviation of Income-Poverty. *European Economic Review 38*(8), 1613–1632.

Kanbur, R. and M. Tuomala (2013). Relativity, Inequality, and Optimal Nonlinear Income Taxation. *International Economic Review 54*(4), 1199–1217.

Kaplow, L. (2010). *The Theory of Taxation and Public Economics*. Princeton University Press.

Kelly, M. (2000). Inequality and Crime. *Review of Economics and Statistics 82*(4), 530–539.

Laffer, A. B. (2004). The Laffer Curve: Past, Present, and Future. *Backgrounder 1765*, 1–16.

Lindbeck, A. (1985). Redistribution Policy and the Expansion of the Public Sector. *Journal of Public Economics 28*(3), 309–328.

Lobeck, M. and M. Støstad (2022). Inequality Externality Beliefs and Redistributive Preferences. *Working paper (Forthcoming)*.

Lockwood, B. B., C. G. Nathanson, and E. G. Weyl (2017). Taxation and the Allocation of Talent. *Journal of Political Economy 125*(5), 1635–1682.

Lollivier, S. and J.-C. Rochet (1983). Bunching and Second-Order Conditions: A Note on Optimal Tax Theory. *Journal of Economic Theory 31*(2), 392–400.

Mankiw, N. G., M. Weinzierl, and D. Yagan (2009). Optimal Taxation in Theory and Practice. *Journal of Economic Perspectives 23*(4), 147–74.

Manning, A. (2015). Top Rate of Income Tax. *Centre for Economic Performance's Election Analysis*.

Mirrlees, J. A. (1971). An Exploration in the Theory of Optimum Income Taxation. *The Review of Economic Studies 38*(2), 175–208.

Mirrlees, J. A. (1976). Optimal Tax Theory: A Synthesis. *Journal of Public Economics 6*(4), 327–358.

Oswald, A. J. (1983). Altruism, Jealousy and the Theory of Optimal Non-Linear Taxation. *Journal of Public Economics 20*(1), 77–87.

Pauly, M. V. (1973). Income Redistribution as a Local Public Good. *Journal of Public Economics 2*(1), 35–58.

Persson, M. (1995). Why are Taxes so High in Egalitarian Societies? *The Scandinavian Journal of Economics*, 569–580.

Piketty, T. and E. Saez (2007). How Progressive is the US Federal Tax System? A Historical and International Perspective. *Journal of Economic Perspectives 21*(1), 3–24.

Piketty, T. and E. Saez (2013). Optimal Labor Income Taxation. In *Handbook of Public Economics*, Volume 5, pp. 391–474. Elsevier.

Piketty, T., E. Saez, and S. Stantcheva (2014). Optimal Taxation of Top Labor Incomes: A Tale of Three Elasticities. *American Economic Journal: Economic Policy 6*(1), 230–71.

Prete, V., A. Sommacal, and C. Zoli (2016). Optimal Non-Welfarist Income Taxation for Inequality and Polarization Reduction. Technical report.

Rothschild, C. and F. Scheuer (2016). Optimal Taxation with Rent-Seeking. *The Review of Economic Studies 83*(3), 1225–1262.

Rueda, D. and D. Stegmueller (2016). The Externalities of Inequality: Fear of Crime and Preferences for Redistribution in Western Europe. *American Journal of Political Science 60*(2), 472–489.

Rufrancos, H., M. Power, K. E. Pickett, and R. Wilkinson (2013). Income Inequality and Crime: A Review and Explanation of the Time Series Evidence. *Sociology and Criminology-Open Access*.

Sadka, E. (1976). On Income Distribution, Incentive Effects and Optimal Income Taxation. *The Review of Economic Studies 43*(2), 261–267.

Saez, E. (2001). Using Elasticities to Derive Optimal Income Tax Rates. *The Review of Economic Studies 68*(1), 205–229.

Saez, E. and S. Stantcheva (2016). Generalized Social Marginal Welfare Weights for Optimal Tax Theory. *American Economic Review 106*(1), 24–45.

Sandmo, A. (1975). Optimal Taxation in the Presence of Externalities. *The Swedish Journal of Economics*, 86–98.

Schmidt, U. and P. C. Wichardt (2019). Inequity Aversion, Welfare Measurement and the Gini Index. *Social Choice and Welfare 52*(3), 585–588.

Seade, J. K. (1977). On the Shape of Optimal Tax Schedules. *Journal of Public Economics 7*(2), 203–235.

Sen, A. (1976). Real National Income. *The Review of Economic Studies 43*(1), 19–39.

Simula, L. and A. Trannoy (2022a). Bunching in Rank-Dependent Optimal Income Tax Schedules. *Social Choice and Welfare*, 1–27.

Simula, L. and A. Trannoy (2022b). Gini and optimal income taxation by rank. *American Economic Journal: Economic Policy Forthcoming*.

Stiglitz, J. E. (1982). Self-selection and Pareto Efficient Taxation. *Journal of Public Economics 17*(2), 213–240.

Thurow, L. C. (1971). The Income Distribution as a Pure Public Good. *The Quarterly Journal of Economics*, 327–336.

Weitzman, M. L. (2009, 02). On Modeling and Interpreting the Economics of Catastrophic Climate Change. *The Review of Economics and Statistics 91*(1), 1–19.