Discussion Paper Series

RIEB

Kobe University

DP2026-03

Optimal Redistribution with Institutional Reference Points

Hirofumi TAKIKAWA

January 22, 2026

Research Institute for Economics and Business Administration

Kobe University

2-1 Rokkodai, Nada, Kobe 657-8501 JAPAN

Optimal Redistribution with Institutional Reference Points

Hirofumi Takikawa*

Kobe University

January 22, 2026

Abstract

Standard optimal tax models typically ignore reference-dependent behavior induced by institutional thresholds. This paper incorporates loss aversion into a Mirrlees optimal income tax framework to analyze how such exogenous reference points, unlike social comparisons, alter optimal redistribution. I show that institutional loss aversion calls for globally higher marginal tax rates and a quantitatively large expansion of the lump-sum transfer. To accommodate behavioral bunching at the reference point, I employ an ironing approach and derive a modified optimal tax formula that remains valid in the presence of a mass point. Simulations calibrated to the U.S. economy imply that the optimal lump-sum transfer increases by 19–32% and yield welfare gains equivalent to 5.8–7.5% of consumption. These results are robust under both paternalistic and non-paternalistic welfare criteria.

Keywords: reference dependent preferences, optimal income taxation, redistribution

JEL codes: D03, H21, H24

*Email: takikawa@econ.kobe-u.ac.jp Address: Kobe University, Graduate School of Economics, 2-1 Rokkodai-cho Nada-ku, Kobe, 6578501, Japan. I am a Junior Research Fellow at the Research Institute for Economics and Business, Kobe University. I would like to thank Alfons J. Weichenrieder for his supervision and continuous support during my PhD studies, and Michael Neugart, Yukihiro Nishimura, Shuichi Tsugawa, Hitoshi Tsujiyama, and the seminar and conference participants for their valuable and helpful comments. All errors remain my own. This work is supported by JSPS KAKENHI Grant Number 25K16667.

1. Introduction

Standard optimal tax models assume that individuals derive utility solely from absolute levels of consumption and labor supply. In contrast, a vast body of evidence in behavioral economics demonstrates that individuals evaluate outcomes relative to a reference point, exhibiting loss aversion when outcomes fall below that reference point (Tversky and Kahneman, 1991; Köszegi and Rabin, 2006). While theoretical literature often treats reference points as determined by the status quo or expectations, in reality, various institutional structures distinct from the income tax schedule frequently create explicit, salient income targets that serve as exogenous reference points. Prominent examples of such institutionally determined reference points include repayment thresholds in income-contingent student loan programs in the UK and Australia (Chapman and Leigh, 2009; Britton and Gruber, 2020), and earnings thresholds in social security systems (Gelber et al., 2020; Seibold, 2021). Similarly, in Japan, the widely publicized "1.03 million yen wall" for spousal deduction creates a strong psychological anchor beyond its pure financial incentive (Fukai and Kondo, 2023).

These institutional features create a conflict between standard optimal tax theory, which typically prescribes smooth marginal tax rates, and the behavioral reality where individuals bunch at specific income levels due to loss aversion. Despite the prevalence of such behavior, optimal tax theory has yet to provide a comprehensive framework for designing tax schedules when labor supply is driven by these exogenous targets. Unlike standard bunching driven by tax kinks, this behavior arises from preferences themselves. However, the normative implications of such preference-driven bunching triggered by institutional thresholds have not been sufficiently addressed in the existing literature. Importantly, I abstract from socially endogenous reference points formed through peer comparisons or network interactions; these mechanisms generate interpersonal externalities and are conceptually distinct from the institutional channel studied here.

To fill this gap, this paper incorporates reference-dependent preferences into the canonical Mirrlees (1971) optimal income tax framework. The key behavioral mechanism is that loss aversion around an institutional threshold generates bunching responses (Kleven, 2016). Individuals earning below the reference point are disproportionately motivated to increase labor supply to avoid the psychological loss of falling short of the threshold. This asymmetry fundamentally alters the elasticity of taxable income. Specifically, agents striving to reach the reference point become less responsive to tax incentives, creating locally distorted responses to taxation and leading to a mass point at the reference income level. I show that institutional

loss aversion calls for globally higher marginal tax rates and a quantitatively large expansion of the lump-sum transfer.

Incorporating this behavioral bunching into the optimal tax problem requires a departure from the standard "first-order approach", which relies solely on the local first-order condition for incentive compatibility while neglecting the second-order condition. This widespread approach assumes the strict monotonicity of income with respect to ability to identify the optimal schedule, typically verifying this assumption ex post. However, the presence of a mass of agents at the reference point violates this strict monotonicity, rendering the first-order approach inapplicable. To rigorously address this technical challenge, I employ the "ironing" technique, adapting the formulation of Brett and Weymark (2017). Instead of proceeding under the assumption of strict monotonicity, this method allows me to treat the agents over the bunching interval as a single mass in the government's optimization problem. By explicitly solving for the optimal allocation without neglecting the second-order condition, I derive a modified optimal tax formula that remains valid even when behavioral loss aversion endogenously generates bunching.

My analysis yields three main results. First, I derive an optimal tax formula that explicitly accounts for loss aversion. Extending the standard ABC formula (Diamond, 1998), I show that reference dependence introduces an additional term that universally increases optimal marginal tax rates. The formula reveals that the government should impose higher marginal tax rates, particularly on low-income earners. The intuition behind this result rests on two mechanisms. On one hand, reference dependence reduces the elasticity of labor supply globally, with a particularly strong effect in the loss domain. This allows the government to tax income with lower efficiency costs, consistent with the standard efficiency principle of taxing inelastic bases. On the other hand, and more importantly, higher tax rates serve to correct the excessive labor supply incentives driven by reference-dependent preferences. By taxing away the gains from intensified labor supply, especially below the reference point, the government can mitigate the behavioral distortion that drives agents to overwork solely to avoid feeling a sense of loss or to seek referencing gains.

Second, using numerical simulations, I quantify the magnitude of these behavioral factors. While the institutional examples motivating the reference points are drawn from various jurisdictions, I calibrate the model using the standard U.S. wage distribution from Mankiw et al. (2009) and elasticity parameters from Chetty (2012). Using this widely accepted benchmark allows for a direct comparison with the existing optimal tax literature and isolates the impact of the behavioral mechanism. I find that higher reference points markedly alter

the optimal tax structure, shifting the bottom of the U-shaped marginal tax rate curve to higher income levels. Crucially, the model suggests that to compensate for the inefficiency of loss-driven overwork, the lump-sum transfer must increase by approximately 19–32%. This substantial increase reflects the high social value of correcting this severe behavioral distortion. Furthermore, accounting for these reference points yields non-trivial welfare gains of approximately 5.8–7.5% compared to the standard benchmark, highlighting the significant value of aligning tax design with behavioral motivations. Finally, to ensure that these results are not driven solely by the discontinuity inherent in the piecewise linear utility, I examine a smoothed specification in the Appendix. I confirm that the key findings, globally higher marginal tax rates and expanded redistribution, remain robust to this generalization.

Third, to address the normative ambiguity inherent in behavioral public finance (Bernheim and Rangel, 2009; Aronsson and Johansson-Stenman, 2018), I conduct a comprehensive welfare analysis under two distinct criteria: a paternalistic criterion, which treats loss aversion as a behavioral bias to be corrected, and a non-paternalistic one, which fully respects individuals' reference-dependent preferences. I derive the optimal tax formulas and perform numerical simulations for both scenarios. Notably, I find that the qualitative prescriptions, higher marginal tax rates and increased redistribution, are robust across these welfare criteria. Although the theoretical rationales differ, as the paternalistic government seeks to correct the behavioral distortion of overworking while the non-paternalistic government aims to redistribute from agents with high subjective utility, the resulting optimal tax schedules are remarkably similar. This suggests that the case for reforming the tax system in the presence of reference points remains strong regardless of the government's specific normative stance.

Related literature. This paper contributes to three strands of literature. First, it adds to the growing field of behavioral public finance. Farhi and Gabaix (2020) provide a unified theory of optimal taxation with behavioral agents, deriving generalized formulas for Ramsey, Pigou, and Mirrlees problems. Beyond this general framework, recent studies have explored optimal taxation under deviations from standard rationality, such as tax inattention (Chetty et al., 2009), sin taxes (Allcott et al., 2019), and present bias (Lockwood, 2020). Specifically regarding prospect theory, Kanbur et al. (2008) analyze optimal taxation under income uncertainty. However, while their model employs the moral hazard framework to study insurance against risk, I incorporate loss aversion into the adverse selection framework (Mirrlees, 1971) to analyze labor supply distortions.

Furthermore, within this field, my work complements studies on relative consumption and social status, such as Aronsson and Johansson-Stenman (2008, 2018, 2024) and Kanbur

and Tuomala (2013). However, a key distinction is that their work focuses on socially driven other-regarding preferences, e.g., status concerns or inequality aversion, where utility depends on the consumption of others. In these models, the reference point is endogenous to the aggregate economy, generating negative externalities. In contrast, this paper focuses on non-social reference dependence, where the reference point is determined exogenously by institutional thresholds. This distinction is crucial because loss aversion around a fixed target generates intra-personal behavioral distortions driven by discontinuous marginal incentives, distinct from the inter-personal externalities arising from social comparisons.

Second, this paper connects the literature on bunching with behavioral reference points. A large empirical literature estimates bunching responses to discontinuities in tax schedules (Saez, 2010; Kleven and Waseem, 2013). More recently, Rees-Jones (2018) empirically documented that taxpayers exhibit loss aversion around tax-balance-due thresholds, leading to bunching in reported income. My paper complements these empirical findings by analyzing the normative implications of such reference-dependent behavior. While this empirical literature typically takes the tax schedule as given to estimate elasticities, normative analysis on how to design the tax schedule itself in the presence of bunching remains scarce. I provide a theoretical framework to determine the optimal tax schedule when such behavioral bunching arises from institutional thresholds (Seibold, 2021).

Finally, from a technical perspective, my analysis draws on the optimal tax literature dealing with the violation of the second-order condition for incentive compatibility. While the standard "first-order approach" assumes strictly monotonic income schedules, seminal contributions by Brito and Oakland (1977), Lollivier and Rochet (1983), Weymark (1986), and Ebert (1992) explicitly characterize optimal schedules when this monotonicity constraint binds and agents bunch. Furthermore, Guesnerie and Laffont (1984) provided a comprehensive analysis of principal-agent problems subject to such constraints, developing the method of "blocked intervals" which serves as the prototype for the modern ironing technique. Typically, in these models, bunching arises due to irregularities in the skill distribution or non-convexities in the feasible set. In this paper, I apply this rigorous theoretical apparatus to a distinct source of non-monotonicity: reference-dependent preferences. Specifically, I employ the "ironing" technique, adapting the formulation by Brett and Weymark (2017), to characterize the optimal tax schedule where behavioral loss aversion endogenously generates a violation of the second-order condition.

Layout. The remainder of the paper is organized as follows. Section 2 presents the model. Section 3 characterizes the optimal tax schedules under both paternalistic and non-

paternalistic social welfare criteria. Section 4 presents numerical simulations to quantify the theoretical findings. Section 5 concludes.

2. Model

2.1 Environment

Consider an economy populated by a continuum of agents with heterogeneous ability $\theta \in [\underline{\theta}, \overline{\theta}]$, which is private information and follows a continuously differentiable cumulative distribution function $F(\theta)$ with density $f(\theta)$. The labor market is perfectly competitive, and ability θ represents the agent's wage rate per unit of labor supply.

Agents' preferences are represented by a reference-dependent utility function $u(c, n; r)$, following Tversky and Kahneman (1991). Utility depends on consumption c, labor supply n, and an income reference point r. Pre-tax income is given by $y = \theta n$. The reference point r is assumed to be exogenous and common across all agents. This reference point can be interpreted as an institutional threshold, such as the student loan repayment trigger in the UK or the social security earnings limit in Japan. These thresholds are governed by specific administrative rules or legislation distinct from the general income tax schedule. Therefore, I focus on optimal tax design, treating these existing institutional thresholds as exogenous constraints. Due to loss aversion, agents have a stronger marginal incentive to increase labor supply when their income is below r than when it is above r.

Specifically, I assume the following quasi-linear utility function containing a gain-loss utility term $\mu(\cdot)$ that satisfies the standard properties outlined by Köszegi and Rabin (2006):[1]

$$u(c, n; r) = c - v(n) + \mu(z), \tag{1}$$

where $z \equiv y - r$ denotes the deviation of income from the reference point. The term $v(n) = \chi \frac{n^{1+1/\varepsilon}}{1+1/\varepsilon}$ denotes the increasing and convex disutility of labor, where $\varepsilon > 0$ represents the constant Frisch elasticity of labor supply and $\chi > 0$ is a scale parameter representing the intensity of labor disutility..

For the gain-loss utility $\mu(z)$, I adopt a piecewise linear specification, as is common in the literature. This linear form, defined in equation (2), ensures analytical tractability and

[1]Köszegi and Rabin (2006), following Bowman et al. (1999), impose assumptions (A0-A4) regarding differentiability, increasingness, loss aversion, and diminishing sensitivity on the gain-loss utility $\mu(\cdot)$.

allows me to isolate the role of loss aversion while abstracting from diminishing sensitivity. Consequently, the utility function exhibits a kink at $z = 0$, which can induce bunching behavior:

$$\mu(z) = \begin{cases} \eta z & \text{if} \quad z \geq 0 \\ \eta \lambda z & \text{if} \quad z < 0 \end{cases} \tag{2}$$

where $\eta \geq 0$ represents the weight on gain-loss utility, and $\lambda > 1$ is the loss aversion coefficient. In the Appendix, I demonstrate that the main results are robust, both qualitatively and quantitatively, to non-linear specifications of the gain-loss utility incorporating diminishing sensitivity.

Furthermore, the utility function (1) satisfies the Spence-Mirrlees single-crossing condition. That is, the marginal rate of substitution between consumption and income, given by $\text{MRS}_{cy}(\theta) \equiv -u_y/u_c = v'(y/\theta)/\theta - \mu'(z)$, is strictly decreasing in ability θ. This implies that higher-ability agents require less compensation in consumption to earn an additional unit of income. Since the piecewise linear specification implies $\mu''(\cdot) = 0$ except at the kink, reference dependence does not violate the single-crossing property.[2]

2.2 Income choice with reference dependence

The government imposes a nonlinear income tax schedule $T(y)$ on income y. I assume that $T(\cdot)$ is continuously differentiable but do not impose a specific functional form. Positive values of $T(y)$ denote tax payments, while negative values represent income transfers, in particular, $-T(0)$ corresponds to the lump-sum transfer to the unemployed.

Given the tax schedule $T(\cdot)$, an agent with ability θ chooses income y to maximize utility subject to the individual budget constraint $c = y - T(y)$. Substituting this constraint and the labor-income relationship $n = y/\theta$ into the utility function, the individual optimization problem is given by:

$$V(\theta) = \max_{y \geq 0} \; u\left(y - T(y), \frac{y}{\theta}; r\right), \tag{3}$$

where $V(\theta)$ denotes the indirect utility of an agent with ability θ.

The agent's optimal income $y(\theta)$ is characterized by the first-order condition of the max-

[2]Under diminishing sensitivity (where $\mu''(\cdot) < 0$ for gains and $\mu''(\cdot) > 0$ for losses), the single-crossing condition is not guaranteed, particularly if the marginal gain-loss utility approaches infinity at the reference point (e.g., $\lim_{z \to 0} \mu'(z) = \infty$). Thus, incorporating diminishing sensitivity requires regularity conditions on $\mu(\cdot)$.

imization problem (3):

$$v'\left(\frac{y(\theta)}{\theta}\right) = \Big[1 - T'(y(\theta)) + \mu'(z(\theta))\Big]\theta \tag{4}$$

where $\mu'(z(\theta))$ represents the marginal gain-loss utility with $z(\theta) \equiv y(\theta) - r$, taking the following piecewise constant values:

$$\mu'(z(\theta)) = \begin{cases} \eta & \text{if} \quad z(\theta) \geq 0 \\ \eta\lambda & \text{if} \quad z(\theta) < 0 \end{cases} \tag{5}$$

The optimality condition (4) equates the marginal disutility of labor supply on the left-hand side with the marginal return to labor supply on the right-hand side. With reference dependence where $\eta > 0$, the marginal return to labor is strictly higher than that in the standard model where $\eta = 0$. This implies that, for a given marginal tax rate, agents derive a higher marginal benefit from income, inducing them to accept a higher marginal disutility of labor. Furthermore, in the loss domain where $z < 0$, the marginal return is further amplified by the loss aversion coefficient λ. This reflects the stronger incentive to increase earnings to avoid losses. Consequently, agents bunch at the reference point r due to the kink in the utility function; that is, income remains constant at r over a range of abilities, implying that labor supply $n(\theta) = r/\theta$ decreases with ability within this interval.

Given the iso-elastic functional form of $v(n)$, the optimal income schedule $y(\theta)$ is given by:

$$y(\theta) = \begin{cases} \theta^{1+\varepsilon}\left(\dfrac{1 - T'(y(\theta)) + \eta}{\chi}\right)^{\varepsilon} & \text{if} \quad \theta > \theta^h \\ r & \text{if} \quad \theta \in [\theta^\ell, \theta^h] \\ \theta^{1+\varepsilon}\left(\dfrac{1 - T'(y(\theta)) + \eta\lambda}{\chi}\right)^{\varepsilon} & \text{if} \quad \theta < \theta^\ell \end{cases} \tag{6}$$

where the ability thresholds are defined as $\theta^h \equiv \left(\frac{r\chi^\varepsilon}{[1-T'(r)+\eta]^\varepsilon}\right)^{\frac{1}{1+\varepsilon}}$ and $\theta^\ell \equiv \left(\frac{r\chi^\varepsilon}{[1-T'(r)+\eta\lambda]^\varepsilon}\right)^{\frac{1}{1+\varepsilon}}$. The optimal income schedule exhibits three distinct regimes. First, high-ability agents with $\theta > \theta^h$ operate in the gain domain ($z > 0$). Second, intermediate-ability agents with $\theta \in [\theta^\ell, \theta^h]$ bunch at the reference point r. Third, low-ability agents with $\theta < \theta^\ell$ operate in the loss domain ($z < 0$). It is important to note that agents in this loss domain face even stronger work incentives due to loss aversion compared to high-ability agents in the gain domain.

Crucially, these ability thresholds θ^ℓ and θ^h are endogenously determined by the agents' optimization given the tax schedule $T(y)$ and the reference point r. Consistent with the assumption of asymmetric information, the government cannot observe ability θ directly but can only design the tax schedule based on observable income, which in turn induces the sorting of types characterized by these thresholds.

A key implication of reference dependence is that the elasticity of taxable income with respect to the net-of-tax rate $1 - T'(y)$, denoted by $\hat{\varepsilon}$, is no longer constant.

$$\hat{\varepsilon} \equiv \frac{\partial y(\theta)}{\partial[1 - T'(y(\theta))]} \frac{1 - T'(y(\theta))}{y(\theta)} = \begin{cases} \varepsilon \left(\dfrac{1 - T'(y(\theta))}{1 - T'(y(\theta)) + \eta} \right) & \text{if} \quad \theta > \theta^h \\ 0 & \text{if} \quad \theta \in [\theta^\ell, \theta^h] \\ \varepsilon \left(\dfrac{1 - T'(y(\theta))}{1 - T'(y(\theta)) + \eta\lambda} \right) & \text{if} \quad \theta < \theta^\ell \end{cases} \tag{7}$$

In the absence of reference dependence where $\eta = 0$, this elasticity is constant at ε, consistent with the standard Mirrlees model. However, with reference dependence where $\eta > 0$, the effective elasticity becomes smaller than the structural parameter ε, implying that taxable income becomes less elastic to the marginal tax rate. For income below the reference point, the elasticity is further reduced due to loss aversion where $\lambda > 1$. Since agents are less responsive to tax changes, i.e., the tax base is less elastic, the government can impose higher tax rates with lower efficiency costs. This result aligns with the standard Ramsey inverse-elasticity rule, which justifies higher tax rates on less elastic tax bases to minimize distortions.

2.3 Incentive compatibility

In addition to the optimality condition for income, the allocation must satisfy the following incentive compatibility constraint for all ability levels:

$$u\left(c(\theta), \frac{y(\theta)}{\theta}; r\right) \geq u\left(c(\theta'), \frac{y(\theta')}{\theta}; r\right) \qquad \forall \theta, \theta' \in [\underline{\theta}, \overline{\theta}] \tag{8}$$

Given that the utility function (1) satisfies the single-crossing condition, the following first-order and second-order conditions for local incentive compatibility, (9) and (10), are sufficient to guarantee global incentive compatibility:

$$V'(\theta) = v'\left(\frac{y(\theta)}{\theta}\right) \frac{y(\theta)}{\theta^2} \tag{9}$$

$$y'(\theta) \geq 0 \tag{10}$$

The first-order condition (9) ensures that the indirect utility $V(\theta)$ increases sufficiently with ability θ to prevent agents from mimicking the behavior of lower abilities. The second-order condition (10) requires income to be nondecreasing in ability θ.

The standard "first-order approach" typically relies solely on equation (9), assuming that the resulting income schedule is strictly monotonic. This monotonicity is usually verified ex-post in numerical simulations. However, reference-dependent preferences introduce a kink in the utility function at the reference income level. This non-differentiability creates a range of abilities where the optimality condition implies constant income $y(\theta) = r$. Since the income schedule is constant rather than strictly increasing over this interval, the standard method of inverting the relationship between ability and income breaks down. Consequently, the standard approach requires modification to handle this interval.

To rigorously address this issue, I consider the second-order condition for incentive compatibility, drawing on the "ironing" technique (Brett and Weymark, 2017). Typically, this technique is applied when the standard "first-order approach" yields a decreasing income schedule, which requires the region to be ironed out. In my context, the violation arises not from a decrease, but from the strict constancy of income due to the kink. However, the fundamental mathematical principle of treating the agents as a single mass remains applicable. Therefore, following the formulation of Brett and Weymark (2017), I adapt this technique to treat the agents bunched at the reference point as a single aggregate entity in the optimization problem.

3. Optimal tax schedule

In this section, I characterize optimal income tax schedules. Following the normative analysis in behavioral public finance (e.g., Aronsson and Johansson-Stenman, 2018; Lockwood, 2020), I consider two distinct social welfare criteria: a paternalistic criterion and a non-paternalistic one. A paternalistic government treats the gain-loss utility $\mu(\cdot)$ derived from reference dependence as a behavioral bias and excludes it from the social welfare function, whereas a non-paternalistic government fully respects individuals' preferences, including the gain-loss utility $\mu(\cdot)$.

I structure the analysis by first focusing on the paternalistic case, which serves as a baseline for correcting behavioral distortions. I begin by solving the relaxed problem, defined as

the maximization problem subject only to the first-order condition for incentive compatibility, thereby ignoring the second-order condition. As discussed, the kinked utility leads to a violation of standard monotonicity assumptions. Nevertheless, solving this relaxed problem provides a necessary benchmark formula. Subsequently, I modify the optimal tax formula to account for the bunching induced by reference dependence. Finally, I extend the analysis to the non-paternalistic case and compare the resulting optimal tax schedules.

3.1 The relaxed problem with a paternalistic government

This subsection focuses on the relaxed problem, where the government maximizes social welfare subject solely to the first-order condition for incentive compatibility, ignoring the second-order condition. By the taxation principle (Hammond, 1979; Guesnerie, 1995), characterizing a nonlinear income tax schedule is equivalent to finding an allocation of income and consumption that satisfies incentive compatibility. In the context of the relaxed problem, I restrict this requirement to the local first-order condition (9). Thus, instead of choosing the tax schedule directly, the government chooses the allocation $y(\theta)$ and $c(\theta)$ to maximize social welfare subject only to the incentive constraint (9) and the resource constraint.

Under the paternalistic criterion, the government maximizes the following social welfare function:

$$\int_{\underline{\theta}}^{\overline{\theta}} G(U(\theta))f(\theta)d\theta \tag{11}$$

where $G(\cdot)$ is an increasing and concave transformation of utility, i.e., $G'(\cdot) > 0$ and $G''(\cdot) < 0$. The argument $U(\theta) \equiv V(\theta) - \mu(z(\theta))$ represents the agent's true utility evaluated by the paternalistic government, which treats the gain-loss utility $\mu(\cdot)$ as a behavioral bias to be disregarded.

The government faces the resource constraint:

$$\int_{\underline{\theta}}^{\overline{\theta}} \Big(y(\theta) - c(\theta)\Big)f(\theta)d\theta \geq E \tag{12}$$

where $E \geq 0$ denotes an exogenous revenue requirement. To simplify the numerical simulations, I assume $E = 0$, implying that all tax revenue is redistributed.

I solve the optimization problem using a standard Hamiltonian approach, where $V(\theta)$ is the state variable and $y(\theta)$ is the control variable. The first-order condition (9) serves as the

law of motion for the state variable. Let γ denote the multiplier on the resource constraint (12), and let $\delta(\theta)$ denote the co-state variable associated with the incentive constraint (9).

The optimal income tax formula is characterized in Proposition 1, which extends the well-known ABC formula. In the following formula, $g_p(\theta) \equiv G'(U(\theta))/\gamma$ denotes the marginal social welfare weight of an agent with ability θ, normalized such that the average weight is 1, and the subscript p indicates the paternalistic case. Using the transversality conditions $\delta(\underline{\theta}) = \delta(\overline{\theta}) = 0$, the multiplier is determined as $\gamma = \int_{\underline{\theta}}^{\overline{\theta}} G'(U(\theta))f(\theta)d\theta$, representing the marginal value of public funds.

Proposition 1. *The optimal marginal tax rate for the relaxed problem at an arbitrary income level $y = y(\theta)$ satisfies:*

$$\frac{T_p'(y) - g_p(\theta)\mu'(z(\theta))}{1 - T_p'(y) + \mu'(z(\theta))} = \underbrace{\frac{1+\varepsilon}{\varepsilon}}_{A} \underbrace{\frac{1-F(\theta)}{\theta f(\theta)}}_{B} \underbrace{\frac{\int_{\theta}^{\overline{\theta}} \Big[1 - g_p(\theta')\Big] f(\theta')d\theta'}{1-F(\theta)}}_{C_p} \tag{13}$$

where $z(\theta) \equiv y(\theta) - r$ denotes the income deviation from the reference point.

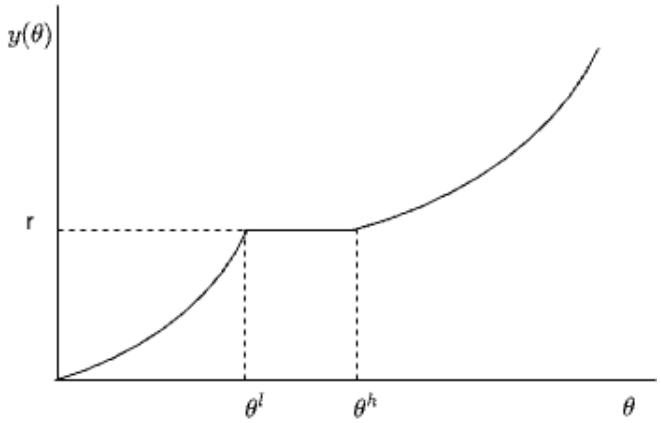
The right-hand side of the optimal tax formula (13) corresponds to the standard ABC formula (Diamond, 1998; Saez, 2001). A represents the efficiency and elasticity term, B captures the shape of the skill distribution and thickness of the right tail, and C_p denotes the desire for redistribution. However, the left-hand side contains additional terms involving $\mu'(z(\theta))$, which capture the effects of reference dependence.

Rearranging equation (13) for the marginal tax rate $T_p'(y)$ provides clearer intuition:

$$T_p'(y) = \frac{[1 + \mu'(z(\theta))]ABC_p + g_p(\theta)\mu'(z(\theta))}{1 + ABC_p} \tag{14}$$

This form highlights two distinct channels through which reference dependence affects optimal taxation. First, the term $1 + \mu'(z(\theta))$ in the numerator reflects the inverse-elasticity rule. Loss aversion increases the marginal return to labor, making agents' labor supply less sensitive to tax changes, rendering it less elastic. Since the efficiency cost of taxation is lower for less elastic tax bases, the government can optimally impose higher marginal tax rates. Second, the term $g_p(\theta)\mu'(z(\theta))$ expresses a Pigouvian motivation to correct the distortion caused by behavioral bias. A paternalistic government views the reference-dependent motivation as a bias that causes agents to overwork relative to the bias-free optimum. To correct this distortion and align labor supply with the true utility function, the government

Figure 1: Optimal income schedule as a function of ability



Note: The solid line illustrates the optimal income schedule according to equation (6). It is divided into three regimes: $\theta > \theta^h$, $\theta \in [\theta^\ell, \theta^h]$, and $\theta < \theta^\ell$. Under appropriate assumptions about income tax, optimal income is strictly increasing for $\theta > \theta^h$ and $\theta < \theta^\ell$, but constant at the reference point r for $\theta \in [\theta^\ell, \theta^h]$.

imposes an additional tax wedge. Thus, if I assume no reference dependence where $\eta = 0$, the optimal tax formula is consistent with the standard form in the literature.

3.2 Optimal tax formula with bunching under paternalism

As discussed in the previous section, reference-dependent preferences introduce a kink in the utility function at $y = r$, which creates a discontinuity in the marginal return to labor. This violates the strict monotonicity assumption required for the standard first-order approach. Consequently, optimal income becomes constant at the reference point over a subinterval of abilities, leading to bunching. While the relaxed optimal tax formula (13) remains valid for agents outside this bunching region, it requires modification for the bunched agents.

As shown in Figure 1, the optimal income schedule consists of three regimes: the gain domain for $\theta > \theta^h$, the bunching regime for $\theta \in [\theta^\ell, \theta^h]$, and the loss domain for $\theta < \theta^\ell$. The regions outside the bunching region are captured by the unified expression using $\mu'(z(\theta))$ derived in Proposition 1. Therefore, I focus here on specifying the optimal tax formula for the bunching region where $y(\theta) = r$.

Following the approach of Brett and Weymark (2017), agents bunched at the reference point r are treated as a single point mass in the optimization problem. Since income is constant at r for all $\theta \in [\theta^\ell, \theta^h]$, the standard optimality condition reflects the discontinuity

in marginal returns: the agent's marginal rate of substitution must lie between the higher marginal return from avoiding losses and the lower marginal return from accruing gains. Specifically, by applying the concept of subdifferentiation, the marginal gain-loss utility at the kink corresponds to the subdifferential set $[\eta, \eta\lambda]$.

To express the optimal tax rate $T_p'(r)$ in a structure comparable to the standard ABC formula, I define a representative ability $\theta^m \in [\theta^\ell, \theta^h]$ for the mass of agents. As detailed in the Appendix, this θ^m is determined such that the aggregate optimality condition, derived using the Mean Value Theorem for Integrals, holds for this representative agent. Accordingly, the modified optimal tax formula employs a specific subgradient value $\kappa^m \in [\eta, \eta\lambda]$ that corresponds to the optimality condition at the reference point. Furthermore, to maintain dimensional consistency with the standard formula, the density term is adjusted to reflect the average density over the bunching interval.

With this definition, the optimal tax formula accounting for bunching is characterized as follows.

Proposition 2. *The optimal marginal tax rate at an arbitrary income level $y = y(\theta)$ is given by:*

(i) *For non-bunching regions, $\theta \in [\underline{\theta}, \theta^\ell) \cup (\theta^h, \overline{\theta}]$:*

$$\frac{T_p'(y) - g_p(\theta)\mu'(z(\theta))}{1 - T_p'(y) + \mu'(z(\theta))} = \frac{1+\varepsilon}{\varepsilon} \frac{1 - F(\theta)}{\theta f(\theta)} \frac{\int_\theta^{\overline{\theta}} \Big[1 - g_p(\theta')\Big] f(\theta') d\theta'}{1 - F(\theta)} \qquad \text{(13 revisited)}$$

(ii) *Over the bunching interval, $\theta \in [\theta^\ell, \theta^h]$, where $y = r$:*

$$\frac{T_p'(r) - g_p(\theta^m)\kappa^m}{1 - T_p'(r) + \kappa^m} = \underbrace{\frac{1+\varepsilon}{\varepsilon}}_{A} \underbrace{\frac{1 - F(\theta^h)}{\theta^m \bar{f}}}_{\tilde{B}p} \underbrace{\frac{\int_{\theta^h}^{\overline{\theta}} \Big[1 - g_p(\theta)\Big] f(\theta) d\theta}{1 - F(\theta^h)}}_{C_p} \qquad (15)$$

where $\theta^m \in [\theta^\ell, \theta^h]$ is the representative ability, $\kappa^m \in [\eta, \eta\lambda]$ denotes a subgradient value of the gain-loss utility $\mu(\cdot)$, and $\bar{f} \equiv \int_{\theta^\ell}^{\theta^h} f(\theta) d\theta / (\theta^h - \theta^\ell)$ represents the average density over the bunching interval.

The optimal tax formula for the bunching region (15) preserves the structural intuition of the standard ABC formula but incorporates specific adjustments for the mass of agents at the reference point. The term $\tilde{B}_p$ reflects the density of the bunched agents, replacing the

local density $f(\theta)$ with the average density $\bar{f}$ over the bunching interval $[\theta^\ell, \theta^h]$. The term κ^m represents the marginal impact of reference dependence at the kink, which takes a value within the subdifferential interval consistent with the optimality condition.

It is important to interpret this formula correctly in the context of asymmetric information. While the expression involves the representative ability θ^m and the interval $[\theta^\ell, \theta^h]$, this does not imply that the tax rate depends on unobservable individual ability. Instead, the formula characterizes the optimal marginal tax rate $T_p'(r)$ based on the government's knowledge of the aggregate ability distribution and the behavioral parameters. The government utilizes this statistical information to set the tax rate at the observable income level r, balancing efficiency and redistribution for the mass of agents pooling at this specific threshold.

3.3 Non-paternalistic government

Next, I consider the non-paternalistic case where the government fully respects individuals' reference-dependent preferences. Unlike the paternalistic case, the non-paternalistic government regards the gain-loss utility $\mu(\cdot)$ not as a behavioral bias but as a legitimate component of individual utility. Consequently, the gain-loss utility is incorporated directly into the social welfare function. The government seeks to maximize:

$$\int_{\underline{\theta}}^{\overline{\theta}} G(V(\theta)) f(\theta) d\theta \tag{16}$$

where $V(\theta)$ is the agent's full indirect utility defined in equation (3), encompassing both the true utility and the gain-loss utility $\mu(\cdot)$. The resource constraint remains the same as in equation (12).

The optimization problem for the non-paternalistic government is formally analogous to the paternalistic case. Following the same procedure, solving the relaxed problem by the Hamiltonian and applying the ironing technique to address bunching, I derive the optimal tax formula valid for the entire ability distribution. In the following proposition, I use the subscript np to denote the tax schedule and component terms specific to the non-paternalistic setting. Additionally, the bunching thresholds, $\hat{\theta}^\ell$ and $\hat{\theta}^h$, and the representative variables, $\hat{\theta}^m$ and $\hat{\kappa}^m$, are denoted with hats to explicitly indicate that these endogenous values differ from the paternalistic case due to the change in the optimal tax schedule.

Proposition 3. *The optimal marginal tax rate at an arbitrary income level $y = y(\theta)$ is given by:*

(i) *For non-bunching regions,* $\theta \in [\underline{\theta}, \hat{\theta}^{\ell}) \cup (\hat{\theta}^{h}, \overline{\theta}]$*:*

$$\frac{T'_{np}(y)}{1 - T'_{np}(y) + \mu'(z(\theta))} = \underbrace{\frac{1+\varepsilon}{\varepsilon}}_{A} \underbrace{\frac{1 - F(\theta)}{\theta f(\theta)}}_{B} \underbrace{\frac{\int_{\theta}^{\overline{\theta}} \Big[1 - g_{np}(\theta')\Big] f(\theta') d\theta'}{1 - F(\theta)}}_{Cnp} \tag{17}$$

(ii) *Over the bunching interval,* $\theta \in [\hat{\theta}^{\ell}, \hat{\theta}^{h}]$*, where* $y = r$*:*

$$\frac{T'_{np}(r)}{1 - T'_{np}(r) + \hat{\kappa}^{m}} = \underbrace{\frac{1+\varepsilon}{\varepsilon}}_{A} \underbrace{\frac{1 - F(\hat{\theta}^{h})}{\hat{\theta}^{m} \tilde{f}}}_{\tilde{B}np} \underbrace{\frac{\int_{\hat{\theta}^{h}}^{\overline{\theta}} \Big[1 - g_{np}(\theta)\Big] f(\theta) d\theta}{1 - F(\hat{\theta}^{h})}}_{Cnp} \tag{18}$$

where $g_{np}(\theta)$ *is the marginal social welfare weight,* $\hat{\theta}^{m} \in [\hat{\theta}^{\ell}, \hat{\theta}^{h}]$ *is the representative ability,* $\hat{\kappa}^{m} \in [\eta, \eta\lambda]$ *denotes a subgradient value, and* $\tilde{f}$ *represents the average density over the bunching interval defined in the same manner as in Proposition 2.*

While the structural form in the bunching region requires the adjustments defined in the optimal tax formula (18), the driving forces behind the optimal tax rate are best understood by examining the non-bunching formula (17). Rearranging it for $T'_{np}(y)$ yields:

$$T'_{np}(y) = \frac{[1 + \mu'(z(\theta))]\, ABC_{np}}{1 + ABC_{np}} \tag{19}$$

A direct structural comparison with the paternalistic formula (14) might suggest that the paternalistic tax rate should be higher, as the paternalistic numerator includes an additional corrective term $g_p(\theta)\mu'(z(\theta))$ to reduce overworking. However, this structural comparison holds only if the distributional terms ABC remain constant, which is not the case. Crucially, the non-paternalistic government treats the reference-dependent gain as a valid source of utility, resulting in higher assessed utility levels for earners above the reference point. Given the concavity of the social welfare function, such as the concave transformation used here, these higher utility levels translate into strictly lower marginal social welfare weights in the non-paternalistic case. These lower weights increase the redistribution factor C_{np} relative to C_p, exerting upward pressure on the tax rate. Since these structural and redistributive effects operate in opposite directions, the overall comparison between T'_{np} and T'_{p} is theoretically ambiguous. Therefore, I rely on the numerical simulations presented in the next section to determine which effect dominates.

4. Numerical simulations

4.1 Parameters and calibration

To examine the quantitative impact of reference dependence on the optimal tax schedule, this paper adopts the calibration framework of Mankiw et al. (2009). While the underlying data is based on the U.S. Current Population Survey (CPS) from March 2007, using this established parameter set serves as a benchmark, allowing for a direct comparison with the standard U-shaped optimal tax schedule derived by Saez (2001) and Mankiw et al. (2009). This strategy ensures that any deviations from the standard result are attributable solely to reference dependence rather than to variations in parameter selection.

Ability θ follows a lognormal distribution with a mean of 2.757 and a variance of 0.3148, interpreted as the hourly wage in U.S. dollars. Consistent with Saez (2001), the upper tail of the distribution is approximated by a Pareto distribution with parameter $\alpha = 2$ for levels above 42.5 dollars per hour to capture the thickness of the right tail of the income distribution. I set the elasticity of labor supply to $\varepsilon = 0.33$ following Chetty (2012) and assume a logarithmic social welfare function, $G(V(\theta)) = \ln V(\theta)$, as in the standard literature. Additionally, I set the scaling parameter $\chi = 0.65$ to normalize the units of labor supply.

Regarding the behavioral parameters for reference dependence, this paper employs the most recent and comprehensive estimate from the meta-analysis by Brown et al. (2024). They report a mean loss aversion coefficient of $\lambda = 1.955$, which provides a more precise measure than the seminal but historically specific estimate of 2.25 by Tversky and Kahneman (1992). Accordingly, I set the weight on the gain-loss utility to $\eta = 0.115$. This choice is justified by the composite strength of reference dependence: with $\lambda = 1.955$ and $\eta = 0.115$, the effective marginal impact of reference dependence, approximated by $\eta(\lambda - 1)$, is 0.1098. This value aligns closely with the structural estimates by Crawford and Meng (2011), who explicitly report a value of approximately 0.11 for the composite term $\eta(\lambda - 1)$ in their analysis of labor supply.

In the simulations, I compute the optimal tax schedules for the paternalistic government, as derived in Proposition 2, and for the non-paternalistic government, as derived in Proposition 3. I consider two scenarios where the exogenous reference point r is set at representative levels of the income distribution corresponding to institutional thresholds. It is important to clarify that these reference points are not empirically estimated from the CPS data but are selected as hypothetical benchmarks to simulate how the optimal tax schedule responds to potential reference-dependent behavior induced by institutional cliffs. Specifically, I set

$r = \{13,000; 40,000\}$.

The first reference point, $r = 13,000$, is chosen to approximate 130% of the HHS Poverty Guideline (13,273 dollars for a single person). This level is selected to represent a safety-net eligibility threshold analogous to the federal gross income limit for the Supplemental Nutrition Assistance Program (SNAP). While I do not claim that SNAP induces a reference point for all agents in the dataset, this threshold serves as a plausible candidate for a loss trigger where individuals earning above it might perceive the loss of eligibility as a psychological discontinuity.

The second reference point, $r = 40,000$, represents the middle-class benefit threshold. While this value aligns with the mean annual wage of 40,690 dollars reported by the Bureau of Labor Statistics, in this simulation context, it acts as a proxy for the institutional boundary where federal support programs for the middle class begin to phase out. For instance, single filers with income above this level face the gradual reduction of education-related benefits, such as the Hope Credit and student loan interest subsidies. Thus, I treat this threshold as an exogenous reference point separating the subsidized middle class from higher-income groups to examine the policy implications of such a potential behavioral response.

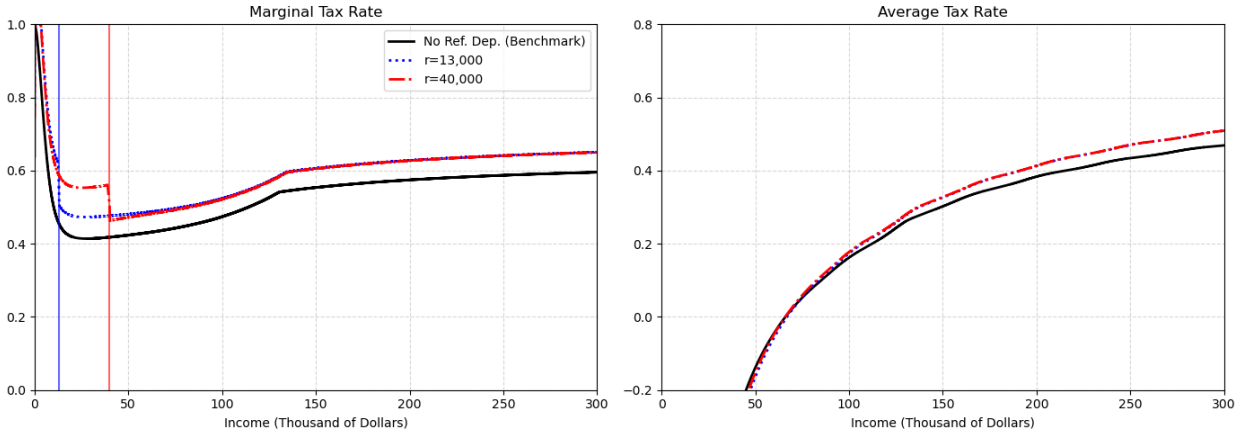
Finally, to assess the welfare implications of ignoring behavioral biases, I evaluate the welfare gain achievable by implementing the optimal paternalistic policy compared to a scenario where the government overlooks reference dependence. In this counterfactual analysis, the government incorrectly assumes that agents possess standard preferences, i.e. $\eta = 0$, and consequently implements the standard optimal tax schedule derived by Saez (2001), even though agents actually behave according to reference-dependent preferences. To quantify the inefficiency arising from this misspecification, I measure the welfare gain in terms of consumption equivalence. This metric calculates the constant percentage increase in consumption that agents would require under the standard tax schedule to attain the same level of social welfare as under the optimal paternalistic tax schedule.

4.2 Simulated tax schedules

This subsection analyzes the optimal tax schedules derived under reference-dependent preferences. I compare the results for the paternalistic criterion and the non-paternalistic criterion sequentially.

Figure 2 displays the optimal tax schedules under the paternalistic criterion. The solid black line represents the benchmark case without reference dependence, while the dotted blue and red lines represent the cases with reference income levels of $r = 13,000$ and $r = 40,000$,

Figure 2: Optimal tax schedules under paternalistic criterion



Note: The left panel displays the optimal marginal tax rates, and the right panel displays the average tax rates. The black solid line represents the benchmark case without reference dependence where $\eta = 0$. The blue and red dotted lines represent the cases with reference points at $r = 13,000$ and $r = 40,000$, respectively. The vertical lines indicate the locations of the corresponding reference points. Parameters are set to $\eta = 0.1$ and $\lambda = 1.955$.

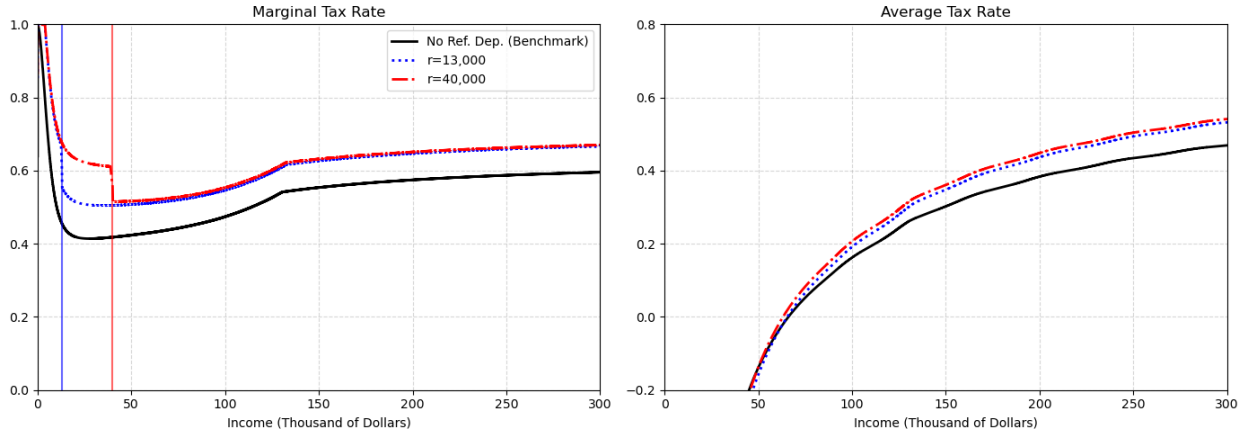
respectively.

The paternalistic tax schedule is characterized by several distinct features relative to the benchmark. One primary feature is that the optimal marginal tax rates are substantially higher than the benchmark at all income levels. While the benchmark marginal tax rate falls to approximately 40% at the bottom of the U-shape, the paternalistic marginal tax rates remain consistently above 50%. This general upward shift occurs because reference dependence, particularly loss aversion, makes labor supply less elastic. Even though the paternalistic government wishes to correct the overworking bias, the inverse-elasticity rule dominates, meaning that the government can impose higher taxes on the less elastic tax base with relatively smaller efficiency costs.

Another defining characteristic is the sharp discontinuity at the reference point. As illustrated by the dotted red line for the $r = 40,000$ case, the marginal tax rate slightly rises as it approaches the reference point and then drops precipitously to just below 50% immediately after passing the threshold. This drastic adjustment is required to handle the mass of agents bunching at the kink. Apart from these localized adjustments at the specific reference points, the tax schedules for $r = 13,000$ and $r = 40,000$ are remarkably similar.

As shown in the figure, outside the bunching regions, the two dotted lines almost over-

Figure 3: Optimal tax schedules under non-paternalistic criterion



Note: The left panel displays the optimal marginal tax rates, and the right panel displays the average tax rates under the non-paternalistic criterion. The color scheme and parameter values correspond to those in Figure 2. The vertical lines indicate the locations of the corresponding reference points.

lap and exhibit a parallel upward deviation from the benchmark. This suggests that the fundamental driver of the tax increase, namely the globally reduced elasticity due to reference dependence, operates similarly across the income distribution, regardless of the specific location of the reference point.

Finally, the right panel of Figure 2 displays the average tax rates. Consistent with the higher marginal tax rates, the average tax rate curves for both reference-dependent cases lie strictly above the benchmark curve for middle- and high-income levels. This upward shift implies that the optimal tax system becomes more progressive under paternalism. The government imposes a higher average tax burden on these income groups not only to correct the behavioral bias but also to finance the expanded lump-sum transfers required to support low-income agents.

Next, Figure 3 presents the results for the non-paternalistic criterion in which the government fully respects reference-dependent preferences. While the general shape of the tax schedules remains similar to the paternalistic case with localized adjustments around reference points, a crucial quantitative difference emerges.

A direct comparison between Figure 2 and Figure 3 reveals that the non-paternalistic government imposes strictly higher marginal tax rates at almost all income levels. As discussed in Section 3, this result is driven by the dominance of the redistributive motive over

the structural correction motive. Since the non-paternalistic government interprets the gain utility from high income as a valid source of welfare, high-income earners are assessed as having higher utility levels. Under a concave social welfare function, these higher utility levels translate into lower social welfare weights for such agents. This valuation strengthens the motive for the government to extract revenue and redistribute it to the lower-income group. Reflecting this logic, the average tax rate curves in Figure 3 shift further upward compared to the paternalistic case, which indicates an even stronger demand for progressivity under the non-paternalistic criterion.

4.3 Redistributive implications and welfare analysis

The structural changes in tax schedules translate into significant implications for redistribution and social welfare. Table 1 summarizes the impact of the optimal tax reform by comparing it to the standard benchmark case where the government ignores reference dependence (i.e., assumes $\eta = 0$). Specifically, it reports the percentage changes in the lump-sum transfer to the unemployed, $T(0)$, and the welfare gains measured in Consumption Equivalent Variation (CEV).

Table 1: Redistributive effects and welfare gains

	Paternalistic		**Non-Paternalistic**	
Variable	$r = 13,000$	$r = 40,000$	$r = 13,000$	$r = 40,000$
Increase in Lump-sum Transfer ($T(0)$)	18.79%	23.28%	23.78%	31.75%
Welfare Gain (CEV)	6.02%	6.77%	5.83%	7.51%

Note: Increase in Lump-sum Transfer represents the percentage change in the intercept of the tax schedule, $T(0)$, relative to the benchmark standard optimal tax schedule derived with $\eta = 0$. Welfare Gain denotes the percentage increase in social welfare in consumption equivalence achieved by optimizing the tax schedule for the given reference parameters, compared to the scenario where the standard tax schedule is applied to reference-dependent agents.

Accounting for reference dependence leads to a substantial expansion in the safety net compared to the standard optimal tax model. Under the paternalistic criterion, the lump-sum transfer increases by approximately 19–23%, while under the non-paternalistic criterion, it increases by 24–32%. This larger expansion under non-paternalism is consistent with the higher tax rates observed in Figure 3, reflecting the stronger redistributive preference derived from the lower social welfare weights.

These welfare figures quantify the cost of overlooking behavioral biases. The benchmark

for this calculation is an economy managed by a government that incorrectly assumes standard preferences ($\eta = 0$) and implements the corresponding standard tax schedule. Because this approach fails to recognize the reduced elasticity of labor supply caused by reference dependence, it sets marginal tax rates too low and provides insufficient lump-sum transfers. Consequently, the reported welfare gains of 5.8–7.5% represent the welfare improvement in terms of consumption equivalent variation that agents would enjoy by shifting from this misspecified policy regime to the optimal one, where the government correctly designs the tax schedule and the safety net to the reference-dependent preferences.

Finally, I examine the robustness of these results with respect to alternative parameter specifications. I conduct sensitivity analyses by varying the labor supply elasticity ε within the range of $[0.2, 0.5]$, which encompasses the standard estimates used in the literature (Chetty, 2012; Piketty and Saez, 2013). Similarly, I vary the gain parameter η within $[0.05, 0.3]$ and the loss aversion parameter λ within $[1.5, 2.5]$, ranges consistent with experimental evidence (Tversky and Kahneman, 1992; Brown et al., 2024). The qualitative findings, specifically the higher marginal tax rates under non-paternalism and the substantial expansion of the safety net, remain robust across these parameter values. Furthermore, while the baseline model assumes a piecewise linear gain-loss utility which generates bunching, I also consider a smoothed specification in Appendix B. As shown in Figure B.1 in the Appendix, although the discrete bunching disappears under the smoothed utility function, the optimal tax schedule continues to exhibit locally high marginal tax rates around the reference point, and the key welfare implications remain unchanged.

5. Conclusion

This paper incorporates reference-dependent preferences into the Mirrlees optimal income tax framework. Motivated by the observation that institutional thresholds often serve as exogenous reference points inducing behavioral bunching, I developed a theoretical approach using the ironing technique to characterize optimal tax schedules.

My analysis yields three key insights. First, the presence of loss aversion fundamentally alters the trade-off between efficiency and equity. The optimal marginal tax rates are consistently higher than the standard benchmark at all income levels. This generalized increase is rationalized by the inverse-elasticity rule, which prescribes higher taxes to minimize the efficiency cost of raising revenue. While this principle operates across the entire distribution,

its effect is further intensified in the loss domain, where loss aversion explicitly lowers the labor supply elasticity, causing marginal tax rates to rise even further.

Second, this increased revenue generating capacity should be directed toward a substantial expansion of the safety net. The simulations calibrated to the U.S. economy demonstrate that the lump-sum transfer to the unemployed should increase by approximately 19–32% to maximize social welfare. Correctly accounting for these behavioral responses yields significant welfare gains, equivalent to 5.8–7.5% of consumption, compared to the standard policy benchmark.

Third, these prescriptions are remarkably robust. I examined two distinct social welfare criteria: a paternalistic view that corrects behavioral bias and a non-paternalistic view that fully respects reference-dependent preferences. While the paternalistic objective is to correct the inefficiency of overworking and the non-paternalistic objective is to redistribute to agents with high marginal utility, both criteria converge on the necessity of higher marginal tax rates and increased redistribution. Furthermore, as shown in the Appendix, these results hold not only under the piecewise linear utility that induces bunching but also under a smoothed specification. This confirms that the findings are not artifacts of the functional form but are driven by the fundamental properties of reference dependence.

The theoretical framework developed here opens several avenues for future research. While this paper focused on a single exogenous reference point, extending the model to heterogeneous reference points or endogenous reference formation based on past income or consumption (status quo) would be a natural next step. Additionally, applying this framework to other institutional contexts, such as the "1.03 million yen wall" for spousal deductions in Japan, could provide valuable policy implications for joint taxation and female labor market participation.

Appendix

A. Proof for propositions

A.1 Proposition 1

The paternalistic government maximizes the social welfare function defined over the standard utility $U(\theta)$, treating the gain-loss utility $\mu(\cdot)$ as a behavioral bias, subject to the resource constraint and the incentive compatibility constraint. Consistent with the individual's optimization problem, we use the total utility $V(\theta)$ as the state variable and income $y(\theta)$ as the control variable.

Using the relationship $U(\theta) = V(\theta) - \mu(z(\theta))$ where $z(\theta) = y(\theta) - r$, and substituting the consumption $c(\theta)$ from the utility function into the resource constraint, the Hamiltonian is formulated as:

$$H = \left[G(U(\theta)) + \gamma\left(y(\theta) - V(\theta) - v\left(\frac{y(\theta)}{\theta}\right) + \mu(z(\theta))\right)\right] f(\theta) + \delta(\theta) v'\left(\frac{y(\theta)}{\theta}\right)\frac{y(\theta)}{\theta^2} \quad (20)$$

The first-order conditions with respect to $y(\theta)$ and $V(\theta)$ are derived as follows. Note that $\frac{\partial U(\theta)}{\partial y(\theta)} = -\mu'(z(\theta))$.

$$\begin{aligned}\frac{\partial H}{\partial y(\theta)} &= \gamma\Big[1 - v'\left(\frac{y(\theta)}{\theta}\right)\frac{1}{\theta} + \mu'(z(\theta))\left(1 - \frac{G'(U(\theta))}{\gamma}\right)\Big] f(\theta) \\ &\quad + \delta(\theta)\left[v'\left(\frac{y(\theta)}{\theta}\right)\frac{1}{\theta^2} + v''\left(\frac{y(\theta)}{\theta}\right)\frac{y(\theta)}{\theta^3}\right] = 0 \quad (21)\\ \frac{\partial H}{\partial V(\theta)} &= \Big[G'(U(\theta)) - \gamma\Big] f(\theta) = -\delta'(\theta) \quad (22)\end{aligned}$$

where $\mu'(z(\theta))$ denotes the marginal gain-loss utility.

Integrating equation (22) over the entire ability space and applying the transversality conditions $\delta(\underline{\theta}) = \delta(\overline{\theta}) = 0$, we obtain the expression for the marginal value of public funds:

$$\gamma = \int_{\underline{\theta}}^{\overline{\theta}} \Big[G'(U(\theta))\Big] f(\theta) d\theta \quad (23)$$

We also find the value of the co-state variable $\delta(\theta)$ by integrating equation (22) from an

arbitrary ability level θ to the upper bound $\overline{\theta}$:

$$\int_{\theta}^{\overline{\theta}} \Big[G'(U(\theta')) - \gamma\Big] f(\theta')d\theta' = -\int_{\theta}^{\overline{\theta}} \delta'(\theta')d\theta' = \delta(\theta) \tag{24}$$

where we used the definition of marginal social welfare weights $g_p(\theta) \equiv G'(U(\theta))/\gamma$.

Substituting $\delta(\theta)$ into the first-order condition (21) and rearranging terms yields:

$$\frac{1 - v'\left(\frac{y(\theta)}{\theta}\right)\frac{1}{\theta} + \mu'(z(\theta))(1 - g_p(\theta))}{v'\left(\frac{y(\theta)}{\theta}\right)\frac{1}{\theta}} = \frac{1 - F(\theta)}{\theta f(\theta)} \frac{\int_{\theta}^{\overline{\theta}} \Big[1 - g_p(\theta')\Big] f(\theta')d\theta'}{1 - F(\theta)} \left(1 + \frac{1}{\varepsilon}\right) \tag{25}$$

where $\varepsilon \equiv \frac{v'(n)}{nv''(n)}$ represents the labor supply elasticity.

Using the individual's first-order condition $v'\left(\frac{y(\theta)}{\theta}\right)\frac{1}{\theta} = 1 - T_p'(y(\theta)) + \mu'(z(\theta))$, we obtain the optimal tax formula for the paternalistic government:

$$\frac{T_p'(y) - g_p(\theta)\mu'(z(\theta))}{1 - T_p'(y) + \mu'(z(\theta))} = \frac{1 + \varepsilon}{\varepsilon} \frac{1 - F(\theta)}{\theta f(\theta)} \frac{\int_{\theta}^{\overline{\theta}} \Big[1 - g_p(\theta')\Big] f(\theta')d\theta'}{1 - F(\theta)} \tag{13 revisited}$$

A.2 Proposition 2

The derivation follows the Hamiltonian framework established in the proof for Proposition 1. Outside the bunching region, the optimal tax schedule is characterized by the pointwise first-order condition (21).

Inside the bunching region for $\theta \in [\theta^{\ell}, \theta^{h}]$, the monotonicity constraint binds, and all agents obtain the same income $y(\theta) = r$. Consequently, the pointwise condition (21), $\partial H/\partial y(\theta) = 0$, is replaced by the integral condition over the bunching interval:

$$\int_{\theta^{\ell}}^{\theta^{h}} \frac{\partial H}{\partial y(\theta)} d\theta = 0 \tag{26}$$

Substituting the pointwise condition (21), and noting that $y(\theta)$ is constant at r within this interval, the condition becomes:

$$\int_{\theta^{\ell}}^{\theta^{h}} \left\{ \gamma \left[1 - \frac{v'\left(\frac{r}{\theta}\right)}{\theta} + \kappa(\theta)\left(1 - g_p(\theta)\right) \right] f(\theta) + \delta(\theta) \left[\frac{v'\left(\frac{r}{\theta}\right)}{\theta^2} + \frac{v''\left(\frac{r}{\theta}\right) r}{\theta^3} \right] \right\} d\theta = 0 \tag{27}$$

where $\kappa(\theta)$ represents a subgradient of the gain-loss utility $\mu(z(\theta))$ at the kink where $z(\theta) = 0$ for an agent with ability θ.

We define the total mass of bunched agents as $M \equiv \int_{\theta^\ell}^{\theta^h} f(\theta)d\theta$ and the width of the bunching interval as $\Delta\theta \equiv \theta^h - \theta^\ell$. Applying the Mean Value Theorem for Integrals, specifically considering the weighted integral for the welfare term and the standard integral for the incentive term, we define a representative ability $\theta^m \in [\theta^\ell, \theta^h]$ and a corresponding representative subgradient $\kappa^m \in [\eta, \eta\lambda]$ such that the aggregate optimality condition holds. This allows us to rewrite equation (27) as:

$$\gamma \left[1 - \frac{v'\left(\frac{r}{\theta^m}\right)}{\theta^m} + \kappa^m(1 - g_p(\theta^m))\right] M + \delta(\theta^m) \left[\frac{v'\left(\frac{r}{\theta^m}\right)}{(\theta^m)^2} + \frac{v''\left(\frac{r}{\theta^m}\right) r}{(\theta^m)^3}\right] \Delta\theta = 0 \tag{28}$$

Note that the welfare term, which is proportional to the population, is scaled by the mass M, whereas the incentive cost term involving the co-state variable δ is scaled by the interval width $\Delta\theta$.

Finally, we derive the optimal tax formula. Using the representative agent's first-order condition to express terms in $T_p'(r)$, and approximating the co-state variable term as $-\delta(\theta^m)/\gamma \approx \int_{\theta^h}^{\overline{\theta}} [1 - g_p(\theta)] f(\theta)d\theta$, we obtain the result. To maintain structural consistency with the standard ABC formula, the term $M/\Delta\theta$ is interpreted as the average density $\overline{f}$ over the bunching interval:

$$\frac{T_p'(r) - g_p(\theta^m)\kappa^m}{1 - T_p'(r) + \kappa^m} = \frac{1+\varepsilon}{\varepsilon} \frac{1 - F(\theta^h)}{\theta^m(M/\Delta\theta)} \frac{\int_{\theta^h}^{\overline{\theta}} \Big[1 - g_p(\theta)\Big] f(\theta)d\theta}{1 - F(\theta^h)} \tag{15 revisited}$$

A.3 Proposition 3

The optimization problem for the non-paternalistic government is analogous to the paternalistic case, with the exception that the social welfare function depends on the full indirect utility $V(\theta)$. The Hamiltonian is formulated as:

$$H = \left[G(V(\theta)) + \hat{\gamma}\left(y(\theta) - V(\theta) - v\left(\frac{y(\theta)}{\theta}\right) + \mu(z(\theta))\right)\right] f(\theta) + \delta(\theta) v'\left(\frac{y(\theta)}{\theta}\right) \frac{y(\theta)}{\theta^2} \tag{29}$$

where $\hat{\gamma}$ is the multiplier for the resource constraint. Note that unlike the paternalistic case, the gain-loss utility $\mu(\cdot)$ enters the Hamiltonian directly through the resource constraint but

is also implicitly included in the state variable $V(\theta)$ within the welfare function $G(V(\theta))$.

(i) Non-bunching regions

The first-order condition with respect to income $y(\theta)$ is:

$$\frac{\partial H}{\partial y(\theta)} = \hat{\gamma}\Big[1 - v'\left(\frac{y(\theta)}{\theta}\right)\frac{1}{\theta} + \mu'(z(\theta))\Big]f(\theta) + \delta(\theta)\left[v'\left(\frac{y(\theta)}{\theta}\right)\frac{1}{\theta^2} + v''\left(\frac{y(\theta)}{\theta}\right)\frac{y(\theta)}{\theta^3}\right] = 0 \tag{30}$$

Crucially, since the state variable is the full utility $V(\theta)$, the derivative of the welfare function $G(V(\theta))$ does not appear in this condition, differing from the paternalistic case where $\partial G(U)/\partial y = -G'(U)\mu'$.

Using the individual's first-order condition (4) and proceeding with the same rearrangements as in the proof for Proposition 1, we obtain the optimal tax formula for the non-bunching region (17).

(ii) Bunching interval

For the bunching region $\theta \in [\hat{\theta}^\ell, \hat{\theta}^h]$, the derivation follows the exact logic established in the proof for Proposition 2. We apply the integral condition $\int_{\hat{\theta}^\ell}^{\hat{\theta}^h} \partial H/\partial y d\theta = 0$ and the Mean Value Theorem for Integrals.

Defining the representative ability $\hat{\theta}^m$ and the subgradient $\hat{\kappa}^m$ to satisfy the aggregate optimality, the integral of the welfare term becomes proportional to $T'_{np}(r)$ due to the simplification shown above. By approximating the co-state variable term using the integral of $1 - g_{np}(\theta)$ from $\hat{\theta}^h$, and adjusting the density to the average density $\tilde{f}$ over the interval, we arrive at the optimal tax formula for the bunching region (18).

B. Robustness to smoothed non-linear specification

In the main text, I assumed a piecewise linear gain-loss utility function. This simplification raises two potential concerns regarding the robustness of the results: first, the results might be driven by the mathematical discontinuity at the kink; and second, the linear model might overestimate the tax increase for high-income earners by ignoring diminishing sensitivity. To address these issues, this appendix examines a smoothed non-linear specification. Specifically, I adopt a shifted power function form inspired by Tversky and Kahneman (1992). To avoid confusion with the notation in the main text, I denote the curvature parameter as σ

and the smoothing shift parameter as ξ. The smoothed gain-loss utility $R(z)$ is defined as:

$$\mu(z) = \begin{cases} \eta \dfrac{(z+\xi)^\sigma - \xi^\sigma}{\sigma} & \text{if } z \geq 0, \\ -\eta\lambda \dfrac{(-z+\xi)^\sigma - \xi^\sigma}{\sigma} & \text{if } z < 0, \end{cases} \tag{31}$$

where $\sigma \in (0,1)$ governs the diminishing sensitivity and $\xi > 0$ is a shift parameter introduced to smooth the kink at $z = 0$ and ensure differentiability. Following the literature, I set $\sigma = 0.8$ and a smoothing parameter $\xi = 1.0$. $\eta = 0.1$ and $\lambda = 1.955$ are maintained as in the main part.

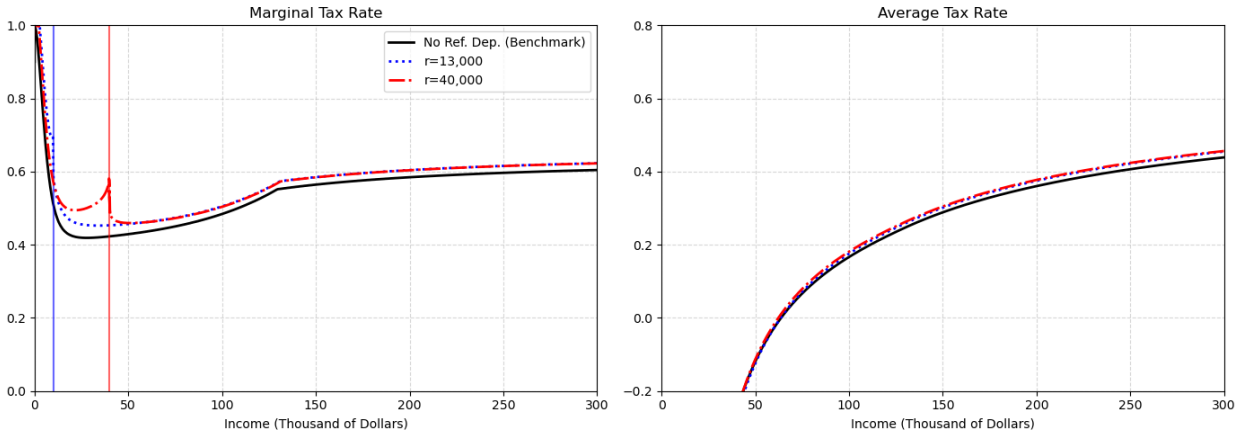
Figure B.1 displays the optimal marginal tax rates under this smoothed utility. The results demonstrate that the main findings remain robust even under this generalized specification. Although the following part confirms robustness only for the paternalistic case, the same argument can be applicable to the non-paternalistic case.

First, regarding the local behavior, the optimal tax schedule exhibits a sharp localized spike around the reference point rather than a vertical drop. Unlike the linear case, the smoothed power function implies that marginal utility increases as income gets closer to the reference point. Consequently, the optimal marginal tax rate rises steeply as it approaches the reference point, confirming that the incentive to correct for local bunching is not an artifact of the discontinuity but a fundamental response to the rapid change in marginal utility.

Second, regarding the global trend, this specification addresses the concern that reference dependence should theoretically fade away for agents far from the reference point due to diminishing sensitivity where $\sigma < 1$. As shown in the figure, the deviation from the benchmark indeed behaves according to this intuition, meaning that the gap is largest near the reference point and gradually attenuates as income increases, converging toward the benchmark at very high income levels. However, crucially, the marginal tax rates remain strictly higher than the benchmark across the relevant income range. This confirms that the qualitative result, globally higher marginal tax rates, is robust even when accounting for the diminishing impact of reference dependence.

Finally, it is verified that under the adopted parameterization and the specific functional form, the single crossing property is satisfied. Since the marginal disutility of labor decreases with wage type θ while the marginal gain-loss utility depends only on income y, the indifference curves satisfy the standard sorting condition, ensuring the validity of the optimization approach.

Figure B.1: Optimal tax schedules under smoothed non-linear utility (paternalistic)



Note: The figure shows the optimal marginal and average tax rates under the paternalistic criterion using a smoothed gain-loss utility function with $\sigma = 0.8$ and $\xi = 1.0$ as well as $\eta = 0.1$ and $\lambda = 1.955$. The solid black line represents the benchmark case without reference dependence. The dotted blue and dash-dotted red lines represent the cases with reference points at $r = 13,000$ and $r = 40,000$, respectively. The vertical lines indicate the locations of the corresponding reference points.

References

[1] **Allcott, Hunt, Benjamin B. Lockwood, and Dmitry Taubinsky.** 2019. "Regressive sin taxes, with an application to the optimal soda tax." *The Quarterly Journal of Economics* 134 (3): 1082–1095.

[2] **Aronsson, Thomas, and Olof Johansson-Stenman.** 2008. "When the Joneses' consumption hurts: Optimal public good provision and nonlinear income taxation." *Journal of Public Economics* 92 986–997.

[3] **Aronsson, Thomas, and Olof Johansson-Stenman.** 2018. "Peternalism against Veblen." *American Economic Journal: Economic Policy* 10 (1): 39–76.

[4] **Aronsson, Thomas, and Olof Johansson-Stenman.** 2024. "Optimal taxation and other-regarding preferences." *Graz Economics Papers* 2024-22.

[5] **Bernheim, B. Douglas, and Antonio Rangel.** 2009. "Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics." *The Quarterly Journal of Economics* 124 (1): 51–104.

[6] **Bowman, David, Deborah Minehart, and Matthew Rabin.** 1999. "Loss aversion in a consumption-savings model." *Journal of Economic Behavior & Organization* 38 155–178.

[7] **Brett, Craig, and John A. Weymark.** 2017. "Voting over selfishly optimal nonlinear income tax schedules." *Games and Economic Behavior* 101 172–188.

[8] **Brito, Degobert L., and William H. Oakland.** 1977. "Some properties of the optimal income-tax." *International Economic Review* 18 (2): 407–423.

[9] **Britton, Jack, and Jonathan Gruber.** 2020. "Do income contingent student loans reduce labor supply?" *Economics of Education Review* 79 102061.

[10] **Brown, Alexander L., Taisuke Imai, Ferdinand M. Vieider, and Colin F. Camerer.** 2024. "Loss aversion in a consumption±savings model." *Journal of Economic Literature* 62 (2): 485–516.

[11] **Chapman, Bruce, and Andrew Leigh.** 2009. "Do very high tax rates induce bunching? Implications for the design of income contingent loan schemes." *Economic Record* 85 (270): 276–289.

[12] **Chetty, Raj.** 2012. "Bounds on elasticities with optimization frictions: A synthesis of micro and macro evidence on labor supply." *Econometrica* 80 (3): 969–1018.

[13] **Chetty, Raj, Adam Looney, and Kory Kroft.** 2009. "Salience and taxation: theory and evidence." *American Economic Review* 99 (4): 1145–1177.

[14] **Crawford, Vincent P., and Juanjuan Meng.** 2011. "New York City cab cab drivers' labor supply revisited: Reference-dependent preferences with rational-expectations targets for hours and income." *The American Economic Review* 101 (5): 1912–1932.

[15] **Diamond, Peter A.** 1998. "Optimal income taxation: An example with a U-shaped pattern of optimal marginal tax rates." *American Economic Review* 88 (1): 83–95.

[16] **Ebert, Udo.** 1992. "A reexamination of the optimal nonlinear income tax." *Journal of Public Economics* 49 (1): 47–73.

[17] **Fukai, Taiyo, and Ayako Kondo.** 2023. "Labor supply of married women, kink-points on tax schedule and social security premium notch: rvidence from municipality tax records in Japan (Japanese)." discussion papers, Research Institute of Economy, Trade and Industry (RIETI).

[18] **Gelber, Alexander M., Damon Jones, and Daniel W. Sacks.** 2020. "Estimating adjustment frictions using nonlinear budget sets: Method and evidence from the earnings test." *American Economic Journal: Applied Economics* 12 (1): 1–31.

[19] **Guesnerie, Roger.** 1995. *A contribution to the pure theory of taxation.* Cambridge University Press.

[20] **Guesnerie, Roger, and Jean-Jacques Laffont.** 1984. "A complete solution to a class of principal-agent problems with an application to the control of a self-managed firm." *Journal of Public Economics* 25 329–369.

[21] **Hammond, Peter.** 1979. "Straightforward individual incentive compatibility in large economies." *The Review of Economic Studies* 46 (2): 263–282.

[22] **Kanbur, Ravi, Jukka Pirttilä, and Matti Tuomala.** 2008. "Moral hazard, income taxation and prospect theory." *The Scandinavian Journal of Economics* 110 (2): 321–337.

[23] **Kanbur, Ravi, and Matti Tuomala.** 2013. "Relativity, inequality, and optimal nonlinear income taxation." *International Economic Review* 54 (4): 1199–1217.

[24] **Kleven, Henrik J.** 2016. "Bunching." *The Annual Review of Economics* 8 435–464.

[25] **Kleven, Henrik J., and Mazhar Waseem.** 2013. "Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan." *The Quarterly Journal of Economics* 128 (2): 669–723.

[26] **Köszegi, Botond, and Matthew Rabin.** 2006. "A model of reference-dependent preferences." *The Quarterly Journal of Economics* 121 (4): 1133–1165.

[27] **Lockwood, Benjamin B.** 2020. "Optimal income taxation with present bias." *American Economic Journal: Economic Policy* 12 (4): 298–327.

[28] **Lollivier, Stefan, and Jean-Charles Rochet.** 1983. "Bunching and second-order conditions: A note on optimal tax theory." *Journal of Economic Theory* 31 392–400.

[29] **Mankiw, N. Gregory, Matthew Weinzierl, and Danny Yagan.** 2009. "Optimal taxation in theory and practive." *Journal of Economic Perspectives* 23.

[30] **Mirrlees, James A.** 1971. "An exploration in the theory of optimum income taxation." *The Review of Economic Studies* 38 (2): 175–208.

[31] **Piketty, Thomas, and Emmanuel Saez.** 2013. "Optimal labor income taxation." *Handbook of Public Economics, Vol. 5* 391–474.

[32] **Rees-Jones, Alex.** 2018. "Quantifying loss-averse tax manipulation." *The Review of Economic Studies* 85 (2): 1251–1278.

[33] **Saez, Emmanuel.** 2001. "Using elasticities to derive optimal income tax rates." *The Review of Economic Studies* 68 (1): 205–229.

[34] **Saez, Emmanuel.** 2010. "Do taxpayers bunch at kink points?" *American Economic Journal: Economic Policy* 2 (3): 180–212.

[35] **Seibold, Arthur.** 2021. "Reference points for retirement behavior: Evidence from German pension discontinuities." *American Economic Review* 111 (4): 1126–1165.

[36] **Tversky, Amos, and Daniel Kahneman.** 1991. "Loss aversion in riskless choice: A reference-dependent model." *The Quarterly Journal of Economics* 106 (4): 1039–1061.

[37] **Tversky, Amos, and Daniel Kahneman.** 1992. "Advances in prospect theory: Cumulative representation of uncertainty." *Journal of Risk and Uncertainty* 5 297–323.

[38] **Weymark, John.** 1986. "Bunching properties of optimal nonlinear income taxes." *Social choice and Welfare* 3 213–232.