Inequality as an Externality: Consequences for Tax Design*

Morten Nyborg Støstad† and Frank Cowell‡

Abstract

Economic inequality may change a wide range of societal outcomes, for example crime rates, economic growth, and political polarization. In this paper we discuss how to model such effects in welfarist frameworks. Our main suggestion is to treat economic inequality itself as an externality, which has wide-ranging implications for classical theory. We show this through the classical optimal non-linear income taxation model, where we focus on a post-tax income inequality externality. Top tax rates are particularly affected by the externality; in our main specification the optimal top marginal tax rate increases from 63% to 81%. Our model also provides a theoretical basis for real-world governmental tax choices that are irrational under standard optimal taxation methods. Finally, we find that the total inequality aversion implied by the current U.S. tax system is insufficient to accommodate both progressive social welfare weights and a significant concern for inequality's externality effects. *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, Eddy Zanoutene, 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. Finally, we are deeply grateful to the Journal of Public Economics' Editor Nathaniel Hendren and two anonymous referees for invaluable comments and suggestions. Version: May 8, 2023.

†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.

1 Introduction

In the last sixty years, economic modeling has regularly used individualistic utility functions and social welfare functions to evaluate policy options. The real-world influence of these models has been considerable; as such, how they treat economic inequality is also of great significance. There are several well-formulated reasons to prevent economic differences in the standard framework, which we will return to shortly, but a crucial factor has also remained neglected; the consequences of economic inequality on society and thus individuals' well-being. Suppose, for example, that higher economic inequality causally changes the crime rate, the amount of social unrest, or the political polarization in a society. If so, even purely self-interested individuals are affected by the economic differences between people – regardless of whether their individual incomes change. Given that virtually all market activities affect the extent of economic inequality, it follows that economic inequality itself could be an externality. This paper explores the consequences of this idea.

The analysis we present can be divided into two primary components. The first, which is also the overarching theme of the paper, is the concept of economic inequality as an externality. This concept is explored in a general fashion. We discuss why an economic inequality term in the utility function is the most appropriate way to model the effects of inequality on society, following both Thurow (1971) and Alesina and Giuliano (2011). As such a term cannot be mathematically approximated by appropriate social welfare functions (SWF) or concave utility functions, most models which preclude externalities also assume that economic inequality does not significantly change society. As this is a potentially large assumption, we discuss how weakening it and allowing for various inequality externalities affects both general economic intuition and optimal taxation frameworks. We also establish micro-foundations for the externality, which are often simple and can exist in the presence of fully self-interested, rational individuals.[1]

The second component of the paper explores a specific case where the no-externality assumption has been influential, namely the Mirrlees (1971) optimal non-linear income taxation model. We calculate optimal marginal tax rates analytically and numerically in the presence of various types of inequality externalities. While we focus on a post-tax income inequality externality, we also introduce other types of inequality externalities into the model (pre-tax income, utility) and vary the inequality metric itself. To pin down plausible magnitudes of a real-world income inequality externality we utilize three distinct methods, all of which imply similar magnitude ranges. Finally, we perform an inverse-optimum exercise to examine how implied transfer progressivity changes in the U.S. tax system if the tax schedule design was influenced by an income inequality externality.

The principal insight of our paper is that the large majority of welfare-based economic frameworks implicitly assume that economic inequality has no meaningful effects on societal outcomes, and that softening this assumption changes model conclusions drastically. We explicitly show these changes in optimal income taxation (OIT), where both theoretical and simulation-based findings are affected, and

[1] That self-interested individuals are affected by the externality is the main difference between our concept and other-regarding preferences. Such preferences are philosophically problematic for policy design as they are based on individuals' emotions (Harsanyi, 1977; Goodin, 1986).

discuss which other frameworks could be similarly fragile. Within the context of OIT we find two main results. First, the presence of an inequality externality has a particularly pronounced impact on *top* optimal marginal tax rates. This is a theoretical finding that is borne through in our numerical simulations; in our main specification the optimal top marginal tax rate changes from 63% to 81% when introducing a median post-tax income inequality externality. Second, our analysis reveals that the total inequality aversion in the current U.S. tax system is insufficient to accommodate both progressive social welfare weights and a significant concern for inequality's externality effects. While the current tax system could be rationalized as prioritizing income transfers to lower-income individuals (Hendren, 2020), it cannot also contain a realistic concern for inequality's externality effects given the aggregate capacity of the tax schedule to mitigate inequality.

Before further discussing our results we will briefly explore what we know about how economic inequality affects various facets of society and individuals' lives. It is difficult to establish causality on the topic for several reasons, the first among them being the lack of exogenous variation in macroeconomic inequality.[2] However, there is no shortage of empirical papers on the subject, and there are overall strong indications that economic inequality acts as an externality in various ways. First, considerable experimental and microeconomic evidence has in recent years indicated that economic inequality between workers or experimental subjects impacts life satisfaction (Card et al., 2012), productivity (Breza et al., 2018), trust (Fehr et al., 2020), and cooperation (Xu and Marandola, 2022). Second, as popularized in Wilkinson and Pickett (2009), there are robust cross-country correlations between economic inequality and various negative societal outcomes.[3] We show two such correlations for general trust and homicides in Figure I. Third, both laypeople and experts often express the belief that inequality does change society; in the United States, the large majority of citizens believe that economic inequality negatively affects a wide range of societal outcomes (Lobeck and Støstad, 2023). Similar concerns have been raised by prominent politicians, philosophers, and economists.[4] Laboratory experiments have also shown that a majority of individuals would forego part of their income to live in more macroeconomically equal societies (Carlsson et al., 2005; Bergolo et al., 2022). Fourth, it is trivial to create realistic microfoundations of various inequality externalities, which we show in Section 5. Other papers have given more attention to specific potential channels; Benabou (1996), Auclert and Rognlie (2018), and Mian et al. (2020) are just a few examples.[5]

So let us assume such effects exist and are welfare-relevant. How would one consider their overall

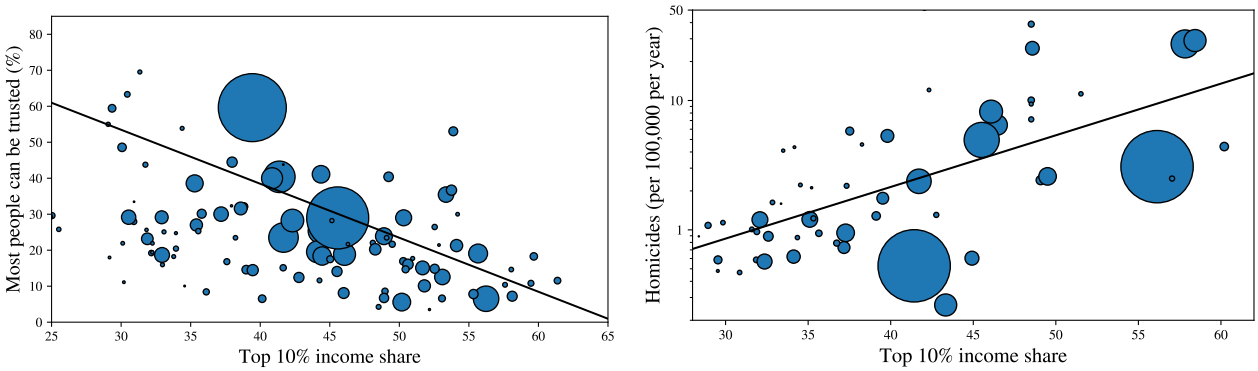
[2]Other concerns include measurement error and missing data on economic inequality, a generally low variability in inequality over time, reverse causality (where outcomes also affect inequality), non-linear effects of inequality on outcomes, the intertwined nature of inequality effects and poverty or individual income effects, the question of *which type* of inequality matters, and so on.

[3]The related literatures are too large to summarize here. 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.

[4]For example Plato (2016): "In a state which is desirous of being saved from the greatest of all plagues [...] here should exist among the citizens neither extreme poverty, nor, again, excess of wealth, for both are productive of both these evils," Greenspan (2014): "You can see the deteriorating impact of [inequality] on our current political system," or Obama (2011): "This kind of inequality – a level that we haven't seen since the Great Depression – hurts us all."

[5]While most of the evidence presented in this paragraph indicates that economic inequality is a negative externality, we do not assume this in general. Jones (2022) discusses how top incomes could drive innovation, for instance. One could imagine such concerns being more prevalent in societies much more economically equal than we see today.

Figure I: The Correlations of Inequality



Note: Left: The cross-country correlation of generalized trust (World Values Survey) and the top 10% income share (World Inequality Database). Right: The cross-country correlation of homicides (World Bank) and the top 10% income share (World Inequality Database). Both correlations are relatively unaffected by standard controls. Data point area is proportional to population.

consequences in a welfarist framework? The most intuitive approach is simply to model every externality individually. It is difficult to imagine that such a model could remain tractable, however, and this strategy would also require an empirically problematic assessment of the actual importance of every externality channel. For general use, then, this route appears overall infeasible. Meanwhile, a standard inequality-averse or Rawlsian social planner is generally insufficient to model inequality externalities based on income or wealth because the externality introduces a wedge between what is individually and societally optimal.[6]

Instead, a more appropriate solution is the inclusion of an inequality metric in the individual's utility function. Such an externality can exist even when individuals are fully rational and selfish, unlike models with traditional other-regarding preferences (ORP). As an example, imagine a perfectly self-interested individual in a society where income inequality increases crime through a Becker (1968)-type opportunity cost framework. 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 variable affected by inequality.

We are not the first to note that economic inequality's effects on society can imply an inequality term in the utility function. The idea was first developed by Thurow (1971), who shows that the First Welfare Theorem fails if the income distribution is a pure public good. Since then the idea has periodically resurfaced. Kaplow (2010), for example, mentions that the economic distribution could affect variables such as crime, which could imply optimal taxation effects. Alesina and Giuliano (2011) briefly consider how inequality's effects on society might affect consumption and thus utility, and Rueda

[6]Non-altruistic individuals will choose their own labor effort without taking into account how this effort impacts the global level of income inequality, for example; if income inequality affects societal factors, there is an externality dimension to this choice that is not well-modeled by simply discounting individual utility.

and Stegmueller (2016) discuss how inequality can act as an externality in the case of crime. We add to this literature by detailing the effect of an inequality externality in a specific, well-known economic model – the Mirrlees model – and by furthering the analysis presented in these works. To advance our understanding of the overarching idea, we (i) clarify the mathematical structure of the externality, classify its key components, and develop a sufficient statistic for the magnitude of the externality given an inequality metric, (ii) make a first approach at estimating the magnitude of a post-tax income inequality externality based on the Gini coefficient, (iii) formulate a set of theoretical consequences of income- and wealth-based inequality externalities that are relevant for the broader economic literature, and (iv) create micro-foundations for various ways in which economic inequality can change pertinent societal factors.

Our case study uses the Mirrlees (1971) model, where we introduce various types of inequality terms into the individual's utility function. As a widely used model describing OIT, the Mirrlees model represents an important pillar of public economics (Diamond and Mirrlees, 1971; Diamond, 1998; Saez, 2001). The original Mirrlees model assumes no externalities, an assumption which has been examined by a wide range of papers. Though we will return to how our work contrasts with the existing literature, the briefest way to describe our technical contribution is that we are the first paper to explore the effect of an income inequality term in the individual utility function in the continuous Mirrlees model. We explore several types of inequality externalities in this framework, focusing mainly on a post-tax income inequality externality, and solve the problem both analytically and numerically.

Our OIT exercise yields two main results. First, top marginal tax rates are particularly sensitive to the inequality externality. This is because the *equality effects* we introduce are more heavily influenced by the distributional location of any tax increase than the standard *revenue effects*. In the standard revenue case, both consequences of a small tax increase – the behavioral responses and mechanical effect described in Saez (2001) – always oppose each other. A tax increase leads to mechanically higher tax revenue at the same time as agents' behavioral shifts away from labor supply decreases tax revenue. In contrast, the two *equality* effects can also harmonize. The mechanical effect is simple, as it is similar to the revenue case and always decreases economic 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 labor increases income inequality. At the top, a behavioral shift away from work effort *reduces* income inequality.

This creates a distributional asymmetry, where the two new equality effects always oppose each other at the bottom of the distribution and always harmonize at the top. Top marginal tax rates are thus 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 taxpayer is crucial when evaluating equality effects, and the social planner's incentive to change incomes is larger towards the ends of the distribution. 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. Applying this median externality value results in the optimal top marginal tax rate increasing from 63% to 81%. 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, particularly when moving away from the Gini coefficient to inequality metrics that weight middle-class differences less heavily.[7]

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.[8] 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.[9] The *individual* concerns that arise from an inequality externality thus differ from the *social* concerns modeled by an inequality-averse SWF. We 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 also 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).

Our second main result is related to this last point and comes from the inverse-optimum exercise popularized by Bourguignon and Spadaro (2012). This method calculates the implied SWWs of real-world tax systems under the assumption that the tax schedule was set optimally. As shown in Lockwood and Weinzierl (2016) and Hendren (2020), SWWs from the U.S. tax schedule are generally decreasing in income. We introduce an inequality externality into this framework, which allows us to understand the ramifications on the implied progressivity of the tax system if the social planner considered inequality as an externality when designing the tax schedule.

Through this exercise we find that the 2019 U.S. tax system is not sufficiently progressive to accomodate both a socially progressive transfer motive and a realistic concern about the societal effects of economic inequality. The intuition is that any tax schedule contains a certain amount of "total" inequality aversion that may explain either progressive SWWs, as in Lockwood and Weinzierl (2016) and Hendren (2020), or a non-negligible concern for inequality's externality effects. The total current inequality aversion in the U.S. tax schedule, however, is too small to explain both. If the U.S. social planner considered inequality as an externality to our median value, implied SWWs are sharply in-

[7]The Gini coefficient has often been criticized for over-weighting middle income inequalities.

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

[9]This is not to say that the inequality externality is an equity concern in the standard equity and efficiency-framework, where it is clearly an efficiency concern.

creasing in income – indicating that one dollar at the bottom of the distribution is worth five dollars at the top. We conclude that the current U.S. tax schedule is either not progressive in transfers or has an implied concern for inequality's externality effects that is significantly lower than our empirical estimates.

We also present a handful of smaller takeaways. The optimal tax schedule is unambiguously more progressive (regressive) under a negative (positive) post-tax income inequality externality, where more progressive is defined as a lower level of post-tax income inequality. The classical U-shape found in the optimal income taxation literature is fragile to the inclusion of a negative (but not positive) post-tax income inequality externality. A moderately sized pre-tax income inequality externality makes the optimal Utilitarian marginal tax rates closer to real-world tax systems, where marginal rates largely increase in income. And more generally, numerical and theoretical results from the OIT model depend strongly on the type and magnitude of the inequality externality. This brings us back to the overarching comment of the paper; the absence of externalities in economic frameworks has lead to a situation where most models implicitly assume that economic inequality itself does not change society, or does so in very limited ways.

We will now briefly outline how our work differs from the existing OIT literature. We have three main technical contributions. First, we introduce a new and simple way to account for inequality terms in individual utility functions in optimal taxation frameworks. This is possible due to the family of inequality metrics we use, which simplifies an analytically intractable externality[10] into a linear combination of consumption externalities with varying marginal effects that depend on the income-rank of the individual.[11] As such, we can use much of the existing externality framework to evaluate what would otherwise be a challenging analytical problem. As our approach assumes separability between the externality and the remainder of the utility function for simplicity,[12] our framework is an extension of the models presented in Oswald (1983) and particularly Kanbur and Tuomala (2013), both of which examine an average income-based externality. The second contribution to the literature is thus to explore the ramifications of an extension from Kanbur and Tuomala (2013) where we allow the marginal externality to depend on the location of the individual in the distribution (and thus also the individual's income). Although both Oswald (1983) and Kanbur and Tuomala (2013) mention this as a possibility, neither paper explicitly explores the issue. We focus on a small-perturbation framework to build intutition, unlike Kanbur and Tuomala (2013) which uses a mechanism design framework (the modified version of which we also solve). Our analysis leads to novel insights relating to the effect of distributional externalities on optimal income taxation, and particularly on optimal *top* tax rates. Third, we solve the inverse-optimum problem (Bourguignon and Spadaro, 2012) in the presence of a global externality and illustrate the consequences for implied SWWs of the 2019 U.S. tax system. Global externalities are

[10]Typical inequality metrics often use absolute values and multiple integrals that depend on endogenous model variables.

[11]This is the same family used in Simula and Trannoy (2022), developed concurrently with this paper. The family itself is general and allows for various types of income inequality metrics.

[12]This is a large assumption due to how it constrains how the externality magnitude relates to the marginal utilities of income and labor. The assumption of separability in externalities is weakened by among others Pirttilä and Tuomala (1997) and Jacobs and De Mooij (2015). We largely also assume separability between utility from income and labor, again for simplicity. For more on the separability assumption see Gauthier and Laroque (2009).

rarely discussed in this literature – we are only aware of Tsyvinski and Werquin (2017), which discusses the compensation principle in a general equilibrium-based framework and is thus both conceptually and mathematically different from our work. Given the large focus on inequality's effects on society in political rhetoric, we believe this is a particularly interesting exercise in our framework.

In general, our work adds to the already large literature on externalities in optimal taxation. This literature has been particularly developed for environmental externalities (e.g. Sandmo, 1975; Bovenberg and van der Ploeg, 1994; Cremer et al., 1998) and relative income concerns/ORP (Boskin and Sheshinski, 1978; Persson, 1995; Aronsson and Johansson-Stenman, 2008, 2015, 2018, 2020). Our analysis is particularly related to Aronsson and Johansson-Stenman (2020), which discusses various types of ORP 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. 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, which contrasts to our continuous distributional metric inside the individual's utility function.

The paper is organized as follows. Section 2 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 3 incorporates an inequality externality in a standard OIT model and investigates the impact of the externality on optimal tax rates. Section 4 conducts numerical simulations in the OIT model. Section 5 discusses the inequality externality concept further, creating micro-foundations and discussing other potential mathemetical formulations. Section 6 concludes.

2 Inequality and Social Welfare: An Externality Approach

Suppose that economic inequality causally affects non-consumption goods individuals care about, the relevant of which we capture in an vector $\overrightarrow{\Psi_i}$. 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 5.1 for a further discussion on various channels. Suppose further that economic inequality can affect individual consumption x_i (Alesina and Giuliano, 2011),[13] and that individuals may have other-regarding preferences over economic inequality $\bar{\theta}$ (Cooper and Kagel, 2016).[14] The individual's utility can thus be written as,

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

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,

[13] 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).

[14] The overbar indicates a society-wide variable.

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 consumption which is not determined by economic inequality, and the term $\bar{\theta}$ represents the total impact of the inequality externality on the individual.[15]

The simplification 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 soon as one of the three components we show in (1) enters the utility function (and 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 and thus emotions into the welfarist framework – as discussed by Harsanyi (1977) and Goodin (1986), among others – 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. This immediately hints at the potential practical significance of the inequality externality, as ORP modifications often have large impacts on standard model conclusions (e.g. Oswald, 1983; Kanbur and Tuomala, 2013).

The concept also needs a well-defined inequality metric $\bar{\theta}$. We return to this later in the paper, but we note that the main type of inequality we will focus on is *income* inequality. However, it could also be intriguing to consider $\bar{\theta}$ as *wealth* inequality (or some combination of the two). This would be a particularly insightful approach in optimal wealth taxation models, where a key practical motivation for additional taxation is arguably a concern for the societal effects of high wealth inequality. For simplicity we 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.

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. We will return to this in Section 3 and 4.

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

[15] 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. We discuss this further in Appendix A.

of social welfare comparisons. These are (i) the cumulative effect of diminishing marginal utilities of income (DMUI), and (ii) social welfare weights (SWWs). We summarize this framework in Table I.

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 SWW is denoted by g_i.

We posit that the inequality externality is mathematically and intuitively distinct from these other two channels; except for special cases,[16] an inequality externality cannot be mathematically captured by the other formulations.[17] We show a simple proof of why using SWW and DMUI cannot account for the inequality externality in Appendix B.I. The intuition is straightforward: as with any other externality, an inequality externality introduces a gap between the socially and individually optimal decisions. The sub-optimality of individual decisions cannot be approximated by suitable SWWs, as discounting *utility* is dissimilar from discounting *income*, and also cannot be approximated by modifications to an individualistic utility function 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 also discuss income-based SWWs in Section 5.3.

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.

3 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 $U(x_i, l_i, \bar{\theta})$. The main discussion will be for a post-tax income (consumption) inequality externality, with extensions for pre-tax income inequality and utility inequality in Section 4.6.

We solve the main optimal tax problem in two ways. In the main text we will discuss the small-perturbations framework (Saez, 2001), which is arguably the more intuitive approach. We assume no income effects for simplicity. In Appendix C we solve the equivalent problem in the more general

[16] We discuss this further in Section 5.3.

[17] This differs from how DMUI *can* be approximated by appropriate SWWs as long as the remaining (individualistic) utility function is appropriately modified to keep the individual's work choice unaffected.

mechanism design framework (Diamond, 1998), which nests the small-perturbation approach under fewer assumptions.

In the small-perturbation solution we will not specify a full functional form of individual utility, following Saez (2001) (we do this in the more general form in Appendix C). Instead 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. We assume that η_i is an individual-level constant. Our approach thus implicitly assumes separability in inequality and income such that individuals' work decisions are independent on the level of income inequality.[18] Due to this, the policy-determining variable is a welfare-weighted average of the individual-level externality, which we define as $\eta = \int_i g_i \eta_i di / \int_i g_i di$ where g_i is the income-based SWW of individual i as in Saez and Stantcheva (2016).[19]

We also need to choose the inequality metric $\bar{\theta}$. Such metrics are often analytically difficult. To simplify the problem we use a particular family of absolute inequality metrics discussed in Cowell (2000). For post-tax income inequality, which will be used in the main specification, this family has the form,

$$\bar{\theta}\left(\boldsymbol{z}, F\right) = \int_{\underline{x}}^{\bar{x}} \kappa(z) x(z) dH(z), \tag{5}$$

where $H(z)$ is the cumulative distribution function of post-tax income z, and $\kappa(z)$ is the weight of the agent in the inequality metric. This weight is, crucially, only dependent on the *rank* of the individual in the distribution. We have used the rank-invariance between pre-tax income z and post-tax income x to specify the weight in terms of z, which simplifies the problem.[20]

The inequality weight $\kappa(z)$ is positive near the top of the income distribution and negative near the bottom, but is otherwise general. For example, the (absolute) Gini coefficient in post-tax income has a weight $\kappa_G(z) = 2H(z) - 1$. 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 $\int_0^\infty \kappa(z) dH(z) = 0$, such as those in the Lorenz (Aaberge, 2000) or S-Gini families (Donaldson and Weymark, 1980). Absolute inequality metrics are used to keep scale invariance.

It is worth mentioning that the true inequality metric for measuring the inequality externality accurately 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

[18] It may be useful to note that a special case such that

$$U_i(x_i, l_i, \bar{\theta}) = x_i - v(l_i) - \eta_i \bar{\theta} \tag{3}$$

fits these criteria for some disutility function of work effort $v(l)$. We could equally present this as $U_i(x_i, l_i, \bar{\theta}) = log\left(x_i - v(l_i) - \eta_i \bar{\theta}\right)$ with a suitably modified SWF. In Appendix C we solve the more general case of

$$U(x, l, \bar{\theta}) = u(x) - V(l) - \Gamma(\bar{\theta}). \tag{4}$$

[19] This is more complicated if g_i depends on the utility of the agent (and thus the externality). We discuss this in Appendix E.IV.

[20] This is particularly important in the more general mechanism design approach. The benefit of this is discussed in Appendices C and D.

effect on political capture is dependent on the proliferation of top incomes. Both relative poverty $\bar{\theta}_p$ and top income proliferation $\bar{\theta}_t$ are distributional metrics, which we represent in our framework by the distributional weights κ_p and κ_t for their respective inequality measurements $\bar{\theta}_p$ and $\bar{\theta}_t$. Take then an example with separability and homogeneity in these externality effects, such as in the simple example of $U = x - \eta_p\bar{\theta}_p - \eta_t\bar{\theta}_t$ where η_p and η_t indicate externality magnitudes. The total externality effect is $-\eta_p\bar{\theta}_p - \eta_t\bar{\theta}_t = -(\eta_p + \eta_t)\int_{\underline{z}}^{\bar{z}} \left(\frac{\eta_p}{\eta_p+\eta_t}\kappa_p(z) + \frac{\eta_t}{\eta_p+\eta_t}\kappa_t(z)\right) x(z)dH(z)$. The modified inequality metric is thus $\bar{\theta}_{true} = \int_{\underline{z}}^{\bar{z}} \left(\frac{\eta_p}{\eta_p+\eta_t}\kappa_p(z) + \frac{\eta_t}{\eta_p+\eta_t}\kappa_t(z)\right) x(z)dH(z)$, 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.

3.1 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; behavioral responses and the mechanical effect. The behavioral responses capture 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. The behavioral response represents a tax revenue loss, while the mechanical effect represents a tax revenue gain. These effects can be denoted by dB and dM, respectively. 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, dW. This effect dampens, but cannot cancel, the revenue-based benefit of the mechanical effect.[21]

These two channels (the behavioral and mechanical) also impact post-tax income inequality directly. This is not considered welfare-relevant in traditional models; in our model the inequality externality creates a welfare effect. In the following we will assume a negative inequality externality for simplicity.[22]

The mechanical (in)equality effect, which we denote $d\bar{I}_M$, gathers income from those above a certain tax bracket and redistributes this income as a flat dividend to all individuals.[23] This effect thus always reduces income inequality regardless of the tax bracket in question. If inequality is a negative externality, we have that $d\bar{I}_M$ incentivizes higher tax rates across the distribution.

[21] The revenue benefit will always equal or exceed the welfare loss of these agents due to the assumption of SWWs that are non-increasing in income. Generally, this channel does not change the relevant intuition.

[22] The same intuition with the opposite welfare direction holds for a positive externality.

[23] 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.

The (in)equality effect of the behavioral responses, which we denote by $d\bar{I}_B$, 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 effect of $d\bar{I}_B$ on optimal tax rates is thus dependent on the location of the tax bracket; it incentivizes lower tax rates at the bottom and higher tax rates at the top. Notably, this means that behavioral responses are welfare-*positive* at the top under a negative externality. The changing sign of $d\bar{I}_B$ across the distribution contrasts with the always negative dB, and is a key difference between the traditional revenue effects and the new equality impacts. The summary of this discussion is in Table II.

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.

Consider, then, an infinitesimally small tax increase $d\tau(z)$ for individuals at income z that leaves marginal tax rates unchanged at all other income levels. There are five welfare-pertinent effects of such a change. Three of these are well-known from the previous literature and discussed in Saez (2001), whereas the two equality consequences are new in our work ($d\bar{I}_M$ and $d\bar{I}_B$). At the optimum, the sum of the welfare effect of these five changes must equal to zero:

$$dM + dB + dW + d\bar{I}_M + d\bar{I}_B = 0 \tag{6}$$

We can now consider the sign change as compared to the standard optimal top marginal tax rates. At the bottom, where $d\bar{I}_M$ and $d\bar{I}_B$ 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 $d\bar{I}_M$ and $d\bar{I}_B$ 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 post-tax income inequality externality and lower with a positive post-tax income inequality externality. These are general results that do not depend on most of the assumptions we use for simplicity.[24]

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 6 and solving for $\tau(z)$. The full derivation is presented in Appendix D. Optimal marginal tax rates at income z in the presence of an inequality externality are;

$$\tau(z) = \frac{1 + \Upsilon(z) - \bar{G}(z)}{1 + \Upsilon(z) + \alpha(z)\epsilon(z) - \bar{G}(z)}, \tag{7}$$

where we use several of the standard parameters from the optimal taxation literature; the local Pareto parameter $\alpha(z) = \frac{zh(z)}{1-H(z)}$, the elasticity of earnings $\epsilon(z)$ (with respect to $1-\tau(z)$), and the average SWW above z denoted by $\bar{G}(z)$.[25] We have also introduced a new variable $\Upsilon(z) = \eta\alpha(z)\epsilon(z)\kappa(z) + \eta\bar{\kappa}(z)$, which indicates how the optimal rates differ from the standard Saez (2001) result. This new term consists of two parts which correspond to $d\bar{I}_B$ and $d\bar{I}_M$ respectively. We will now discuss these two terms in-depth.

The behavioral response: A Pigouvian tax The first term, $\eta\alpha(z)\epsilon(z)\kappa(z)$, comes from $d\bar{\theta}_B$ in Equation 6 and represents the behavioral responses of the individuals who are located at income z.[26] 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 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.[27]

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.

[24] This holds under any standard SWF and with income effects.

[25] $\alpha(z) = \frac{zh(z)}{1-H(z)}$ 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.

[26] Agents above z do not change their labor choice due to the assumption of no income effects.

[27] 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).

The term is affected by four parameters. First, how the 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 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 a consequence of the Mirrlees (1971) model assuming that economic equality in itself is not impactful for 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.

[28] The local Pareto parameter $\alpha(z) = \frac{zh(z)}{1-H(z)}$ 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 tax depending on the direction of the inequality externality. The optimal marginal tax rate at the top of a bounded distribution is the $\tau(z) = \frac{\eta\kappa(z)}{1+\eta\kappa(z)}$.

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.[31] 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.

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) 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 model parameters are the size of the inequality externality (represented by η) and the choice of the relevant inequality metric (represented by κ).

4 Optimal Income Taxation: Numerical Simulations

4.1 Numerical Specification

In this section we use numerical calculations to find optimal marginal tax rates in the presence of a post-tax income inequality externality. We assume quasi-linear utility in consumption $x = nl - T(nl)$, where n is the wage-earning ability distribution and l is work effort, a constant labor elasticity, and a linear homogeneous inequality externality. This implies the utility function

$$U(x, l, \bar{\theta}) = x - \frac{l^{(1+\frac{1}{E_c})}}{\left(1 + \frac{1}{E_c}\right)} - \eta\bar{\theta}, \tag{8}$$

which we will logarithmically scale in the SWF to introduce an inequality-aversion motive (such that the Utilitarian case is equal to $W = \int_i log(U_i)di$, for example). We solve the full Mirrlees non-linear optimal tax problem by using a mechanism design approach in Appendix C, and show that the resulting

[30] $\alpha(z)$ contributes to both channels. See footnote 28.
[31] For the Gini, for example, $\bar{\kappa}(z) = H(z)$.

tax rates are,

$$\frac{t(n)}{1-t(n)} = \eta\kappa(n) + \eta\left(1+\frac{1}{E_L}\right)\frac{F(n)}{\alpha(n)} + \left(1+\frac{1}{E_L}\right)\frac{1}{f(n)n}\int_n^\infty \left[1-\frac{W_{U(p)}}{\lambda}\right] dF(p),$$

where λ is the marginal value of public funds, E_L is the elasticity of labor supply, and $W_{U(p)}$ is the derivative of the SWF with respect to utility (capturing the aforementioned inequality aversion).

The underlying wage-earning ability distribution n is found through inverting the observed income distribution and using the known solution to the individual's maximization problem, following Saez (2001) and others. We use the US Distributional National Accounts micro-files to measure the 2019 U.S. labor income distribution.[32] We use the NBER TAXSIM model to find marginal tax rates on labor income for any given tax unit in the DINA files.[33] The main focus of the numerical simulations will be on how the inequality externality changes the results from the no-externality case; we thus largely follow the existing literature for the remaining model specification. For more details on the simulation procedure see Appendix E.I.

There are two further choices that are crucial for the simulations that are specific to the inequality externality. These are the choice of the relevant inequality metric $\bar{\theta}$ (e.g. the Gini coefficient in post-tax income) and the magnitude and direction η of the inequality externality. We detail these choices below.

4.1.1 Inequality metric The two inequality metrics we show in the main text both follow the general form in Equation 5. In the main specification (Section 4.2) we use the Gini, which has the following form:

$$\kappa_G(z) = 2H(z) - 1. \tag{9}$$

We also show results for a generalized Gini with weights of the following form (Section 4.3),

$$\kappa_T(z) = (q+1)H(z)^q - 1, \tag{10}$$

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. The weights of the Gini and the generalized Gini with $q = 4$ are plotted in Figure II.[34] We also show 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.III.

4.1.2 Inequality externality magnitude 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

[32] Described in Piketty et al. (2018), accessed at https://gabriel-zucman.eu/usdina/ on March 22nd 2023.

[33] Described in Feenberg and Coutts (1993), accessed at https://taxsim.nber.org/ on April 20th 2023. More details in Appendix E.I.

[34] The figure shows the relative weight of the income of any agent when calculating the specified inequality metric.

Figure II: Weights of Inequality Metrics

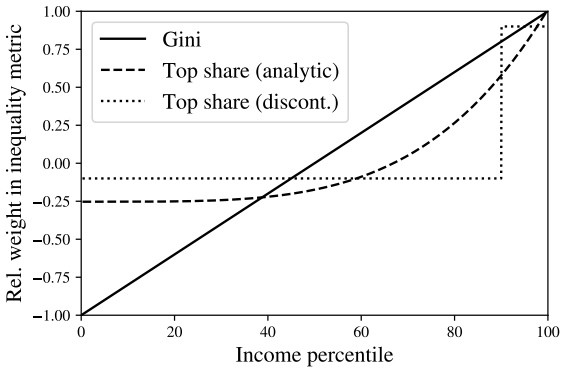
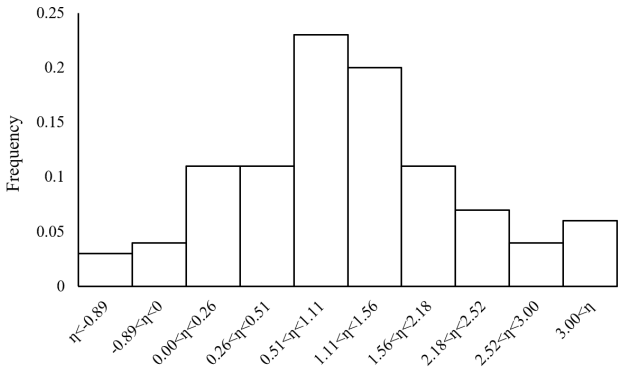


Figure III: Estimated η_G



Notes: Figure II 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 IV and A4, respectively). This corresponds to $\kappa(z)$ in the general expression $\bar{\theta} = \int_{\underline{z}}^{\bar{z}} \kappa(z)x(z)dH(z)$. More inequality metrics are explored in Appendix E.III. Figure III 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).

are unavoidable empirical challenges in calibrating such a number,[35] we do not aim to strongly argue for any one value. We instead use a range of realistic η_G to illustrate the potential tax policy consequences of various income inequality externalities. We present three different methods to understand the magnitudes of these η_G.

Correlation-based estimates 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_G. 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.[36] The correlation between income inequality and intentional homicide is strongly positive, and through this very simple approach we find $\eta_{G,homicides} \approx 0.07$.

This only represents a single externality channel, and the full η_G estimate would be found as $\eta_G = \sum_i \eta_{G,i}$. Extending this method to find all $\eta_{G,i}$, however, requires internationally comparable outcome data.[37] This is not a trivial requirement, and precludes the use of more detailed crime data.[38]

Other internationally comparable outcomes usually lack willingness-to-pay estimates. Despite this,

[35] 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 macroeconomic inequality.

[36] 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.

[37] We note that this approach assumes that the inequality externality operates on the country-level.

[38] Harrendorf (2018) notes the following: "Crime levels are not a valid measure of crime in different countries, with the possible exception of completed intentional homicide. Total crime rates depend mainly on the internationally differing quality of police work."

a reasonably-chosen willingness-to-pay can yield intuition through interpretable results. We will add another reference table here soon to show such estimates.

Experimental 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.[39] 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.[40]

The distribution is presented in Figure III. The median respondent in the survey has approximately $\eta_G = 1.00$. A majority of respondents have $0.26 < \eta_G < 2.18$.[41] 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.

Hypothetical exercise 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?*[42]

Answering the question creates equivalent incomes for differing inequality levels.[43] 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.

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.

4.2 Main Results: The Gini Externalities

Our main specifications, using the Gini as the post-tax income inequality metric, are presented in Figure IV. The introduction of even a small post-tax income inequality externality substantially changes the

[39] **?** finds comparable numbers for Uruguayan university students

[40] 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 assumption of a homogeneous inequality externality matches the respondents' perceived utility functions. All these reasons contribute to why we are using a *range* of η.

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

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

[43] While this method obscures the fact that this percentage is significantly different for other incomes – due to the assumption of a homogeneous inequality externality – the same percentage can be thought of as the total society-level income increase that would be required for indifference under a welfare function such that $W = \sum_i \left(x_i - \eta\bar{\theta}\right)$ (as in Sen, 1976).

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 8.

optimal tax structure. The effect is larger towards the top of the income distribution.

First, note that at the very top the Utilitarian and Rawlsian results converge under any externality value, as in the classical literature.[44] The magnitude of the inequality externality, however, is naturally impactful for the optimal top tax rate. This illustrates that a Rawlsian SWF, in itself, does not imply a maximum dislike of inequality.

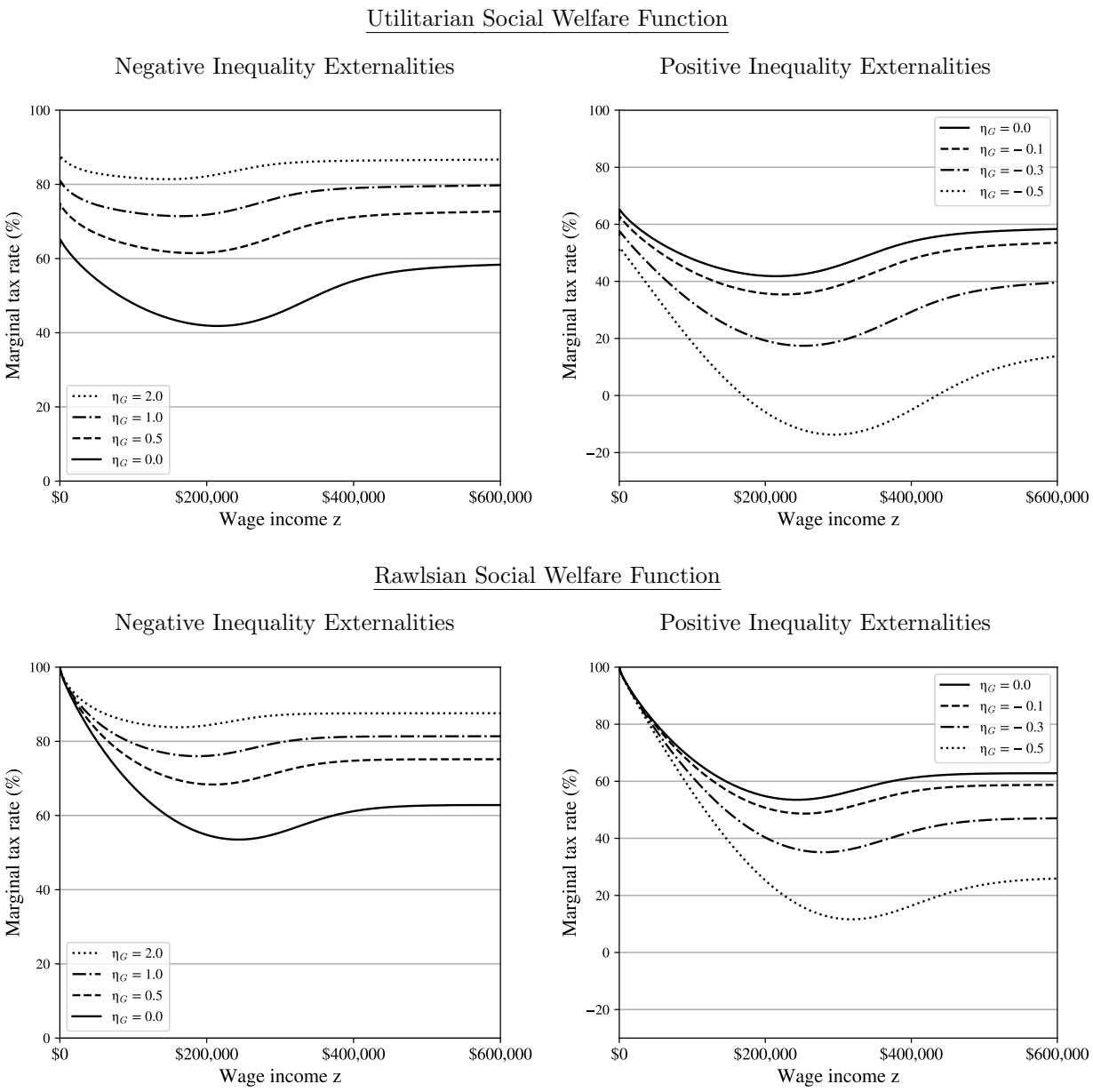
We thus begin by discussing optimal top marginal tax rates, which are the same in both the Rawlsian and the Utilitarian case. With no inequality externality, the optimal top rate is 63%. For $\eta_G = 1.00$, the value closest to the empirical externality estimate taken from Carlsson et al. (2005), it is 81%. When assuming a larger negative inequality externality, $\eta_G = 2.0$, the top rate increases to 88%. With a small positive inequality externality ($\eta_G = -0.5$), the optimal top marginal tax rate is only 26%. The inequality externality magnitude thus has a large impact on the optimal top tax rate; we will discuss this further in Section 4.4.

In the Utilitarian case, the marginal tax rates are shifted up or down from the no-externality case across 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.[45] The effects are larger near the top, which is particularly noticeable around the 95$^{\text{th}}$ percentile. 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 3.1. 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

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

[45] 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.II, 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 4.6).

Figure IV: 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 Utilitarian and Rawlsian cases converge when moving towards the top for a given externality value. 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.

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).[46] The effects of the inequality externality are mostly located above the 90^{th} percentile for both negative and positive externalities. Under a positive externality, top marginal tax rates approach zero around the 97^{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 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.[47] 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.

4.3 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 where we use the general top income share

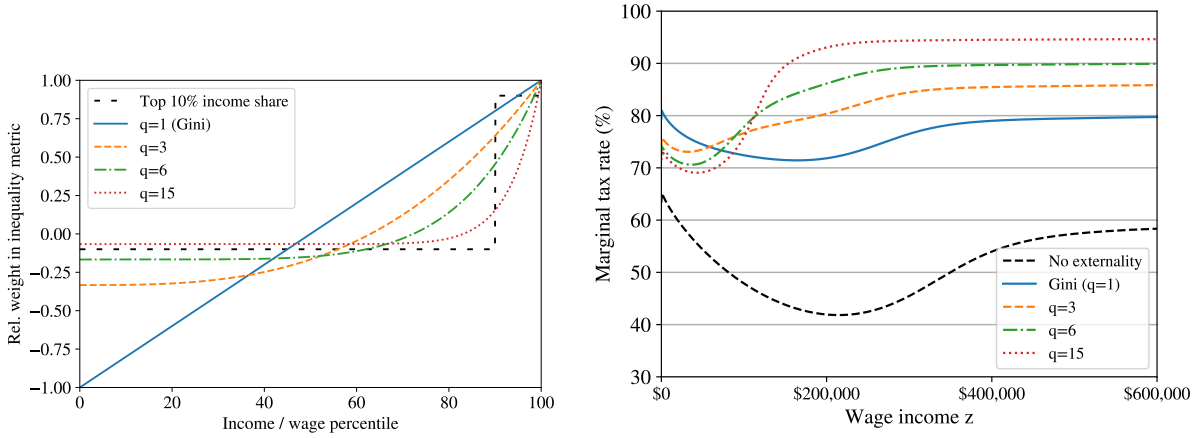
[46] The high optimal rates at the bottom of all the Rawlsian simulations are due to the large positive mechanical revenue effects of increasing bottom marginal tax rates. When one only cares about the very bottom agent, as in the Rawlsian case, redistributing away from any other agent is a net positive 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.

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

metric family $\kappa(z) = (q+1)H(z)^q - 1$, $q \in \mathbb{N}$as the relevant inequality measurement, with increasingly larger q. After $q = 1$, which defines the Gini coefficient, this inequality metric becomes increasingly top-focused and approximates top income share metrics.

We show a set of such inequality metrics and the effect of using them in the optimal tax calculation in Figure V. The externality in the optimal tax calculation is kept constant at the median result from Carlsson et al. (2005).[48]

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 no-externality case is shown as a reference in dotted black. The wage-earning ability distribution is the empirical income distribution, and the SWF is Utilitarian.

When we move away from the Gini towards a top income share, the effects of the externality are increasingly concentrated towards the top of the distribution. This should not be surprising given the increasing weight of top incomes in the inequality metric, although the *magnitude* of the effect is large. The inequality metric debehavioral 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' behavioralfined by $q = 15$ coupled with the median inequality externality from Carlsson et al. (2005) leads to optimal top marginal tax rates above 95%.

It is also noticeable 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 $\bar{\kappa}_{bottom}$).[49] In the numerical simulations, the mechanical effect is more powerful, indicating that the average marginal externality above is more

[48]The actual values of η change, as estimating η from the Carlsson et al. (2005) data requires an assumption about which inequality metric to use. Changing this inequality metric also changes the calculated η.

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

impactful than marginal externality of the tax bracket itself. Due to this, tax rates for the majority of Americans would be closer to the no-externality case under inequality metrics that focus more on top income shares.

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.III. 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.

4.4 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 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.

4.4.1 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%.[50]

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

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 the standard parameter values $1/\alpha$ or E_L, which we show in Tables A1 and A2. 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 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.

4.4.2 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.[51]

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 7,[52]

[51] The Rawlsian simulations in Section 4 provide numerical examples.

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

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

that is, if the equality effects of taxation are larger than the welfare effects.[53] 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 11 is zero above the very bottom earner. Thus, using the Gini values and a negative externality, the inequality simplifies to

$$\frac{H(z)}{\alpha(z)\epsilon(z)} > 1 - 2H(z), \tag{12}$$

which is independent of η and holds for any income above the median. 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.[54]

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.

4.5 U.S. social welfare weights with an inequality externality

As shown in Bourguignon and Spadaro (2012), it is possible to calculate the implied SWWs of the observed tax schedule given the relatively large assumption that the social planner is welfare-maximizing under the constraints of the optimal income tax problem.[55] This method is applied to the U.S. in Lockwood and Weinzierl (2016) and Hendren (2020), both of which generally find decreasing SWWs with income. Hendren (2020), which has more granular data, also notes an increase in SWWs towards the very top of the distribution.

These methods implicitly assume that no inequality externality is taken into account by the social planner when setting the tax schedule. However, U.S. citizens generally believe that inequality has negative consequences (Lobeck and Støstad, 2023). Such beliefs have also been voiced by prominent U.S. politicians.[56] It is thus natural to think that some concern for inequality itself could be included

$$\gamma\left[\kappa(n) + \frac{\zeta u_{x(n)}}{f(n)n}\int_n^\infty \left[\frac{\kappa(p)}{u_{x(p)}}\right] f(p)dp\right] > \frac{\zeta u_{x(n)}}{f(n)n}\int_n^\infty \left[W'(U(p))\right] f(p)dp, \tag{11}$$

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

[53] This follows from comparing Equation 7 to the revenue-maximizing tax rate, which is the same equation when $\bar{G}(z) = 0$ and $\eta = 0$.

[54] For a positive externality the inequality changes directions.

[55] This is an unlikely assumption, as discussed in Lockwood and Weinzierl (2016). Nonetheless, it is useful to see how current tax systems can be rationalized in the framework of optimal taxation.

[56] For example Obama (2011): "This kind of inequality – a level that we haven't seen since the Great Depression – hurts

in the income tax schedule design. If so, we show in Appendix F that the implied SWW $g(z)$ is,

$$g(z) = -\frac{1}{h(z)}\frac{d}{dz}\left[(1-H(z))\,(1+\Upsilon(z)) - \frac{\tau(z)}{(1-\tau(z))}zh(z)\epsilon(z)\right],$$

which differs from the standard case by $\Upsilon(z) = \eta\alpha(z)\epsilon(z)\kappa(z) + \eta\bar{\kappa}(z)$.[57] Intuitively, the implied inequality aversion in a given tax system can come from either the SWF $g(z)$ or externality motivations $\Upsilon(z)$. If externality motivations to avoid inequality were greater when designing a given tax schedule, the same tax schedule will imply that the SWWs in the same design process were less progressive.

In Figure VI we show $g(z)$ of the 2019 U.S. tax system under standard specifications, assuming the social planner has taken into account various negative post-tax Gini income inequality externalities. The model specification is further discussed in Appendix F.

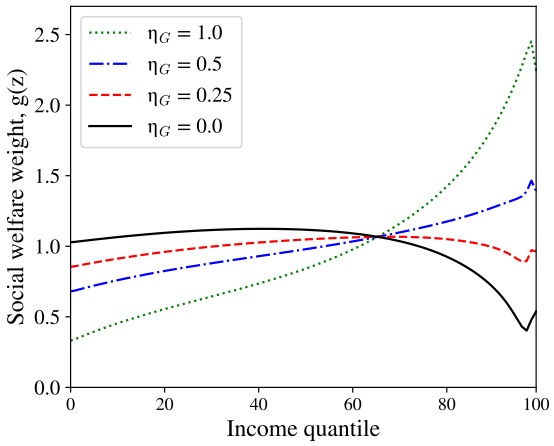


Figure VI: Implied social welfare weights $g(z)$ from the 2019 U.S. tax system under various negative inequality externalities η_G.

The standard case of no inequality externality ($\eta = 0$) has generally decreasing welfare weights with income with an upward bend towards the top of the distribution, similar to Hendren (2020).[58] Introducing a negative inequality externality ($\eta > 0$) changes implied SWWs quickly, however. Implied SWWs are relatively flat for $\eta_G = 0.25$, indicating that all the inequality aversion in the tax system is accounted for by such an inequality externality.[59] The implied SWWs are increasing for $\eta_G = 0.5$, and even more so for $\eta_G = 1.0$. For $\eta_G = 1.0$, the social planner values one dollar at the top equally to five dollars at the bottom.[60]

This illustrates a key finding. The current U.S. tax schedule cannot accomodate both a socially progressive transfer motive and be significantly concerned with inequality's societal effects. The social planner may have progressive $g(z)$, implying that the government prefers to transfer one dollar from the poor to the rich ceteris paribus (as in Lockwood and Weinzierl (2016) and Hendren (2020)). The social

us all."

[57] A few technical points: We use the income density directly, as in Lockwood and Weinzierl (2016), instead of the "virtual" earnings density, as employed in Hendren (2020) and the rest of this work. Due to this the elasticity we use is technically defined to include the circularity between the "virtual" earnings density and the observed income density (Jacquet et al., 2013). This is unlikely to significantly change results due to the absence of pronounced bunching in the actual U.S. income distribution (Saez, 2010). We assume no income effects and no extensive margin behavioral responses for simplicity. A more detailed approach for the no-externality case can be found in Jacobs et al. (2017), which also notes that these factors are empirically small.

[58] The social welfare weights at the top in Hendren (2020) are larger than in our case under no inequality externality (a minimum of $g(z) = 0.25$ in our case versus $g(z) \approx 0.57$ in Hendren (2020)). This is largely due to differences in the real-world marginal tax rates used to calculate the inverse optimum.

[59] It is useful to find the η_G above which social welfare weights become regressive. There are various ways to do this. The full linear trend is flat at roughly $\eta \approx 0.21$. As $g(z)$ is slightly increasing below the median, it is also useful to note the set of η_G which correspond to $G(z_{median}) > g(z_{median})$, which indicates that the average social welfare weight above the median is higher than that of the median. The corresponding externality magnitudes are $\eta_G > 0.28$.

[60] For $\eta = 2.0$ we find negative SWWs at the bottom, indicating that the social planner would want to remove income at the bottom if this did not also increase inequality itself.

planner may also have $\eta_G \geq 0.25$, implying a negative inequality externality of a potentially sizable magnitude. However, it cannot have both. The inequality aversion in the system as a whole is simply too small for this to be the case.

But the U.S. social planner may also have a *positive* inequality externality in mind. An inequality externality focusing on positive benefits from top-incomes could explain the puzzle of increasing SWWs at the top from Hendren (2020) (a result which is also visible in Figure VI). If the social planner believes top-income inequality is strongly beneficial for society – through increasing innovation, economic growth, or charitable giving, for example – the implied SWWs may still be everywhere decreasing. We illustrate this graphically in Figure A7.

Several other conclusions from the inverse optimal tax literature could change if inequality externality beliefs are a salient feature of policy-making. Lockwood and Weinzierl (2016) note that TRA86 implies a substantial change in SWWs over a short time period, which could be resolved if TRA86 instead represented a change in the *inequality externality belief* of the social planner (beliefs that are arguably more malleable than the SWWs themselves, Lobeck and Støstad (2023)). Lockwood and Weinzierl (2016) also calculate the welfare cost of the inequality in income growth between 1980 and 2010 as 4.3% of total economic growth in the period, an estimate which depends on inequality not being an externality.[61] Similarly, Hendren (2020) creates a preference ordering of countries' income distributions based on implied SWWs that is only valid if inequality is not an externality. More generally, the inverse-optimum literature is an example of a welfare-based framework that is relatively fragile to the inclusion of an inequality externality.

4.6 Other types of inequality externalities

The preceding sections have discussed a *post-tax income* inequality externality. While such an externality could be reasonable through several motivations – some of which we outline in Section 5.1 – 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.

[61] At least two parts of this calculation would be affected by an inequality externality. First, the implied SWWs from the inverse-optimum method would change under an inequality externality, as shown in this section. Second, the total welfare implications of income changes would be affected by an inequality externality.

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.[62] The full derivation is in Appendix G.I.

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 VII, 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 47% at the bottom to 85% at the top when $\eta_{pre} = 0.6$.[63]

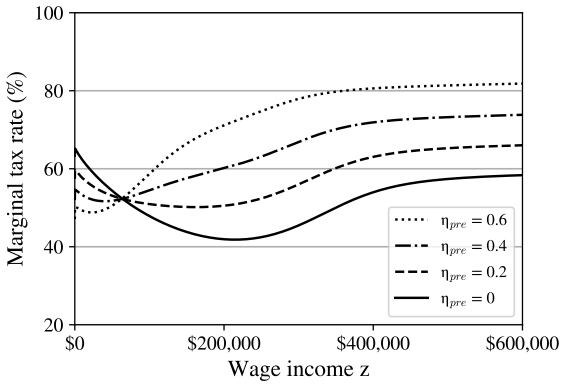


Figure VII: Optimal income tax rates with a pre-tax income inequality externality. The social planner is Utilitarian, and the remaining specification is identical to Figure IV.

Interestingly, the pre-tax income inequality externality almost removes the well-known U-shape of optimal marginal tax rates from the classical literature. Instead, the marginal tax rates generally increase 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.[64]

[62] There is a subtle point to be made here about the magnitude of η_{pre}. Pre-tax income inequality is generally higher than post-tax income inequality, which influences the shadow price of each unit of inequality and hence η. To keep externality sizes similar we thus use a lower set of η_{pre} in Figure VII than the corresponding η_G in Figure IV.

[63] This corresponds roughly to $\eta_G = 2.0$ in Figure IV.

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

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 G.II. Assuming that negative weights are acceptable, using the modified SWWs $\bar{G}'(z) = \bar{G}(z) - \eta_U \cdot \bar{\kappa}(z)$ allows this result to be simplified to the standard Mirrlees case without the need for empirical variables in the modified income-based welfare weights.[65] Further, this result can be approximated in the mechanism design case through utility-based SWWs, unlike both the pre-tax and post-tax externality results.

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.[66] 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 SWWs.

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 IV summarizes these results.

Table IV

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.

5 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:

[65] 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 $\epsilon(z)$.

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

- Equality itself becomes policy-relevant and has an associated shadow price.[67] The trade-off between income maximization at the bottom and the preferred inequality level becomes relevant.

- Introducing an inequality externality presents an efficiency-based reason for the state to distributionally interfere in otherwise well-functioning markets.

- 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 SWWs.

- 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.

- 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 3 showed. It is negative if inequality increases ($\frac{\partial \theta}{\partial x_i} > 0$)[68] 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.[69] The total effect depends on the relative importance of equality and consumption, a version of the familiar equity-efficiency trade-off.

This last 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

[67] This shadow price corresponds to γ in Equation 41 and η in Equation 7. In general the shadow price is endogenous to the system solution; in the simplified small-perturbation solution it is constant.

[68] 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 $\frac{\partial I_{\text{Gini}}}{\partial x_i} = \kappa(x_i)$.

[69] 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.

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.

5.1 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:[70]

- 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 $\bar{P} = \varphi(\boldsymbol{O})$, 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.[71]

- Innovation / Economic growth: Assume that agents view high inequality as an incentive to work such that l_i and thus x_i are increasing functions of income inequality $\bar{\theta} = I(\boldsymbol{x})$. If so, utility can be written as $U_i(x_i(\bar{\theta}), l_i(\bar{\theta}), ...)$ 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 variable $p_{i,j}$ 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.[72] We also create micro-foundations for inequality effects on trust,

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

[71] 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.

[72] 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

crime, and political capture in Appendix H. 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.

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).

5.2 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 when income inequality increases. 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 exter-

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.

nality. 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.

Third, cost-benefit analysis-type results that depend on SWWs may be fragile to the inclusion of an inequality externality. Our results in Section 4.5 imply that SWWs with both a progressive SWF and

5.3 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 SWWs cannot approximate the effects of an economic inequality externality. The optimal marginal tax rates in the mechanism design case, shown in Equation 40, illustrates one case where even best-designed SWWs would fail to approximate the inequality externality.

There are two main reasons for why SWWs 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, SWWs cannot change individual behavior. In addition to these two points, approximating inequality's societal effects – real-world phenomena – through SWWs 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 SWWs 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 7, 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. The weights could also have negative values without breaking the Pareto principle.[73]

An additive resource inequality in the social welfare function If we move away from strict SWWs one could imagine a SWF of the form $\int_i g_i U_i(x_i, l_i, ...)di - \Gamma(\bar{\theta})$, 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 3 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.

6 Conclusion

This paper has introduced the concept of an *inequality externality* and has particularly focused on an *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 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

[73] The negative weights would imply breaking the Pareto principle in *income*, but not *utility*.

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. Given that policy makers believe that inequality itself is concerning, the analysis presented here thus recommends more progressive taxes than those previously suggested by Saez (2001), Piketty et al. (2014), and others. We present two main new insights to the optimal income taxation literature, both 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 81% 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.

Second: The inequality aversion implied by the current U.S. income tax system is insufficient to explain both progressive social welfare weights *and* a realistic concern for inequality's effects on society. While the tax system may imply a preference for progressive redistribution *or* a negative inequality externality of a substantial magnitude, it is currently not able to accommodate both objectives effectively.

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.

References

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

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

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). Paternalism against Veblen: Optimal Taxation and Non-Respected Preferences for Social Comparisons. *American Economic Journal: Economic Policy 10*(1), 39–76.

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

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

Auclert, A. and M. Rognlie (2018). Inequality and Aggregate Demand. *National Bureau of Economic Research* (24280).

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

Benabou, R. (1996). Inequality and Growth. *National Bureau of Economic Research Macroeconomics Annual 11*, 11–74.

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 (2022). Dissecting Inequality-averse Preferences. *Journal of Economic Behavior & Organization 200*, 782–802.

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

Bourguignon, F. and A. Spadaro (2012). Tax-benefit Revealed Social Preferences. *The Journal of Economic Inequality 10*, 75–108.

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.

Breza, E., S. Kaur, and Y. Shamdasani (2018). The Morale Effects of Pay Inequality. *The Quarterly Journal of Economics 133*(2), 611–663.

Card, D., A. Mas, E. Moretti, and E. Saez (2012). Inequality at Work: The Effect of Peer Salaries on Job Satisfaction. *American Economic Review 102*(6), 2981–3003.

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.

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.

Feenberg, D. and E. Coutts (1993). An introduction to the taxsim model. *Journal of Policy Analysis and management 12*(1), 189–194.

Fehr, D., H. Rau, S. T. Trautmann, and Y. Xu (2020). Inequality, Fairness and Social Capital. *European Economic Review 129*, 103566.

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

Gauthier, S. and G. Laroque (2009). Separability and Public Finance. *Journal of Public Economics 93*(11-12), 1168–1174.

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

Greenspan, A. (2014). Comments: National Association of Business Economists Conference. https://www.thefiscaltimes.com/Articles/2014/02/25/These-Top-Economists-All-Agree-Biggest-Problem-US-Faces.

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.

Harrendorf, S. (2018). Prospects, Problems, and Pitfalls in Comparative Analyses of Criminal Justice Data. *Crime and Justice 47*(1), 159–207.

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.

Hendren, N. (2020). Measuring Economic Efficiency using Inverse-Optimum Weights. *Journal of Public Economics 187*, 104198.

Jacobs, B. (2018). The Marginal Cost of Public Funds is One at the Optimal Tax System. *International Tax and Public Finance 25*, 883–912.

Jacobs, B. and R. A. De Mooij (2015). Pigou meets Mirrlees: On the Irrelevance of Tax Distortions for the Second-best Pigouvian Tax. *Journal of Environmental Economics and Management 71*, 90–108.

Jacobs, B., E. L. Jongen, and F. T. Zoutman (2017). Revealed Social Preferences of Dutch Political Parties. *Journal of Public Economics 156*, 81–100.

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

Jacquet, L., E. Lehmann, and B. Van der Linden (2013). Optimal Redistributive Taxation with Both Extensive and Intensive Responses. *Journal of Economic Theory 148*(5), 1770–1805.

Jones, C. I. (2022). Taxing Top Incomes in a World of Ideas. *Journal of Political Economy 130*(9), 2227–2274.

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.

Lobeck, M. and M. Støstad (2023). Inequality Externality Beliefs and Redistributive Preferences. *Working paper (Forthcoming)*.

Lockwood, B. B. and M. Weinzierl (2016). Positive and Normative Judgments Implicit in US Tax Policy, and the Costs of Unequal Growth and Recessions. *Journal of Monetary Economics 77*, 30–47.

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*.

Mian, A. R., L. Straub, and A. Sufi (2020). The Saving Glut of the Rich. Working Paper 26941.

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

Obama, B. (2011). Remarks by the President on the Economy in Osawatomie, Kansas. https://obamawhitehouse.archives.gov/the-press-office/2011/12/06/remarks-president-economy-osawatomie-kansas.

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

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.

Piketty, T., E. Saez, and G. Zucman (2018). Distributional National Accounts: Methods and Estimates for the United States. *The Quarterly Journal of Economics 133*(2), 553–609.

Pirttilä, J. and M. Tuomala (1997). Income Tax, Commodity Tax and Environmental Policy. *International Tax and Public Finance 4*, 379–393.

Plato (2016). *The Laws of Plato*. CreateSpace Independent Publishing Platform. Estimated written 360 B.C.. Translated by Benjamin Jowett.

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. (2010). Do Taxpayers Bunch at Kink Points? *American Economic Journal: Economic Policy 2*(3), 180–212.

Saez, E., J. Slemrod, and S. H. Giertz (2012). The Elasticity of Taxable Income with Respect to Marginal Tax Rates: A Critical Review. *Journal of Economic Literature 50*(1), 3–50.

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 (2022). Gini and optimal income taxation by rank. *American Economic Journal: Economic Policy 14*(3), 352–379.

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.

Tsyvinski, A. and N. Werquin (2017). Generalized Compensation Principle. Technical report, National Bureau of Economic Research.

Tuomala, M. (2016). *Optimal Redistributive Taxation*. Oxford University Press.

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.

Wilkinson, R. and K. Pickett (2009). *The Spirit Level: Why More Equal Societies Almost Always Do Better*. Allen Lane.

Xu, Y. and G. Marandola (2022). The (Negative) Effects of inequality on Social Capital. *Journal of Economic Surveys*.

[FOR ONLINE PUBLICATION]

Appendix

A Discussion on Equations 1 and 2

Equations 1 and 2 show the following simplification:

$$U_i(x_i(\bar{\theta}), \bar{\theta}, \overrightarrow{\Psi}(\bar{\theta}), ...) \rightarrow \tilde{U}_i(\tilde{x_i}, \bar{\theta}, ...). \tag{13}$$

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.

This does not change the intuition of the problem, however. 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 necessarily the only solution to every inequality externality channel.

What does this mean in a practical example? We return to the example of inequality and crime. Suppose the social planner wishes to deal with inequality externality problems indirectly (such as through crime prevention). The required revenue $R = R_0 + p(c(\bar{\theta}))$ is a function of original revenue requirements R_0 and crime prevention p, which is a function of crime c, which is a function of inequality $\bar{\theta}$. In other words, the required revenue R is a function of inequality θ.

The social planner is then faced with a problem of maximizing social welfare $W = \int_i g_i U(x_i, l_i) di$ under the constraint $R_0 + p(c(\bar{\theta})) \leq \int_{\underline{n}}^{\overline{n}} T(nl) f(n) dn$. Crucially, minimizing inequality also increases "effective" revenue R_0 if total revenue collected R is kept constant. In other words, the social planner's goal once again becomes to maximize revenue $\int_{\underline{n}}^{\overline{n}} T(nl) f(n) dn$ and minimize inequality $\bar{\theta}$. Although the magnitude of the revenue-inequality trade-off has changed – in accordance with how expensive crime prevention programs are compared to reducing inequality itself – the core intuition of the problem has not changed. A naïve social planner would choose the suboptimal tax schedule, not taking into account that its choice of tax schedule changes inequality which affects the effective revenue collected.[74]

The main takeaway is that inequality's societal consequences come with a cost. Whether the social planner directly or indirectly deals with this cost is relatively unimportant to the main conclusions sketched in the remainder of the article. Again, the choice simply changes the *magnitude* of the inequality externality.

As an aside, we also note that a similar simplification as (13) is usually made implicitly when including consumption x_i in the utility function. The benefit of consumption to individuals is often

[74] Interestingly, such a social planner can also have optimal tax rates above the revenue-maximizing rate. The foregone revenue comes with the benefit of lowered inequality, leaving more revenue for the "standard" revenue requirements R_0 – in other words, potentially raising effective revenue collection.

not just consumption *per se*, but also what consumption brings them – such as improved health, social status, and so on. As in our case, there are many such potential channels that are usually not explicitly modeled, but can be captured in a vector $\overrightarrow{\Psi}'$. In effect, the following simplification is implicitly made,

$$U_i(x_i, \overrightarrow{\Psi}'(x_i), ...) \rightarrow \hat{U}_i(x_i, ...), \tag{14}$$

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.[75]

B Varying welfare weights

Another approach to introducing a dislike of inequality, common in the optimal income taxation literature, is varying the utility-based SWWs. 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* SWWs, 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 SWWs, the choice of a social planner is no longer a purely philosophical concept. This is problematic as it conflicts with

[75] 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.

standard welfarist traditions in several ways. The most obvious case is when large inequality effects lead to negative SWWs, which breaks the Pareto principle. Another counter-intuitive example is how a truly Utilitarian social planner would need to use non-Utilitarian SWWs 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, SWWs, and an inequality externality. The former two are distinct from the inequality externality, 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.[76]

(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*.[77]

More generally, diminishing marginal utility of income and SWWs 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 5.3.

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

[76]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.

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

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

The social planner aims to maximize:

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

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, l_i is work effort, and $\theta(\boldsymbol{x})$ is inequality as a function of all incomes $\boldsymbol{x}$.

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

$$\frac{\partial U(x_i, l_i, \theta(\boldsymbol{x}))}{\partial l_i} = \frac{\partial U(x_i, l_i, \theta(\boldsymbol{x}))}{\partial x_i} + \frac{1}{g_i} \int_j g_j \frac{\partial U(x_j, l_j, \theta(\boldsymbol{x}))}{\partial \theta(\boldsymbol{x})} \frac{\partial \theta(\boldsymbol{x})}{\partial x_i} dj \tag{15}$$

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, $\frac{\partial \hat{U}(x_j, l_j)}{\partial \theta(\boldsymbol{x})} = 0$ and the second term on the right-hand side of Equation 15 is zero. The solution to the social planner's problem would thus involve $\frac{\partial \hat{U}(x_i, l_i)}{\partial x_i} = \frac{\partial \hat{U}(x_i, l_i)}{\partial l_i} \forall \hat{g}_i \neq 0$, which is equivalent to $\frac{\partial U(x_i, l_i, \theta(\boldsymbol{x}))}{\partial x_i} = \frac{\partial U(x_i, l_i, \theta(\boldsymbol{x}))}{\partial l_i} \forall \hat{g}_i \neq 0$. However, in the correct solution we are trying to approximate, $\frac{\partial U(x_i, l_i, \theta(\boldsymbol{x}))}{\partial x_i} \neq \frac{\partial U(x_i, l_i, \theta(\boldsymbol{x}))}{\partial h_i} \forall g_i \neq 0$. 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 variables in the utility function. If x_j is changed, Equation 15 implies that it will affect the FOC for i. In the modified solution with $\hat{U}$, it has no effect. To correctly specify $\hat{U}(x_i, l_i)$, one would need x_j or l_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 first solve the problem in a mechanism design framework, where we fully specify the utility function as,

$$U(x, l, \bar{\theta}) = u(x) - V(l) - \Gamma(\bar{\theta}), \tag{16}$$

where u is the utility of consumption (after-tax income) x, $V(l)$ is the disutility of work l, and Γ is disutility from post-tax income inequality $\bar{\theta}$ (a society-wide parameter, indicated by the overbar). The functions $u(x)$, $V(l)$ and $\Gamma(\bar{\theta})$ are continuous and second-order differentiable in their arguments. The function $u(x)$ is strictly concave in x, $V(l)$ is strictly convex in l, and $\Gamma(\bar{\theta})$ has no restriction. We also have that $u_x > 0$ and $V_l > 0$ where subscripts indicate partial derivatives. Equation (16) 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. However, their actions have welfare-pertinent effects as the change in income inequality affects every other agent. Note that due to this assumption (and separability), the individual choice of (x, l) does not require that the individual is aware of or estimates the magnitude of the inequality externality.

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 nl, and total consumption is $x = nl - T(nl)$, where $T(\cdot)$ is the tax schedule. The individual maximizes utility by choosing work effort l given n and $T(\cdot)$. The utility-maximising values of consumption and hours worked are written as

$$x(n), l(n). \tag{17}$$

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,

$$\max_{T(\cdot)} \int_{\underline{n}}^{\overline{n}} W(U(x(n), l(n), \bar{\theta}))dF(n). \tag{18}$$

Notice that formulating individual utility as (16) avoids the complication of potentially heterogeneous effects of inequality if the social planner is strictly Utilitarian – 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 (18) 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:

$$R \leq \int_{\underline{n}}^{\overline{n}} T(nl)f(n)dn. \tag{19}$$

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:

$$u(x(n)) - V(l(n)) \geq u(x(\hat{n})) - V(l(\hat{n})) \tag{20}$$

where $x(\hat{n})$ and $l(\hat{n})$ are, respectively, the consumption and hours worked if the agent masquerades as someone with ability $\hat{n}$, possibly different from n. The IC constraint (20) 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 (17). But it is also partly the result of

decisions by the social planner, captured in the tax function T and therefore embedded in (17). We can represent this relationship as the following *inequality condition:*

$$\bar{\theta} = I\left(\boldsymbol{x}, F\right) \tag{21}$$

where $I\left(\cdot,\cdot\right)$ is an inequality measure, $\boldsymbol{x}\left(\cdot\right)$ is the full set of consumption choices from (17) and $F\left(\cdot\right)$ is the distribution function for n.

To complete the model we need an inequality metric $I\left(\cdot,\cdot\right)$. We begin with a specific form of the (absolute) Gini coefficient in after-tax income taken from Cowell (2000):

$$I_{\text{Gini}}\left(\boldsymbol{x}, F^x\right) = \int_{\underline{n}}^{\overline{n}} \kappa^x(x(n))x(n)dF^x(x(n)), \tag{22}$$

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

$$\kappa^x(x) = 2F^x(x) - 1 \tag{23}$$

is an expression for the weight of the agent in the Gini dependent on the cumulative density of post-tax income.

This form of the inequality metric presents a difficult endogeneity problem when taking derivatives for x, namely that the weight itself depends on the distribution of post-tax income.[78] It can, however, be modified to a simpler form. If there is rank-equivalency between income and ability, we can use the fact that $F^x(x) = F^n(n)$ to see that $\kappa^x(x) = \kappa^n(n)$. Thus we can re-write the inequality metric as,

$$I_{\text{Gini}}\left(\boldsymbol{x}, F\right) = \int_{\underline{n}}^{\overline{n}} \kappa(n)x(n)dF(n), \tag{24}$$

where

$$\kappa(n) = 2F(n) - 1. \tag{25}$$

Here we have removed the superscripts for notational simplicity. This shows that the absolute Gini in post-tax income can be calculated as a sum of weighted post-tax incomes in the population, where the weight $\kappa(n)$ depends only on the *rank* of the agent in the wage-earning ability distribution $F(n)$, which is constant and exogenous by assumption. Simula and Trannoy (2022), 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.[79]

[78] In short, the inequality weight $\kappa^x(x_i)$ depends on x_i and thus $T(z_i)$, which is problematic since all other $\kappa^x(x_j)$ also depend on $\kappa(x_i)$.

[79] In the small perturbation approach we use that inequality weights are unchanged with small perturbations around the optimum given that the individual's second-order conditions hold.

Using (22), condition (21) becomes

$$I_{Gini} = \int_{\underline{n}}^{\overline{n}} \left[2F(n) - 1\right] x(n) dF(n). \tag{26}$$

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), which simply changes $\kappa(n)$.

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: $x = c(nl)$. The individual maximization implies,

$$\frac{dU}{dl} = 0 = u'c'n - V', \tag{27}$$

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

$$\frac{dU}{dn} = u'c'l \tag{28}$$

Taken together these two imply :

$$\frac{dU}{dn} = \frac{V'l}{n} =: g\left(n\right) \tag{29}$$

We can write $T = nl - x$, where x is after-tax consumption.[80] From this and the IC constraint, we observe that the tax schedule implicitly defines both work hours and total individual utility. Instead of setting the tax schedule T, then, we can say that the social planner chooses work hour schedules $l(n)$, utility schedules $U(n)$, and the inequality level $\bar{\theta}$.

The Lagrangian of the full problem classified in Equations 18–21 is,

$$\begin{aligned} L = \int_{\underline{n}}^{\overline{n}} W(U(n)) f(n) dn + \lambda(\int_{\underline{n}}^{\overline{n}} \left[nl(n) - x\right] f(n) dn) \\ + \int_{\underline{n}}^{\overline{n}} \alpha(n) \left[\frac{dU}{dn} - g(n)\right] dn + \gamma \left[\bar{\theta} - I_{Gini}\right] \end{aligned} \tag{30}$$

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

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

account and combining the rest of the integrals, we have:

$$L = \int_0^\infty \left[W(U(n)) + \lambda(nl(n) - x)\right] f(n) - \alpha(n)g(n) - \alpha'(n)U(n)dn \\ +\alpha(\infty)U(\infty) - \alpha(0)U(0) + \gamma\left[\bar{\theta} - I_{Gini}\right] \tag{31}$$

Substituting in (26) for I_{Gini}, the Lagrangian becomes:

$$L = \int_0^\infty \Bigg[\left(W(U(n)) + \lambda\left[nl(n) - x\right] - \gamma\kappa(n)x\right) f(n) - \alpha(n)g(n) \\ -\alpha'(n)U(n)\Bigg]dn + \alpha(\infty)U(\infty) - \alpha(0)U(0) + \gamma\bar{\theta} \tag{32}$$

From this we can find the first-order conditions with respect to $l(n)$, $U(n)$, and $\bar{\theta}$, as these variables together will implicitly set the tax schedule.[81] Before we begin, note that we can rewrite $x = y(l, U, \bar{\theta}) = u^{-1}(U + V(l) + \Gamma(\bar{\theta}))$, and find expressions for the derivatives y_l, y_U, and $y_{\bar{\theta}}$.[82] The first order conditions are the following:

$$U: \quad 0 = \left[W_{U(n)} - \lambda y_U\right] f(n) - \alpha'(n) - \gamma\kappa(n)f(n)y_U \tag{33}$$

$$l: \quad 0 = \lambda(n - y_l)f(n) - \alpha(n)\frac{V_{ll}l + V_l}{n} - \gamma\kappa(n)f(n)y_l \tag{34}$$

$$\bar{\theta}: \quad 0 = \gamma - \int_0^\infty \gamma\kappa(n)f(n)y_{\bar{\theta}}dn - \int_0^\infty \lambda y_{\bar{\theta}}f(n)dn \tag{35}$$

In the FOC for l we have used that $g = \frac{V_l l}{n}$ from Equation (29), and that $\frac{dg}{dl} = \frac{V_{ll}l + V_l}{n}$.

Equation 35 implies,

$$\frac{\gamma}{\lambda} = \frac{\int_0^\infty \frac{\Gamma_{\bar{\theta}}}{u_x} f(n)dn}{1 - \int_0^\infty \frac{\Gamma_{\bar{\theta}}}{u_x}\kappa(n)f(n)dn} \tag{36}$$

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. Note that $\lambda = 1$ in an optimal tax system (Jacobs, 2018). 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. If $\Gamma(\bar{\theta}) = 0$, as in the standard case when the inequality externality does not exist, then $\gamma = 0$, and the optimal tax rates in 40 become identical to those from Diamond (1998). If we assume quasi-linearity in consumption and a linear inequality externality of the form $\Gamma(\theta) = \eta\theta$, then $\frac{\gamma}{\lambda} = \eta$ as the integral term in the denominator vanishes when the externality is homogeneous.[83] We explore individual-specific η_i in Appendix E.IV.

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

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

[83] With a squared inequality externality, which we discuss specifically in Appendix C.II, the term in the utility function

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 33:[84]

$$\alpha(n) = \int_n^\infty \left[\frac{\lambda + \gamma\kappa(p)}{u_{x(p)}} - W_{U(p)} \right] f(p)dp \tag{37}$$

And substitute this into Equation (34):

$$0 = \lambda(n - y_l)f(n) - \gamma\kappa(n)f(n)y_l - \frac{V_{ll}l + V_l}{n} \int_n^\infty \left[\frac{\lambda + \gamma\kappa(p)}{u_{x(p)}} - W_{U(p)} \right] f(p)dp \tag{38}$$

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

We have that $\frac{n-y_l}{y_l} = \frac{nu_{x(n)}}{V_l} - 1 = \frac{t(n)}{1-t(n)} - 1 = \frac{t(n)}{1-t(n)}$, so we quickly have the expression for optimal marginal tax rates:

$$\frac{t(n)}{1 - t(n)} = \frac{\zeta_n u_{x(n)}}{f(n)n} \int_n^\infty \left[\frac{1}{u_{x(p)}} - \frac{W_{U(p)}}{\lambda} \right] dF(p) + \frac{\gamma}{\lambda} \left[\kappa(n) + \frac{\zeta_n u_{x(n)}}{f(n)n} \int_n^\infty \frac{\kappa(p)}{u_{x(p)}} dF(p) \right], \tag{40}$$

$\zeta_n = \frac{V_{ll}l}{V_l} + 1$ is a term closely related to the inverse compensated elasticity of labor.[85] The first two terms are equivalent to the traditional OIT terms. However, there is a potentially subtle difference in the term containing $W_{U(p)}$. Depending on the SWF, the weights implied by this term could be dependent on the inequality externality term itself. Take, for example, $W = \int_i log(U_i)di$. Here the introduction of an inequality externality would change $W_{U(p)}$ and thus the optimal tax rates. The numerical simulations we consider in Section 4 take this into account, and in practice this is not what empirically drives the differences to the no-externality solution.

By denoting the part of the optimal tax function found in Diamond (1998) as $\frac{t(n)_{orig}}{1-t(n)_{orig}}$, we can isolate and evaluate the effect of the inequality externality.

$$\frac{t(n)}{1 - t(n)} = \frac{\gamma}{\lambda} \left[\kappa(n) + \frac{\zeta}{f(n)n} \int_n^\infty \frac{u_{x(n)}}{u_{x(p)}} \kappa(p) dF(p) \right] + \frac{t(n)_{orig}}{1 - t(n)_{orig}} \tag{41}$$

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 ($\kappa(n)$,

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} .

[84] 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.

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

which changes sign across the distribution), and

- (ii) a change to the redistributive benefit of the tax, $\frac{\gamma}{\lambda}\frac{\zeta_n u_{x(n)}}{f(n)n}\int_n^\infty \frac{\kappa(p)}{u_{x(p)}}dF(p)$, in effect modifying the SWWs. Beyond standard Mirrleesian parameters, this latter 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, $\int_n^\infty \frac{\kappa(p)}{u_{x(p)}}dF(p)$.

As noted in the discussion around Equation 36, $\frac{\gamma}{\lambda}$ is the shadow price of inequality in terms of public funds. If inequality is a negative externality (a public bad), $\frac{\gamma}{\lambda}$ will generally be large and positive.[86] The agent's weight $\kappa(n)$ in an inequality metric – for example the Gini coefficient discussed above – is negative at the bottom and positive at the top.

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 SWW 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 40 illlustrates that, when post-tax income inequality is an externality – or when the 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).

For clarity let us assume a linear homogeneous inequality externality ($\Gamma(\bar{\theta}) = \eta\bar{\theta}$) and quasi-linearity in consumption.[87] The optimal tax rate condition simplifies to:

$$\frac{t(n)}{1-t(n)} = \eta\kappa(n) + \eta\left(\frac{1}{E_L}+1\right)\frac{F(n)}{\alpha(n)} + \frac{t(n)_{orig}}{1-t(n)_{orig}}, \tag{42}$$

where we denote the local Pareto parameter $\frac{f(n)n}{1-F(n)}$ as $\alpha(n)$. This is what we employ in the main numerical simulations.

C.I. Equivalence of income rankings

In using the modified Gini in Equation 24, 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

[86] If we assume a linear inequality externality of the form $\Gamma(\theta) = \eta\theta$ then $\frac{\gamma}{\lambda} = \eta$ (see Equation 36).

[87] The resulting utility function is

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

Note that with quasi-linearity, $\int_n^\infty \kappa(p)dF(p)$ in (41) simplifies as $\int_n^\infty (2F(n)-1)\,dF(n) = F(n) - F(n)^2$.

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

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

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

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

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

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.[88] 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:

$$U(x, l, \bar{\theta}) = x - \frac{l^{\left(1+\frac{1}{E_c}\right)}}{\left(1 + \frac{1}{E_c}\right)} - \eta(\bar{\theta} - \bar{\theta}_{opt})^2 \tag{43}$$

The resulting analytical optimal tax rates are:

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

$$\frac{t(n)}{1-t(n)} = 2\eta(\bar{\theta} - \bar{\theta}_{opt}) \left[\kappa(n) + \frac{\zeta}{f(n)n}\int_n^{\infty} \kappa(p)f(p)dp\right] + \frac{t_{orig}}{1-t_{orig}} \tag{44}$$

Comparing these tax rates to Equation 41, 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.

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 $1 - H(z)$ agents above z under the assumption of no income effects. Thus it collects $[1 - H(z)]\, dz\partial\tau$ 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 given that second-order conditions hold, each decrease in one unit of post-tax income at z changes absolute post-tax income inequality by $\kappa(z)h(z)$ (from Equation 5).[89] The mechanical effect thus has a differing equality effect of $-\kappa(z_j)h(z_j)dz\partial\tau$ at each point j above z, where z_j is the income of the agent and $h(z_j)$ 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 $\overline{\kappa}(z)\,[1 - H(z)] = \int_{\{j:z_j>z\}} \kappa(z)h(z)dj$ and write that the mechanical effect changes income inequality by $d\bar{\theta}_M = -\overline{\kappa}(z)\,[1 - H(z)]\, dz\partial\tau$.[90]

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 $\partial z = -\epsilon(z)z\partial\tau/\,(1 - \tau(z))$. $\epsilon(z)$ is the elasticity of earnings z with respect to $1 - \tau(z)$. There are $h(z)dz$ individuals in the tax bracket who were taxed at $\tau(z)$ before the perturbation, so total revenue decreases by $-dz\partial\tau \cdot \epsilon(z)zh(z)\tau(z)/\,(1 - \tau(z))$. 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.[91]

[89] Note that $\kappa(z)$ is negative at the bottom of the distribution.

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

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

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 $d\bar{\theta}_B = -\kappa(z) \cdot dz\partial\tau \cdot \epsilon(z)zh(z)$.

The total revenue effects are:

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

The direct welfare effect through the individual income channels is $\int_j g_j dR dj$ 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 $d\bar{\theta} = d\bar{\theta}_M + d\bar{\theta}_B$:

$$\partial\bar{\theta} = dz\partial\tau \left(-\overline{\kappa}(z)\left[1 - H(z)\right] - \kappa(z)\epsilon(z)zh(z)\right) \tag{45}$$

In terms of utility, this affects individuals as $\int_j \frac{\partial U_j}{\partial\bar{\theta}} \cdot \partial\bar{\theta} \cdot dj$. We have that $\eta_i = MRS_{x\bar{\theta}} = -\frac{\partial U/\partial\bar{\theta}}{\partial U/\partial x_i}$, and thus the total welfare effect of the inequality change is $dI = \int_j (-g_j\eta_j) \cdot \partial\bar{\theta} \cdot dj = -\partial\bar{\theta} \cdot \int_j \eta_j g_j dj$.

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

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

Note that in the main text we denote $dI = dI_B + dI_M$ where dI_B and dI_M correspond to the welfare-weighted versions of $d\bar{\theta}_B$ and $d\bar{\theta}_M$ respectively. Thus, using the expressions for dR and dI, and the expression $\bar{G}(z)\left(1 - H(z)\right) = \int_{\{j:z_j \geq z\}} g_j dj / \int_j g_j dj$, we have:

$$\begin{aligned} dz\partial\tau \int_j g_j dj \left[1 - H(z) - h(z)\epsilon(z)z\frac{\tau(z)}{1 - \tau(z)}\right] - dz\partial\tau\bar{G}(z)\left(1 - H(z)\right)\int_j g_j dj \\ + \int_j \eta_j g_j dj \cdot \left[dz\partial\tau\left(\overline{\kappa}(z)\left[1 - H(z)\right] + \kappa(z)\epsilon(z)zh(z)\right)\right] = 0 \end{aligned}$$

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

$$\frac{\tau(z)}{1 - \tau(z)} = \eta \cdot \kappa(z) + \frac{1 - H(z)}{z \cdot h(z)} \frac{\left(1 - \bar{G}(z) + \eta\overline{\kappa}(z)\right)}{\epsilon(z)}, \tag{46}$$

where we have used the weighted average of the externality $\eta = \int_i g_i\eta_i di / \int_i g_i di$. By using the local Pareto parameter $\alpha(z) = \frac{z \cdot h(z)}{1 - H(z)}$ and $\Upsilon(z) = \eta\alpha(z)\epsilon(z)\kappa(z) + \eta\bar{\kappa}(z)$, we find the optimal marginal income tax rates as specified in Equation 7.

E Additional notes for Section 3

E.I. Numerical simulation specifications

Calibrating the model In the traditional optimal tax literature, tax rates are largely determined by three factors (Mankiw et al., 2009); (i) the shape of the wage-earning ability distribution, (ii) the social

welfare function, and (iii) labor or earnings elasticities.

The first factor is the shape of the wage-earning ability distribution $f(n)$,which is well-known to be important in such simulations (see e.g. Tuomala, 2016). Our main specification backs out the wage-earning ability distribution from the observed pre-tax labor income distribution. We use the DINA microfiles detailed in Piketty et al. (2018) to measure the U.S. pre-tax labor income distribution in 2019.[92] We show the local Pareto parameter for pre-tax income in the DINA files, $\frac{z \cdot h(z)}{1-H(z)}$, in Figure A1a.[93]

We then apply the NBER TAXSIM model to find marginal tax rates for any given tax unit in the DINA files.[94] In applying the TAXSIM model we add the number of dependents, the age of the taxpayer, and marital status for each representative tax unit in the DINA files to calculate corresponding real-world marginal tax rates. We then add an assumed 5% state tax, a 2.9% tax rate for Medicare, and a 2.4% sales tax rate, following Saez et al. (2012) and Hendren (2020).[95] We show a Kernel-smoothed version of the TAXSIM marginal tax rates in Figure A1b. Given these marginal tax rates and the empirical pre-tax income data, we assume individuals have correctly optimized according to the utility function in Equation 8 and back out the resulting wage-earning ability of each observation.[96] We then estimate the full post-tax wage-earning ability distribution from through a Kernel density estimator with a wage-earning ability bandwidth of \$5,000.[97] We assume a constant Pareto distribution for the last 0.5% of the distribution (above ∼\$600,000 in income), where data is sparse. The local Pareto parameter for this top region is set equal to the value immediately before the cut-off. This yields the final wage-earning ability distributions $f(n)$. In addition to this empirical wage-earning ability distribution, we also present two standard theoretical distributions in Appendix E.II.

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 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 elasticitieWhiles. We keep these homogeneous 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. While this choice is naturally

[92] As the Mirrlees model focuses on labor effort, we focus our analysis on labor income.

[93] This is calculated by taking a Kernel density estimator before backing out the underlying wage-earning ability distribution. The Kernel estimator bandwidth is \$80,000. This is the income distribution used in the inverse optimum exercise in Section 4.5.

[94] Described in Feenberg and Coutts (1993), accessed at https://taxsim.nber.org/ on April 20th 2023.

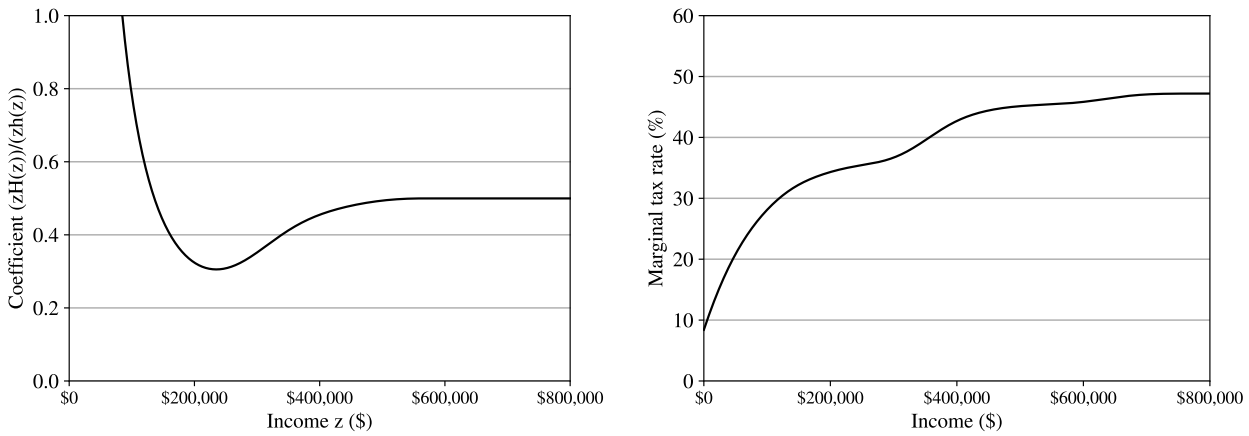
[95] We do not take into account state-based EITC benefits and other deductions; as discussed in Hendren (2020) this is unlikely to significantly affect results.

[96] We use that $(1 - \tau(z)) = -U_z/U_x$ from the individual's first-order condition. This together with Equation 8 indicates that;

$$n = \frac{x(z)^{1/(1+E_c)}}{(1-\tau(z))^{E_c/(1+E_c)}}$$

[97] This bandwidth corresponds to roughly \$80,000 in the income distribution.

Figure A1: (a) Hazard rate from 2019 U.S. income distribution, (b) 2019 marginal tax rates



Note: Left: Hazard ratio $(1 - H(z))/(zh(z))$ for the U.S. pre-tax labor income distribution in 2019. Right: 2019 U.S. marginal income tax rates for the simulations, taken from the NBER TAXSIM tool.

crucial for the optimal tax rates themselves, the numerical effects of introducing an inequality externality is relatively similar across different values of E_L (not shown).

The numerical simulations were performed in Python through an iterative process.[98] 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.

E.II. Theoretical ability distributions

We present Rawlsian optimal marginal income tax rates from two theoretical skill distributions in Figure A2, 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.[99] 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.

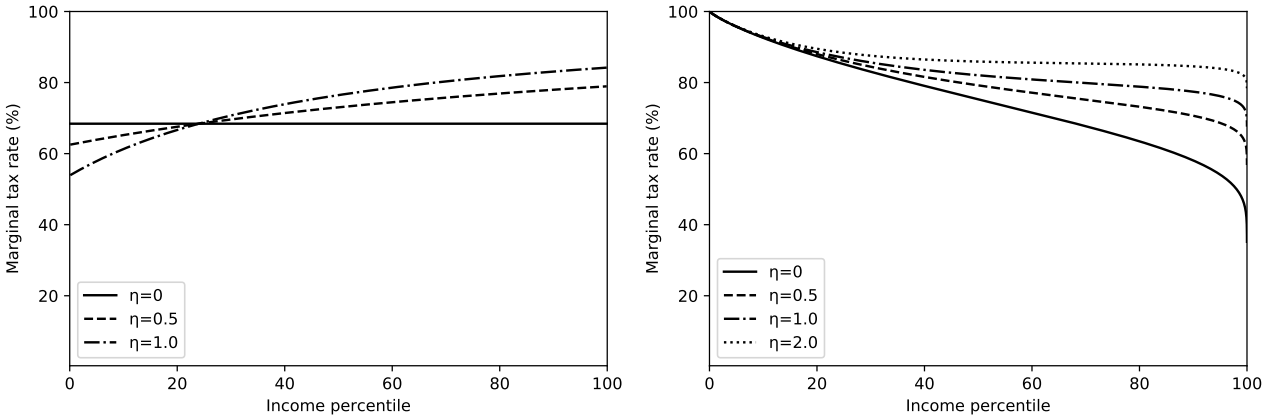
The Pareto case in Figure A2a illustrates the potentially positive effect of behavioral responses at the bottom. It is socially beneficial for low-income individuals to increase their incomes – so that inequality is reduced – which leads to a small income subsidy at the bottom as compared to the no-externality 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 $\eta = 2.0$

[98] 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. The method is further discussed in the Appendix of Mankiw et al. (2009). Note that the Rawlsian case can be solved analytically and thus do not require an iterative loop.

[99] 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 A2: Optimal Taxation with Inequality Externalities: Theoretical Ability Distributions



While

Note: Optimal marginal tax rates for various negative Gini-based post-tax income inequality externality magnitudes η_G. 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 $\eta_G = 1.0$. The solid line, $\eta = 0$, is the standard case of no inequality externality. See Table III for further explanation of the inequality externality magnitudes. The $\eta_G = 2.0$ case is excluded from the Pareto simulation because second-order conditions fail at the bottom. The elasticity of labor E_L is 0.3.

that increases to 67%. This illustrates the Pigouvian correction at the top, and is salient 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$^{\text{th}}$ percentile the marginal tax rate increases from 39% in the standard case to 79% when $\eta = 2.0$.

E.III. 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 A3.[100]

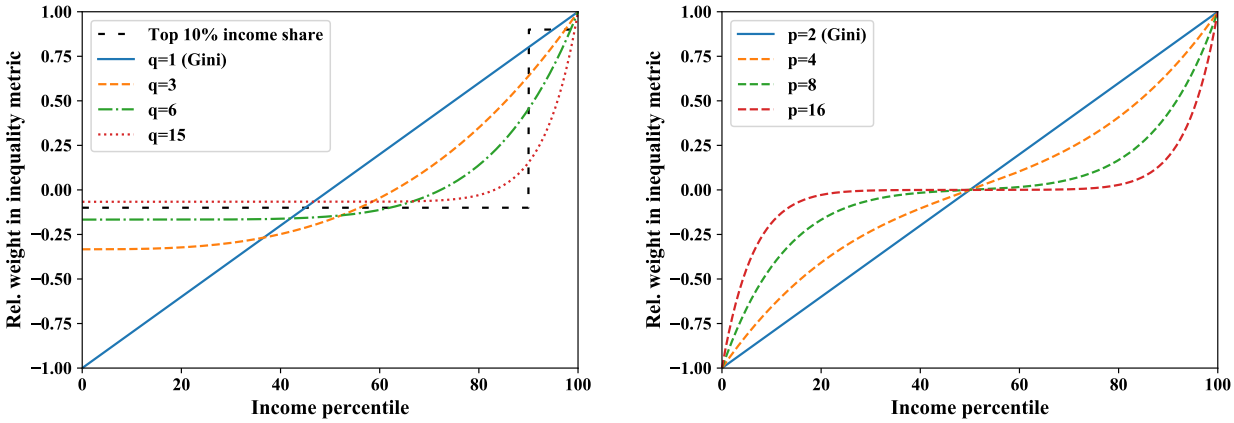
E.III.1 Approximating top income shares The first family of inequality metrics, also used in the main robustness test, has some of the properties of top income shares. It is,

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

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 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,

[100]The weights in Figure A3 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 A3: Weights for Families of Inequality Metrics



Note: Consumption weights for inequality metrics used in Appendix E.III. 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.

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 8 become,

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

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.[101]

In Figure A4 we replicate Figure IV for this inequality metric and $q = 4$. 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 IV, 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 incre ase 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

[101] 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.

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.

To further illustrate how increasing q has a large effect on top marginal tax rates, 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 A5. 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.[102]

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.

E.III.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 8 are,

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

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

In Figure A6 we show the effect of changing p on $\frac{t}{1-t}$ with the same methodology as in Figure A5. Increasing p again leads to larger effects towards the top of the distribution and relatively small

[102] 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.

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.

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 also 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.

E.IV. Heterogeneous inequality externalities

In the small-perturbation model we used heterogeneous inequality externalities η_i, where the policy-determining variable was the weighted average $\eta = \int_j \eta_j g_j dj / \int_j g_j dj$. This does not fully capture the potential complexity of the problem, however. The issue is that g_i often depends on the utility of agent i. To explain we go to the mechanism design framework, where this problem is immediate as the SWF is defined as $\int_{\underline{n}}^{\overline{n}} W(U(x(n), l(n), \bar{\theta}))dF(n)$. As we note in Appendix C, the most general optimal tax rates are,

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

where the shadow price of inequality $\frac{\gamma}{\lambda}$ is,

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

We can see that $\frac{\gamma}{\lambda}$ is potentially affected by a heterogeneous inequality externality. This is particularly clear in the denominator, where quasi-linearity and a homogeneous $\Gamma_{\bar{\theta}} = \eta$ would set the integral equal to zero whereas a heterogeneous $\Gamma_{\bar{\theta}} = \eta_i$ would not. There is a related issue in the solution for $\frac{t(n)}{1-t(n)}$ for the term containing $W_{U(p)}$. As utilities are differentially affected by the externality, $W_{U(p)}$ will

also be affected.[103] The solution is thus analytically complicated. We run numerical simulations with a heterogeneous $\Gamma(\bar{\theta})$ to explore the extent to which such heterogeneous inequality externalities could affect our results. We use $\Gamma(\bar{\theta}) = \eta \frac{n}{\int_{-\infty}^{\infty} n' f(n') dn'} \theta$. This implies that the externality is proportional to the wage-earning ability n of the agent, such that higher-earning agents are willing to pay more for a reduction in inequality, but still on average equal to $\eta\bar{\theta}$. The results of these simulations will be included shortly.

F Inverse-optimal social welfare weights

Re-arranging Equation 7, we can quickly find an expression for $\bar{G}(z)$:

$$\bar{G}(z) = (1 + \Upsilon(z)) - \frac{\tau(z)}{(1 - \tau(z))} \alpha(z) \epsilon(z) \tag{51}$$

Using that $\bar{G}(z) = \frac{1}{1-H(z)} \int_z^{\infty} g(j) dH(j)$, we can multiply by $1 - H(z)$ and take derivatives to find:

$$g(z) = \frac{1}{h(z)} \frac{d}{dz} \left[(1 - H(z)) (1 + \Upsilon(z)) - \frac{\tau(z)}{(1 - \tau(z))} z h(z) \epsilon(z) \right] \tag{52}$$

To calculate the inverse-optimal SWWs implied by the U.S. income tax schedule shown in Figure VI we used Equation 52, taking the numerical derivative of the bracketed expression. The pre-tax income distribution and tax specification are detailed in Appendix E.I and shown in Figure A1. We also assumed an elasticity of $\epsilon(z) = 0.3$ and a Gini post-tax income inequality externality. Finally we smoothed the resulting $g(z)$ to 99 quantile bins by taking the weighted mean of data inside each quantile boundary.[104]

In Figure A7a we show the implied SWWs under the same specification for a set of positive post-tax income inequality externalities.

In Figure A7b we show that a positive inequality externality could lead to the implied SWWs from the 2019 U.S. tax system being everywhere decreasing. We use the top-income inequality metric from Equation 10 with $q = 24$.[105] The mix of everywhere decreasing SWWs and a belief in top-end inequality as a positive externality could arguably describe conservative U.S. politics around the latest large-scale tax schedule reform in 1986 (TRA86).[106]

[103] As we noted in Appendix C, this is true for a homogeneous inequality externality as well. Here the issue is more serious, as the externality affects $W_{U(p)}$ not just through the level of U but also through the differential effect of $\Gamma_{\bar{\theta}}$.

[104] We ignored the top 0.5% to avoid our results being affected by the assumption of a constant Pareto distribution at the top as discussed in Appendix E.I.

[105] This indicates a positive inequality externality particularly focused on top incomes. Note that values for η_T are not comparable to values for η_G. Indeed, the magnitudes of η_T have been specifically chosen to illustrate this point.

[106] Statements from Ronald Reagan could indicate such a world view; (i) "Entrepreneurs and their small enterprises are responsible for almost all the economic growth in the United States", and (ii) "I believe we're meant to use wisely what is ours, make it grow, then help others to share and benefit from our success."

G 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.

G.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 $\partial z = -\epsilon(z) z \partial\tau / (1-\tau(z))$.[107] The behavioral response thus has an effect on the post-tax income inequality metric as $d\bar{\theta}_B = -\kappa(z) \cdot dz \partial\tau \cdot \epsilon(z) z h(z) / (1-\tau(z))$. 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 a homogeneous 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:

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

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

[107] 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 $1-\tau(z)$.

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

Which implies, after substituting $\alpha(z)=zh(z)/\left(1-H(z)\right)$,

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

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 SWWs, whether in utility or income.

Similarly, changing the analytical specification in Section C to a pre-tax inequality externality modifies Equation 41 to;

$$\frac{t}{1-t}=\frac{\gamma}{\lambda}\left[\frac{\kappa(n)u_x n}{V_l}\right]+\frac{t_i}{1-t_i}. \tag{53}$$

G.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,

$$\bar{\theta}_{\mathrm{U}}\left(\boldsymbol{z},H\right)=\int_{\underline{U}}^{\bar{U}}\kappa_U(U(z))U(z)dH'(U(z)), \tag{54}$$

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

$$U(x,l,\bar{\theta}_U)=x-v(l)-\eta_U\bar{\theta}_U. \tag{55}$$

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 $\int_{\underline{U}}^{\bar{U}} \kappa_U(U) dH'(U) = 0$ 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,

$$\bar{\theta}_{\mathrm{U},mech}(\boldsymbol{z}, F) = \int_{\underline{z}}^{\bar{z}} \kappa(z) x(z) dH(z), \tag{56}$$

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 SWWs $\bar{G}'(z) = \bar{G}(z) - \eta_U \cdot \bar{\kappa}(z)$ this can be simplified to the no-externality case without the need for $\alpha(z)$ or $\epsilon(z)$ in the modified SWWs.

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

$$\bar{\theta}'_{\mathrm{U},mech}(\boldsymbol{z}, F) = \int_{\underline{z}}^{\bar{z}} \kappa(z) u(x(z)) dH'(U). \tag{57}$$

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 $-\kappa(z)u_x(x(z))h'(z)$ (from Equation 56). The total decrease is thus $\int_{z'}^{\bar{z}} -u_x(x(z))\kappa(z)dz\partial\tau dH(z)$. 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 $\int_{\underline{z}}^{\bar{z}} u_x(x(z))\kappa(z)dR'dH(z)$. 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 $\int_{z'}^{\bar{z}} -u_x(x(z))\kappa(z)dz\partial\tau dH(z) = -\overline{\kappa}(z)\left[1 - h(z)\right]dz\partial\tau$, and the flat increase in income has no effect as $\int_{\underline{z}}^{\bar{z}} u_x(x(z))\kappa(z)dR'dH(z) = dR'\int_{\underline{z}}^{\bar{z}} \kappa(z)dH(z) = 0$. We

can thus write that the total utility inequality change from the perturbation is $d\bar{\theta}_U = -\overline{\kappa}(z)\left[1 - H(z)\right] dz\partial\tau$, which is equal to the mechanical effect from the standard externality case.

H 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.[108]

- 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.

[108] 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.

Table A1

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 7 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.

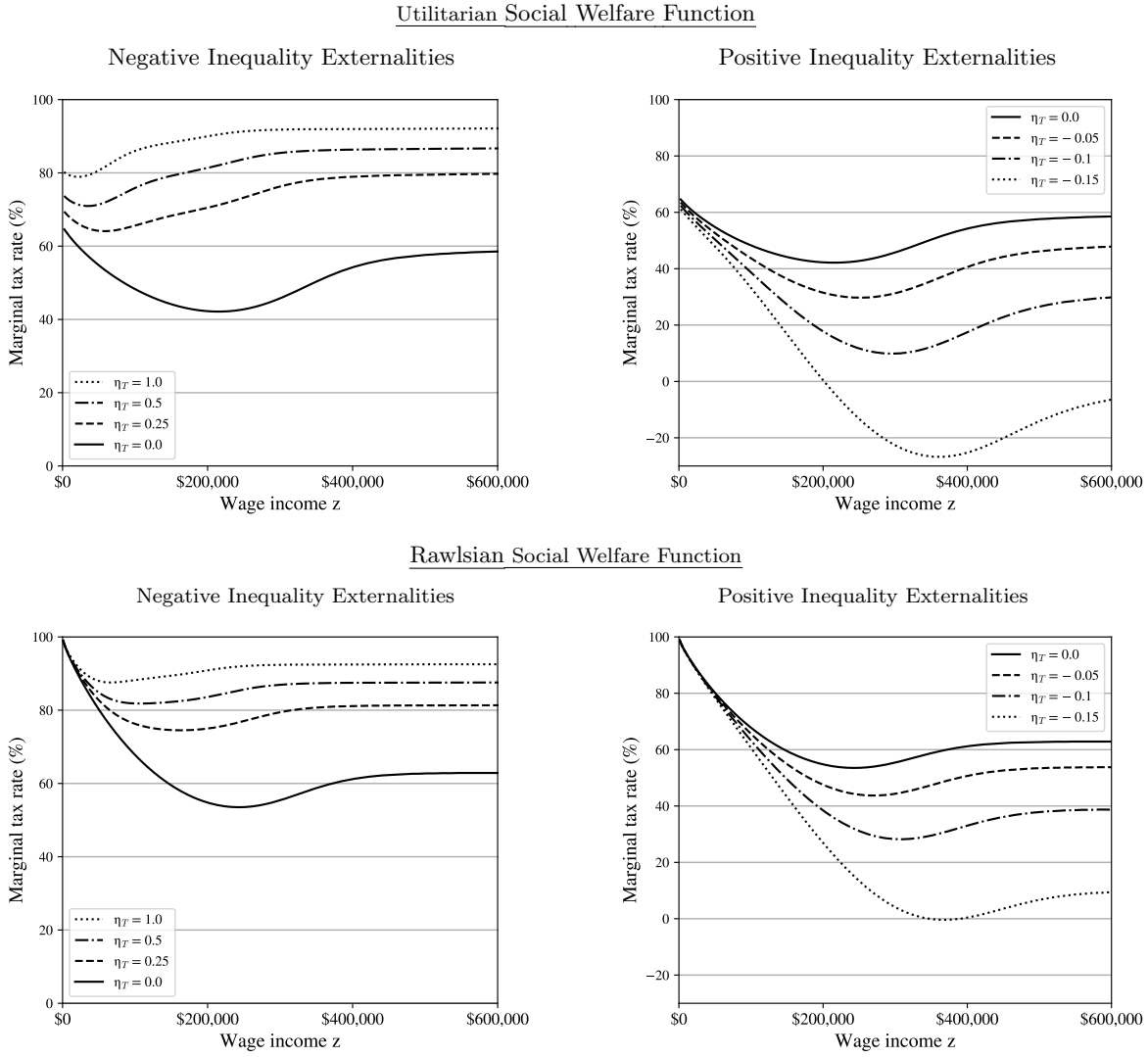
Table A2

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

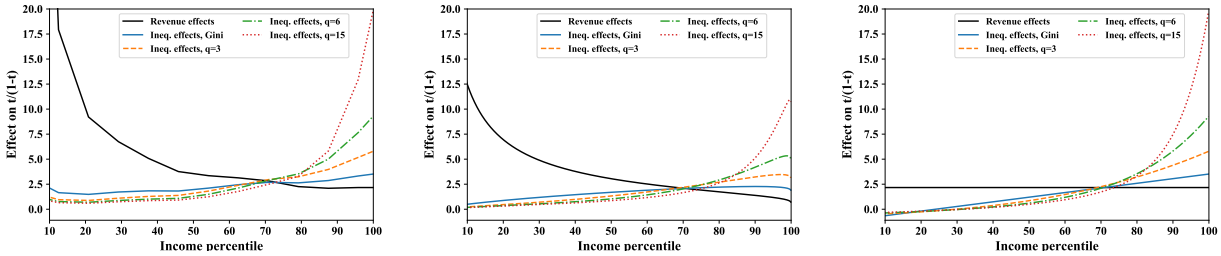
Note: Top marginal tax rates from Equation 7 with varying values of an inequality externality and elasticity of labor E_L. The social planner is Rawlsian. The inverse local Pareto parameter $1/\alpha$ is 0.5 in these calculations. The elasticity of labor E_L is 0.3 in the remainder of the paper. The standard no-externality case is in bold.

Figure A4: Optimal Marginal Income Tax Schedules with Top Share Inequality Externalities



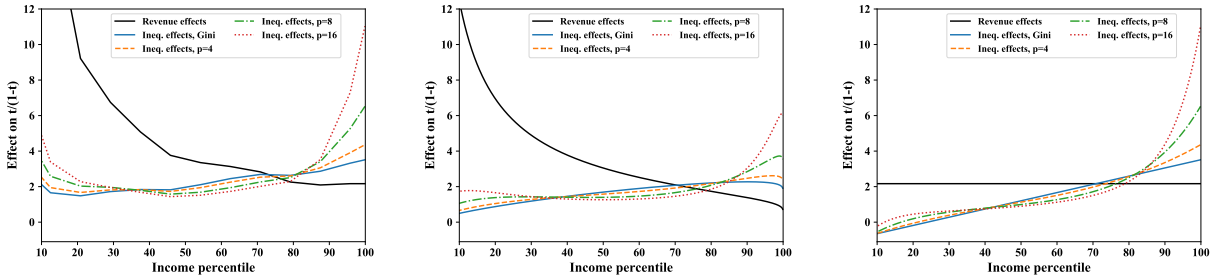
Notes: Optimal marginal tax rates for various top share-based inequality externalities with magnitudes η_T where inequality is either a negative externality (left) or a positive externality (right). The social planner is Utilitarian (above) and Rawlsian (below). The two cases converge when moving towards the top. Empirical estimates indicate $\eta_T \approx 0.5$. 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.

Figure A5: Effects on $\frac{t}{1-t}$: Top Income Share Externalities



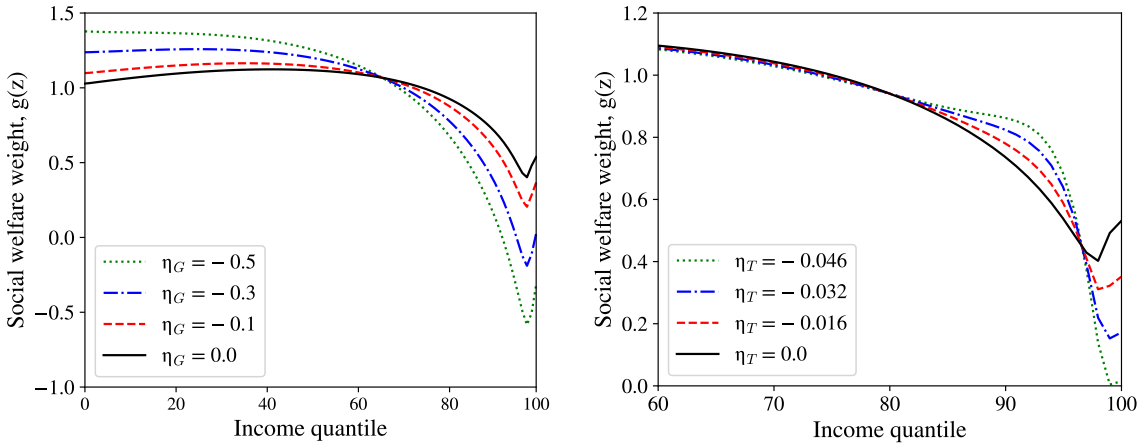
Note: Effects on $\frac{t}{1-t}$ for various negative inequality metrics $\int_0^\infty \left[(q+1)F(n)^q - 1\right] x(n)dF(n)$, $q \in \mathbb{N}$. The social planner is Rawlsian. The magnitude of the inequality externality is in each case calculated as the median value from the empirical inequality aversion estimates in Carlsson et al. (2005). This is done for comparability across inequality metrics. The productivity distribution is (a) the empirical income distribution, (b) a log-normal distribution with $\sigma = 0.39$ and $\mu_{log} = -1$, and (c) a Pareto distribution with $a = 2$. See Figure A3 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.

Figure A6: 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 income distribution, (b) a log-normal distribution with $\sigma = 0.39$ and $\mu_{log} = -1$, and (c) a Pareto distribution with $a = 2$. See Figure A3 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.

Figure A7: Implied $g(z)$ from the 2019 U.S. tax system across inequality externalities



Note: Left: Implied social welfare weights $g(z)$ from the 2019 U.S. tax system under various positive inequality externalities η_G (inequality is a social benefit). Right: Implied top social welfare weights $g(z)$ from the 2019 U.S. tax system under different positive top-share inequality externality magnitudes η_T (top inequality is a social benefit). Note that η_G is not comparable to η_T.