NHH

INSTITUTT FOR FORETAKSØKONOMI

DEPARTMENT OF BUSINESS AND MANAGEMENT SCIENCE

FOR 33 2014

ISSN: 1500-4066

September 2014

Discussion paper

Optimal Redistribution and Monitoring of Labor Effort

BY

Floris T. Zoutman AND **Bas Jacobs**

Norges Handelshøyskole

NORWEGIAN SCHOOL OF ECONOMICS

Optimal Redistribution and Monitoring of Labor Effort[*]

Floris T. Zoutman[**] Bas Jacobs[***]

September 10, 2014

Abstract

This paper extends the Mirrlees (1971) model of optimal non-linear income taxation with a monitoring technology that allows the government to verify labor effort at a positive, but non-infinite cost. We analyze the joint determination of the non-linear monitoring and tax schedules and the conditions under which these can be implemented. Monitoring of labor effort reduces the distortions created by income taxation and raises optimal marginal tax rates, possibly above 100 percent. The optimal intensity of monitoring increases with the marginal tax rate and the labor-supply elasticity. Our simulations demonstrate that monitoring strongly alleviates the trade-off between equity and efficiency as welfare gains of monitoring are around 1.4 percent of total output. The optimal intensity of monitoring follows a U-shaped pattern, similar to that of optimal marginal tax rates. Our paper can explain why large welfare states optimally rely on work-dependent tax credits, active labor-market policies, benefit sanctions and work bonuses in welfare programs to redistribute income efficiently.

Keywords: optimal non-linear taxation, monitoring, costly verification ability/effort, optimal redistribution
JEL-codes: H21, H26, H24, H31

[*] The authors were awarded the IIPF Young Economist Award 2013 for economists under age 40 for this paper. The authors would like to thank Luca Micheletto, Jean-Marie Lozachmeur, Aart Gerritsen, Katherine Cuff, Dirk Schindler and Robin Boadway for very useful suggestions and comments on an earlier version. Furthermore, this paper benefitted from comments and suggestions made by participants at the 66th and 69th IIPF Conferences in Uppsala and Taormina; the CPEG Conference, Quebec; and seminar participants at the Erasmus School of Economics. All remaining errors are our own. The Matlab programs used for the computations in this paper are available from the authors on request.

[**] NHH Norwegian School of Economics, Department of Business and Management Science. E-mail: floris.zoutman@nhh.no

[***] Erasmus University Rotterdam, Tinbergen Institute, Netspar, and CESifo. Corresponding author: Erasmus School of Economics, Erasmus University Rotterdam, PO box 1738, 3000 DR Rotterdam, The Netherlands. Phone: +31–10–4081481/1491. E-mail: bjacobs@ese.eur.nl. Internet: http://people.few.eur.nl/bjacobs

> "Informational frictions are a specification of a particular type of technology. For example, when we say "effort is hidden", we are really saying that it is infinitely costly for society to monitor effort. The desired approach would be to devise optimal tax systems for different specifications of the costs of monitoring different activities and/or individual attributes. To be able to implement this approach, we need to ... extend our modes of technical analysis to allow for costs of monitoring other than zero or infinity." Kocherlakota (2006, pp. 295-296)

1 Introduction

Redistribution of income is one of the most important tasks of modern welfare states. However, redistribution is expensive as it distorts the incentives to supply labor. As a result, there is a trade-off between equity and efficiency. On a fundamental level, Mirrlees (1971) demonstrates that the trade-off between equity and efficiency originates from an information problem. Earnings ability and labor hours are private information, and the government cannot condition redistributive taxes and transfers on earnings ability. Therefore, the government cannot distinguish individuals that are unable to work from individuals that are unwilling to work, and redistribution from high-income to low-income earners inevitably distorts the incentives to supply labor hours.

In practice, labor supply is not completely non-verifiable, as assumed by Mirrlees (1971). Indeed, some welfare states do condition the tax burden on some measure of hours worked. For example, in the UK low-income individuals receive a tax credit if they work more than 30 hours. This policy can only be implemented if the government is able to verify hours worked. Similar restrictions apply to in-work tax credits in Ireland and New Zealand, see also OECD (2011). Clearly, the assumption that hours worked and earnings ability are not verifiable is a too strong assumption. In the real world, the government does verify hours worked of some individuals to some extent, albeit at a cost. Consequently, the government can – to some extent – separate shirking high-ability individuals from hard-working low-ability individuals.

This paper extends Mirrlees (1971) by allowing the government to operate a monitoring technology. The monitoring technology allows the government to verify labor hours of an individual at a positive, but finite cost. The government optimally sets the *monitoring schedule* as a function of gross income. That is, the probability that an individual is monitored depends (possibly non-linearly) on his/her gross labor earnings. If an individual is monitored, the government perfectly verifies his/her labor supply and can thus deduce a worker's ability. By monitoring hours worked the government can thus provide incentives to individuals to change their labor supply in a direction that the government desires. How exactly these incentives are provided is immaterial. We can formulate our model such that individuals receive a work bonus when they meet a certain reference level of labor hours. This would correspond to the type of work bonuses observed in the UK, Ireland and New Zealand. Alternatively, we can formulate our model such that monitored individuals receive a penalty when their hours worked fall short of a minimum reference level of hours worked, which corresponds to observed work requirements and conditional welfare benefits in most advanced welfare states.

Each individual is aware of the monitoring and tax schedules before making labor-supply decisions. Hence, individuals can alter their monitoring probability by adjusting their hours worked. The total *wedge* on labor supply consists of the explicit income tax rate and an implicit subsidy on labor supply due to monitoring. Monitoring of hours worked acts as an implicit subsidy on labor supply for two reasons. First, the expected bonus increases (penalty decreases) in the difference between hours worked and the reference level of hours worked. Second, the monitoring intensity may decrease with gross earnings, depending on the shape of the monitoring schedule. For a given tax rate, monitoring can thus reduce the distortions of the income tax on labor supply, thereby increasing both equity and efficiency.

The government maximizes social welfare by optimally setting the non-linear monitoring intensity, alongside the optimal non-linear income tax.[1] We solve for the optimal non-linear tax and monitoring schedules by decentralizing the optimal, incentive-compatible direct mechanism that induces truthful revelation of ability types. We do not deviate from Mirrlees (1971) in that individuals always truthfully report earnings.[2] The schedule of optimal non-linear labor wedges is affected in two important ways in comparison to Mirrlees (1971). First, the monitoring intensity reduces the efficiency costs of the labor wedge, and thus allows for higher marginal tax rates. Second, a decrease in labor supply directly increases the penalties (or decreases work bonuses). Monitoring generates within-ability inequality between monitored and non-monitored individuals. Therefore, higher marginal taxes result in a distributional loss due to monitoring activities. The net effect of monitoring on the optimal wedge is thus theoretically ambiguous.

In Mirrlees (1971) tax rates at, or above, 100 percent can never be optimal. In contrast to Mirrlees (1971), we demonstrate that marginal tax rates could optimally be larger than 100 percent due to optimal monitoring. In particular, individuals may supply labor even if the marginal income tax rate is above 100 percent, as long as the total wedge on labor remains below 100 percent. This could explain why effective marginal tax rates of close to, or even higher than, 100 percent are observed in real-world tax-benefit systems in the phase-out range of means-tested benefits. See Immervoll (2004), Spadaro (2005), Brewer et al. (2010) and OECD (2011) for examples in OECD countries. The non-linear monitoring schedule is set so as to equate the marginal cost of monitoring to the marginal efficiency gain associated with monitoring at each gross income level. The efficiency gain of monitoring is increasing in the distortion created by the wedge on labor. Therefore, the optimal monitoring intensity increases with both the total labor wedge and the labor-supply elasticity.

Unfortunately, there is no closed-form solution for the optimal tax and monitoring schedules. Therefore, we resort to numerical simulations based on a realistic calibration of the model to US data. Our simulations demonstrate that the optimal tax schedule follows a U-shape, which closely resembles those in the simulations of Saez (2001). Moreover, the monitoring schedule also follows a U-shape. This confirms that the monitoring intensity should indeed be large

[1] In our model, first-best can generally not be obtained, because the penalty function is exogenous. If the government would be able to optimize the penalty function a trivial first-best outcome would result by either raising the penalty to infinity or adjusting the penalty function such that the implicit subsidy on work exactly off-sets the explicit tax on work.

[2] We realize that the assumption of truthful reporting of earnings is not always realistic due to, for example tax evasion and avoidance. This issue has been discussed in, amongst others, Cremer and Gahvari (1996), Schroyen (1997) and Chander and Wilde (1998). In most developed countries, however, firms are required to report gross labor earnings directly to the tax authorities, which prevents underreporting of earnings for a very large fraction of labor earnings (see e.g. Kleven et al., 2011).

when tax distortions on labor supply are large. The simulations demonstrate that the marginal tax rates with monitoring are generally higher than without monitoring. Hence, monitoring always results in more redistribution of income from high- to low-ability individuals, despite the inequality within-ability groups that results due to monitoring and penalizing individuals.

Strikingly, our simulations demonstrate that the optimal tax rate at the bottom end of the income scale is substantially above 100 percent. This implies that the implicit subsidy on labor supply due to monitoring is very effective in reducing the total tax wedge on labor supply at the lower end of the income scale. Indeed, the optimal monitoring probability is close to one at the bottom, but it drops substantially towards middle-income levels. There is a slight increase in the monitoring probability towards the top, as tax rates increase. We conclude from our simulations that monitoring is most important at the bottom of the income distribution. Strongly redistributive governments should therefore optimally employ a high monitoring intensity at the low end of the income scale, for example, via job-search requirements, benefit sanctions, work bonuses, and active labor-market programs. Moreover, our findings suggest that work-dependent tax credits for low-income earners, like those in the UK, Ireland and New Zealand, are indeed part of an optimal redistributive tax policy.

The welfare gains of monitoring are shown to be large. Compared to the optimal non-linear tax schedule without monitoring, monitoring increases total output by 1.35 percent in our baseline simulation. Moreover, the transfer increases by about 4 percent. The monetized welfare gain of monitoring is about 1.4 percent of total output. The optimal monitoring probability does not exceed 20 percent anywhere except at the lower end of the income distribution. In our baseline simulations, the cost of monitoring is a small fraction of average labor earnings. Extensive sensitivity analyses demonstrate that the results are robust to parameter changes in the monitoring technology, on which little empirical evidence exists.

The setup of the paper is the following. The next section gives a brief overview of the related literature. The third section introduces the model and derives the conditions for first- and second-order incentive compatibility. The fourth section derives the optimality conditions for monitoring and redistribution. The fifth section presents the simulations. Finally, the sixth section concludes.

2 Review of the literature

Our model builds upon two strands in the mechanism-design literature. Mirrlees (1971), Diamond (1998) and Saez (2001) develop the theory of the optimal non-linear income tax under the assumption that both hours worked and ability are completely private information, implicitly assuming that verification of either hours worked or ability is infinitely costly. On the other hand, the literature on costly state verification develops principal-agent models where the outcome of a project is a function of both the state of the world and the action of the agent (see, e.g., Mirrlees, 1999, 1976, Holmstrom, 1979, and Townsend, 1979). The outcome is observed, but the action and the state of the world can only be verified through costly monitoring. Monitoring can then improve the ex-ante utility of both the principal and the agent. We apply the theory of costly state verification to the Mirrlees (1971) model and show that monitoring of labor supply can increase welfare significantly.

In a related paper, Armenter and Mertens (2013) study the effect of optimal monitoring of ability types on the optimal tax schedule. They analyze a dynamic model of optimal taxation where the government can use a monitoring technology to establish the ability of an agent. In their model, the monitoring intensity is exogenous, while penalties are endogenous. In equilibrium, individuals do not misreport their ability, and are therefore never penalized. Indeed, the economy is shown to converge to first best in an infinite-horizon setting. We, instead, analyze the case where monitoring is endogenous and penalties are exogenously given. Because penalties are exogenously given, individuals may misreport their ability type in equilibrium. Consequently, our model does not converge to a first-best outcome. An advantage of allowing for an endogenous monitoring intensity is that we do not need to worry about a tax-riot equilibrium in which all individuals misreport their type when they expect other individuals to do the same (Bassetto and Phelan, 2008).

The effect of monitoring has also been studied in the literature on tax evasion and the literature on unemployment insurance. The literature on tax evasion (see, e.g., Allingham and Sandmo, 1972, Sandmo, 1981, Mookherjee and Png, 1989, Slemrod, 1994, Cremer and Gahvari, 1994, 1996, Chander and Wilde, 1998, and Slemrod and Kopczuk, 2002) extends the Mirrlees (1971) framework by allowing individuals to underreport their earned income to the tax authorities.[3] Compared to the standard Mirrlees (1971) model, income taxation is more distortionary, because it not only reduces labor supply, but also increases tax evasion. However, the government can monitor individuals by auditing their tax returns and fine them when they evade taxes. In a two-type economy with non-linear taxation and monitoring Cremer and Gahvari (1994, 1996) show that the welfare-maximizing policy is to levy a positive marginal tax rate on the bottom type and a zero tax rate at the top. All individuals reporting income below a threshold level should be monitored with positive probability. The tax rate and the monitoring schedules are strategic complements for the government, because a higher tax rate induces an increase in tax evasion, thereby increasing the social value of monitoring. In our model the only choice variable of individuals is their labor supply.[4] The monitoring instrument is therefore aimed at measuring hours worked instead of evasion. We extend the literature by considering optimal non-linear tax and monitoring under a continuum of skill types. This allows us to derive an elasticity-based formula for the optimal non-linear tax and monitoring schedule in the spirit of Diamond (1998) and Saez (2001). Moreover, we can determine the shape of non-linear tax and monitoring schedules over the entire income distribution through simulations.

In the literature on unemployment insurance, Ljungqvist and Sargent (1995a,b) study the effect of monitoring on equilibrium employment in welfare states.[5] In their model, unemployed workers may receive a job offer each period. In the absence of monitoring, the benefits induce workers to decline an inefficiently large number of job offers. Monitoring can help raising efficiency by punishing those workers who decline job offers. Simulations using Swedish data

[3] A comprehensive survey of the literature can be found in Slemrod and Yitzhaki (2002).

[4] An alternative interpretation would be that individuals exogenously supply labor, but can use a costly evasion technology.

[5] A large literature exists on optimal unemployment insurance, see Fredriksson and Holmlund (2006) for a survey of this literature. However, this literature typically does not consider monitoring of search effort.

demonstrate that welfare states with large benefits and progressive taxation can have low equilibrium unemployment rates, provided the monitoring probability and sanctions are sufficiently large. In a model of optimal income redistribution with search, Boadway and Cuff (1999) determine the welfare-maximizing monitoring probability and demonstrate that it is increasing in the level of the benefits. Boone and Van Ours (2006) and Boone et al. (2007) develop a search model where the government can actively monitor and sanction job-search effort. They show that monitoring and sanctioning may be more effective in reducing unemployment than cutting the replacement rate. In addition, they show that monitoring may be effective, even when the duration of unemployment benefits is limited. This literature has focused on monitoring the search effort of unemployed workers. We contribute to this literature by studying the effect of monitoring on employed workers.

Finally, we contribute to the literature on optimal non-linear tax simulations (see, for example, Mirrlees, 1971, Tuomala, 1984, Saez, 2001, Brewer et al., 2010 and Zoutman et al., 2013). We show that monitoring can lead to significant improvements in both equity and efficiency.

3 Model

3.1 Households

The setup of our model closely follows Mirrlees (1971). Individuals are heterogeneous in their earnings ability, n, which denotes the productivity per hour worked. Ability is distributed according to cumulative distribution function $F(n)$ with support $[\underline{n}, \overline{n}]$, where $\overline{n}$ could be infinite. The density function is denoted by $f(n)$. Workers are perfect substitutes in production and the wage rate per efficiency unit of labor is constant and normalized to one. n therefore corresponds to the number of efficiency units of labor of each worker. Gross labor income of an individual is the product of his/her ability and his/her labor hours $z_n = nl_n$. Individuals derive utility from consumption c_n and disutility from hours worked l_n.

We introduce the model using a formulation where individuals may receive a work bonuses when they supply a given level of work effort. Then, we demonstrate that a tax implementation where individuals receive a penalty if they fail to meet a given level of working hours is equivalent. The critical part of our analysis is therefore the monitoring of labor supply, not the particular tax implementation through bonuses or penalties.

To fix ideas, we assume that the tax schedule consists of two parts. First, individuals pay income taxes $\hat{T}(z_n)$ based on their earned income z_n. Second, individuals can apply for a working tax credit, $\mathcal{T}$, if their hours worked exceed an exogenously given work requirement, l^*. The work requirement is the same for all individuals.[6] This tax schedule corresponds closely to what we observe in the UK, New Zealand and Ireland – see the remarks in the introduction. Consequently, total tax payments for individuals applying for the tax credit are given by $T(z_n) \equiv \hat{T}(z_n) - \mathcal{T}$. Similarly, tax payments of the individuals who do not apply for the tax credit are simply $\hat{T}(z_n)$.

[6] Zoutman and Jacobs (2014) show that it is straightforward to extend the analysis with a non-linear work requirement that is dependent on ability. However, no additional insights are gained and the analysis becomes more complex as incentive-compatibility constraints will be affected by the work-requirement schedule as well.

We make the technical assumption that all individuals apply for the tax credit. This assumption is nearly without loss of generality as we can always ensure that all individuals apply for the credit by simultaneously adjusting tax payments without the tax credit $\hat{T}(z_n)$ and the tax credit $\mathcal{T}$ by similar amounts.[7] Below we demonstrate that such a policy is in the best interest of the government, since monitoring effectively alleviates the equity-efficiency trade-off and moves the optimal second-best allocation closer to the first-best allocation.

Individuals can misreport their hours worked to the tax authorities, and can therefore claim the tax credit while not satisfying the minimum-hours requirement. The government, however, can operate a *monitoring technology* to verify actual hours worked of an individual applying for the tax credit. $\pi(z_n)$ denotes the probability that an individual with earnings z_n is monitored by the government. $\pi(z_n)$ is also referred to as the *monitoring intensity*. We assume the government receives a perfect signal of the individual's labor supply l_n if the individual is monitored. The government only monitors the individuals that claim the tax credit.

Monitored individuals receive a penalty if they are found to misrepresent their hours worked. The size of the penalty depends on the difference between the required working hours l^* and actual hours l_n worked:

$$P \equiv \begin{cases} P(l^* - l_n) & \text{if} \quad l^* > l_n \\ 0 & \text{if} \quad l^* \leq l_n \end{cases}, \quad P(\cdot), P'(\cdot) \geq 0. \tag{1}$$

We will refer to $P(\cdot)$ as the penalty function. We restrict penalties to be non-negative. The penalty function $P(\cdot)$ is exogenously given and assumed to be continuous and twice differentiable. Penalties increase when individuals are found to supply less labor than the hours requirement ($P'(\cdot) > 0$). Therefore, penalties decrease in hours worked. For a given gross income level z_n penalties thus increase in ability, since higher ability individuals need to supply less hours in order to attain a given gross income level. Finally, we assume that the government does not penalize individuals that applied to the working tax credit and supplied the required minimum amount of hours.

We believe that constraining the penalty function $P(\cdot)$ is realistic for two reasons. First, the legal system practically imposes limitations on the government's ability to use infinite penalties.[8] Second, we assume perfect monitoring as the labor supply of each individual is verified with perfect certainty. If we would more realistically assume that monitoring is imperfect, hard-working individuals could inadvertently be monitored as shirking individuals. Then, we would be able to endogenize both the penalty function and the monitoring function, and infinite penalties would never be socially optimal, see e.g. Stern (1982), Diamond and Sheshinski (1995), and Jacquet (2014). We leave this extension for future research as it would severely complicate our analysis without affecting the main result: monitoring alleviates the equity-efficiency trade-off.

[7] To see this, suppose that the tax credit $\mathcal{T}$ and the tax schedule $\hat{T}(z_n)$ are given. Next, add an arbitrarily large number to both. The incentive to apply for the tax credit then increases, but it does not affect total tax payments $T(z_n)$. Consequently, there always exists a level of the tax credit $\mathcal{T}$ beyond which everyone applies for it.

[8] A more thorough discussion on these issues can be found in Schroyen (1997), Mirrlees (1997), and Mirrlees (1999).

Figure 1 displays an example of a penalty function. As can be seen, the penalty decreases quadratically in labor supply up to $l_n = l^*$, after which it remains constant at 0. Such a penalty function will be used in the simulations later.

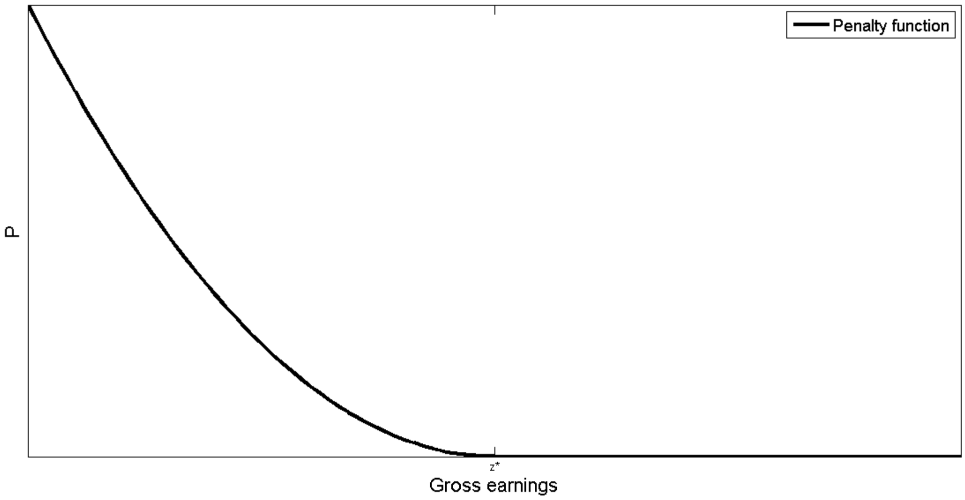

Figure 1: Example of a penalty function

Rather than using work bonuses, the government could, equivalently, use a negative tax credit $\mathcal{T}$, i.e. a work penalty, for all individuals not supplying the optimal level of labor. Individuals would then be required to report to the government whether they met the work requirement to avoid having to pay $\mathcal{T}$, and the government then needs to verify whether these work reports are indeed truthful. The total tax schedule $T(z_n)$ would remain the same. Hence, the particular tax implementation with either work bonuses or non-work penalties is immaterial to our main findings. In the remainder of this paper we will, therefore, focus on determining the optimal total tax schedule $T(z_n)$ *including* the tax credit.

The consumption of an individual who is not monitored is thus given by $c_n^U \equiv z_n - T(z_n)$. The consumption of a monitored (and penalized if hours worked are less than required for the credit) individual is given by $c_n^P \equiv z_n - T(z_n) - P(l^* - l_n)$. Individuals are assumed to maximize expected utility subject to their budget constraints in monitored and unmonitored states. We follow Diamond (1998) by assuming that all individuals have an identical quasi-linear expected utility function:

$$\begin{aligned} u(z_n, n) &\equiv \pi(z_n)c_n^P + (1 - \pi(z_n))c_n^U - v(l_n), \quad v'(\cdot) > 0, \quad v''(\cdot) > 0, \qquad (2) \\ &= z_n - T(z_n) - \pi(z_n)P(l^* - z_n/n) - v(z_n/n), \qquad \forall n, \end{aligned}$$

where we substituted the household budget constraint and $l_n = z_n/n$ in the second line. An important analytical advantage of this quasi-linear-in-consumption utility function is that individuals are risk-neutral.[9] The first term in the first line represents the non-monitoring probability

[9] We could allow for risk-aversion in the utility function. In that case we are only able to solve for the optimal non-linear tax and monitoring schedules if the social welfare function is utilitarian. Intuitively, the problem becomes analytically untractable if the government has a different degree of risk-aversion – which is implied by a non-utilitarian social welfare function – than households have. Without risk aversion, this problem is always

times the consumption of an individual that is not monitored. The second term in the first line is the monitoring probability times the consumption of an individual that is monitored. The last term in the first line is the disutility of labor supply.

Individuals choose the optimal amount of gross income based on their productivity n, the tax function $T(\cdot)$, the monitoring function $\pi(\cdot)$, and the penalty function $P(\cdot)$. An income level z_n is incentive compatible if it maximizes $u(z_n, n)$. The first-order condition for optimal labor supply is given by:

$$v'(z_n/n) = \big(1 - T'(z_n) - \pi'(z_n)P(l^* - z_n/n)\big)n + \pi(z_n)P'(l^* - z_n/n), \qquad \forall n. \tag{3}$$

On the right-hand side, we see that policy drives a wedge between the private and social benefits of labor supply. The *total labor wedge* $\mathcal{W}_n$ is given by:

$$\mathcal{W}_n \equiv \frac{n - v'(z_n/n)}{n} = \underbrace{T'(z_n)}_{\text{explicit tax}} + \underbrace{\pi'(z_n)P(l^* - z_n/n) - \frac{\pi(z_n)}{n}P'(l^* - z_n/n)}_{\text{implicit tax}}, \qquad \forall n. \tag{4}$$

In a laissez-faire equilibrium the right-hand side of eq. (3) equals n and the total labor wedge $\mathcal{W}_n$ is zero. The total labor wedge consists of the explicit marginal tax on labor (T') and the *implicit* marginal tax (subsidy) on labor due to monitoring ($\pi'P - \pi P'$). If $T' + \pi'P - \pi P' > 0$, the redistributive tax and monitoring policy reduces optimal labor supply below the laissez-faire level, and vice versa if it is smaller than zero. The wedge is naturally increasing in the explicit marginal rate T'. Furthermore, it increases in the marginal monitoring probability, π', if penalties are positive, i.e. $P > 0$. π' gives the marginal increase in the monitoring probability as a function of gross earnings. If the monitoring probability increases (decreases) with income, this reduces (increases) the incentive to supply labor, because a higher labor income increases (decreases) the probability of receiving a penalty. Therefore, an increase in the marginal monitoring probability decreases the incentive to supply labor.

Proposition 1 shows that without loss of generality we can assume that expected consumption, $\mathcal{C}(z_n) \equiv z_n - T(z_n) - \pi(z_n)P(l^* - z_n/n)$, is non-decreasing in earnings z_n. Consequently, the total labor wedge $\mathcal{W}_n$ can never be larger than one, i.e. larger than 100 percent.

Proposition 1 *All implementable continuous allocations can be implemented through a continuous non-decreasing expected consumption function $\mathcal{C}(z_n)$, $\forall n$. If $\mathcal{C}(z_n)$ is continuous and differentiable, the wedge $\mathcal{W}_n$ can never exceed* 1.

Proof. The proof directly follows Mirrlees (1971). Let $\tilde{\mathcal{C}}(z)$ be any continuous expected consumption function. The individual maximization problem is given by:

$$z_n = \arg\max_{z_n} \tilde{\mathcal{C}}(z_n) - v(z_n/n), \qquad \forall n. \tag{5}$$

Now consider function $\mathcal{C}(z_n) = \max_{\tilde{z}_n \leq z_n} \tilde{\mathcal{C}}(\tilde{z}_n)$. Clearly, $\mathcal{C}(\cdot)$ is non-decreasing and continuous,

absent and we can allow for any degree of inequality aversion in the social welfare function.

because $\tilde{\mathcal{C}}(\cdot)$ is continuous. Now, consider the maximization problem:

$$\max_{z_n} \mathcal{C}(z_n) - v(z_n/n) = \max_{z_n} \left[\max_{\tilde{z}_n \leq z_n} \tilde{\mathcal{C}}(\tilde{z}_n) \right] - v(z_n/n), \qquad \forall n. \tag{6}$$

Assume z_n is the solution to problem (5). The solution to this second maximization problem must also be z_n. To see this, evaluate $\mathcal{C}(\cdot)$ at z_n: $\mathcal{C}(z_n) = \max_{\tilde{z}_n \leq z_n} \tilde{\mathcal{C}}(\tilde{z}_n)$. Either $\mathcal{C}(z_n) = \tilde{\mathcal{C}}(z_n)$ or $\mathcal{C}(z_n) = \mathcal{C}(\bar{z}_n)$ with $\bar{z}_n < z_n$. In the first case, maximization problems (6) and (5) are equivalent, and hence, they must have the same solution. In the second case, because $v'(\cdot)$ is strictly increasing in z_n, $\bar{z}_n$ must give a higher value to the objective function in eq. (5) than does z_n. Hence, we arrive at a contradiction, because z_n could not have been the solution to problem (5) in the first place. Therefore, without loss of generality we can focus on non-decreasing functions $\mathcal{C}(\cdot)$. Now, suppose $\mathcal{C}(\cdot)$ is differentiable and consider its derivative.

$$\mathcal{C}'(z_n) = 1 - T'(z_n) - \pi'(z_n)P(l^* - z_n/n) + \frac{\pi(z_n)}{n}P'(l^* - z_n/n) = 1 - \mathcal{W}_n, \qquad \forall n. \tag{7}$$

$\mathcal{C}(z_n)$ is non-decreasing if its derivative is greater than or equal to zero: $\mathcal{C}'(z_n) \geq 0 \Leftrightarrow \mathcal{W}_n \leq 1$. ■

Proposition 1 has an intuitive interpretation. Suppose, an individual has a budget constraint such that expected consumption is decreasing in gross income over some interval. Then, this individual will never choose gross income in this interval, because he can work less and consume more, both yielding higher utility. Consequently, the government can never increase social welfare by setting the wedge $\mathcal{W}_n$ above 1. The explicit marginal tax rate $T'(z_n)$, however, could be above 1, provided that monitoring implies a sufficiently large implicit marginal subsidy on work, i.e. $\pi'P - \pi P' < 0$, such that the overall wedge remains below 1. This is the case if the expected penalty increases sufficiently fast in the difference between expected and required labor supply such that $\pi P' > \pi' P$. Therefore, monitoring can improve the incentives for supplying labor.

3.2 Government

The government designs an optimal income tax system and monitoring schedule so as to maximize social welfare, subject to resource and incentive constraints. The government's objective function is a concave sum of individual utilities:

$$\int_{\underline{n}}^{\overline{n}} (1 - \pi(z_n))G(u_n^U) + \pi(z_n)G(u_n^P)\mathrm{d}F(n), \qquad G'(\cdot) > 0, \qquad G''(\cdot) < 0, \tag{8}$$

where $u_n^U \equiv c_n - v(z_n/n)$ and $u_n^P \equiv u_n^U - P(l^* - z_n/n)$ denote the utility levels of the penalized and unpenalized individuals respectively. $G(\cdot)$ is the social welfare function. Redistribution from high-income individuals to low-income individuals raises social welfare because the government is inequality averse. Due to quasi-linearity of private utility there is no social desire to redistribute income if the social welfare function is utilitarian. The government is constrained in its ability to redistribute income, because the ability of individuals is private information. However, the government can infer the ability of an individual from costly monitoring activities or it can induce self-selection by sacrificing on redistribution.

The total cost of monitoring is given by:

$$\int_{\underline{n}}^{\overline{n}} k(\pi(z_n))\mathrm{d}F(n), \qquad k(0)=0, \qquad k'(\cdot), k''(\cdot)>0. \tag{9}$$

The cost of monitoring is increasing and convex in the monitoring probability π. Since there is a perfect mapping between skill n and labor earnings z_n, we can also write $\pi(\cdot)$ as a function of the skill level n, where we use the short-hand notation $\pi(z_n) = \pi_n$. However, $\pi'(z_n) \equiv \frac{\mathrm{d}\pi_n}{\mathrm{d}z_n}$ always denotes the derivative of monitoring with respect to gross earnings.

The economy's resource constraint implies that total labor earnings equal aggregate consumption plus monitoring costs:

$$\int_{\underline{n}}^{\overline{n}} z_n \mathrm{d}F(n) = \int_{\underline{n}}^{\overline{n}} \big((1-\pi(z_n))c_n^U + \pi(z_n)c_n^P + k(\pi(z_n))\big)\mathrm{d}F(n). \tag{10}$$

By defining unpenalized consumption as $c_n \equiv c_n^U = c_n^P + P(z_n, n)$, we can write for aggregate consumption:

$$\int_{\underline{n}}^{\overline{n}} \big((1-\pi(z_n))c_n^U + \pi(z_n)c_n^P\big)\mathrm{d}F(n) = \int_{\underline{n}}^{\overline{n}} (c_n - \pi(z_n)P(l^* - z_n/n))\mathrm{d}F(n). \tag{11}$$

Hence, using eq. (11) the economy's resource constraint (10) can be rewritten as:

$$\int_{\underline{n}}^{\overline{n}} (z_n + \pi(z_n)P(l^* - z_n/n))\mathrm{d}F(n) = \int_{\underline{n}}^{\overline{n}} (c_n + k(\pi(z_n)))\mathrm{d}F(n). \tag{12}$$

We do not need to consider the government budget constraint, since it is automatically implied by Walras' law if the individual budget constraints and the economy's resource constraint are satisfied.

The timing of the model is as follows:

1. The government announces the exogenously given penalty function, as well as the optimal non-linear income tax and monitoring schedules.

2. Each individual optimally chooses hours worked.

3. The government observes the labor incomes chosen by each individual and taxes income and monitors individuals accordingly. The government penalizes all monitored individuals according to the penalty function.

4. Individuals receive utility from consumption and leisure.

By the revelation principle any indirect mechanism can be replicated with an incentive-compatible direct mechanism (Myerson, 1979; Harris and Townsend, 1981). Therefore, we can find the optimal second-best allocation by maximizing welfare subject to feasibility and incentive-compatibility constraints. We can decentralize the optimal second-best allocation as a competitive market outcome through the non-linear tax and monitoring schedules.

3.3 First-order incentive compatibility

By using the envelope theorem we can derive a differential equation for the indirect utility function u_n which is a necessary condition for incentive compatibility. The next subsection derives the conditions under which the first-order condition is indeed sufficient. The incentive compatibility constraint is found by totally differentiating eq. (2) with respect to n:

$$\frac{\mathrm{d}u_n}{\mathrm{d}n} = \frac{\partial u(z_n, n)}{\partial n} + \frac{\partial u(z_n, n)}{\partial z_n}\frac{\mathrm{d}z_n}{\mathrm{d}n} = \frac{l_n}{n}(v'(l_n) - \pi(z_n)P'(l^* - l_n)), \quad \forall n, \tag{13}$$

where $\frac{\partial u(z_n,n)}{\partial z_n} = 0$ due to the individual's first-order condition in eq. (3). Thus, if the optimal allocation satisfies eq. (13), individuals' first-order conditions for utility maximization are also satisfied.

3.4 Second-order incentive compatibility

Without further restrictions we cannot be certain that the optimal allocation derived under the first-order incentive compatibility constraint (13) is also implementable. An implementable allocation should satisfy additional requirements to ensure that the first-order approach also respects the second-order conditions for utility maximization. The next Lemma summarizes the requirements for second-order incentive compatibility.

Lemma 1 *Second-order conditions for utility maximization are satisfied under the first-order approach if the following conditions hold at the optimal allocation for all n:*
i) single-crossing conditions on the utility and penalty functions are satisfied:

$$\frac{\partial(v'(l_n)/n)}{\partial n} - \frac{\pi(z_n)P'(l^* - l_n)}{n^2}\left(\varepsilon_n^P - 1\right) + \frac{l_n \pi'(z_n)}{n}P'(l^* - l_n) \leq 0, \tag{14}$$

where $\varepsilon_n^P \equiv \frac{P''(l^* - l_n)l_n}{P'(l^* - l_n)}$ is the elasticity of the penalty function,
ii) z_n is non-decreasing in ability:

$$\frac{\mathrm{d}z_n}{\mathrm{d}n} \geq 0. \tag{15}$$

Proof. The second-order condition for the utility-maximization problem (2) is given by:

$$\frac{\partial^2 u(z_n, n)}{\partial z_n^2} \leq 0, \quad \forall n. \tag{16}$$

This second-order condition can be rewritten in a number of steps. Totally differentiating the first-order condition (3) gives:

$$\frac{\partial^2 u(z_n, n)}{\partial z_n^2}\frac{\mathrm{d}z_n}{\mathrm{d}n} + \frac{\partial^2 u(z_n, n)}{\partial z_n \partial n} = 0, \quad \forall n. \tag{17}$$

Substitution of this result in eq. (16) implies that the second-order condition is equivalent to:

$$\frac{\partial^2 u(z_n,n)}{\partial z_n \partial n}\left(\frac{\mathrm{d}z_n}{\mathrm{d}n}\right)^{-1} \geq 0, \qquad \forall n. \tag{18}$$

Differentiating the first-order condition (3) with respect to n and substituting the result yields:

$$\left(\frac{\partial(v'/n)}{\partial n} + \frac{\pi P'}{n^2}\left(1 - \frac{P'' l_n}{P'}\right) + \frac{l_n \pi'}{n}P'\right)\left(\frac{\mathrm{d}z_n}{\mathrm{d}n}\right)^{-1} \leq 0, \qquad \forall n. \tag{19}$$

The inequality holds if all conditions of the Lemma are satisfied. ■

The single-crossing condition and the monotonicity of gross earnings are well-known from the Mirrlees model (Mirrlees, 1971; Ebert, 1992). The single-crossing condition ensures that – at the same consumption-earnings bundle – individuals with a higher ability have a larger marginal willingness to work. In our model, the single-crossing condition contains three elements. The first is the standard Spence-Mirrlees condition on the utility function, i.e. $\frac{\partial(v'(l_n)/n)}{\partial n} < 0$. If this term is negative, the marginal disutility of work for individuals with a higher ability level is lower. Most utility functions considered in the literature exhibit this property, including our own. The sign of the second term is determined by $\pi P'(\varepsilon_n^P - 1)/n^2$. Intuitively, it is more costly for high-ability individual to mimick a low-ability individual if $\frac{\partial(P'/n)}{\partial n} > 0$. That is, the marginal penalty of earning a lower income increases with ability. $\frac{\partial(P'/n)}{\partial n} > 0$ is equivalent to $\varepsilon_n^P > 1$. Intuitively, if the elasticity of the marginal penalty is larger, penalties become increasingly more severe for high-ability individuals mimicking low-ability individuals. The third term, $l_n \pi' P'/n$, concerns the slope of the monitoring schedule, and its sign is determined by the monitoring schedule, since $P' > 0$. If the marginal monitoring probability decreases in gross earnings ($\pi' < 0$) individuals will work harder in order to decrease the probability of being monitored and penalized. The sign of the last term is determined by the endogenous monitoring schedule. Hence, high-ability individuals can be induced to self-select into higher income-consumption bundles, unless the monitoring probability increases too fast with ability.

A second requirement to induce self-selection is that gross earnings are indeed increasing with ability at the optimal schedule. Consequently, a tax schedule that provides higher income to higher ability individuals induces self-selection of higher ability types into higher income-consumption bundles. In the remainder we assume that all the conditions derived in Lemma 1 hold at the optimal allocation. In our simulations, we check the second-order sufficiency conditions ex-post and we always confirm that they are respected.

4 Optimal second-best allocation with monitoring

The optimization problem with monitoring can be specified formally as:

$$\max \int_{\underline{n}}^{\overline{n}} [(1-\pi_n)G(u_n + \pi_n P(l^* - z_n/n)) + \pi_n G(u_n - (1-\pi_n)P(l^* - z_n/n))] f(n) \mathrm{d}n, \tag{20}$$

$$\text{s.t.} \quad \int_{\underline{n}}^{\overline{n}} [z_n + \pi_n P(l^* - z_n/n) - c_n - k(\pi_n)] f(n) \mathrm{d}n = 0, \tag{21}$$

$$\frac{\mathrm{d}u_n}{\mathrm{d}n} = \frac{l_n}{n}(v'(l_n) - \pi_n P'(l^* - z_n/n)), \tag{22}$$

$$u_n = c_n - \pi_n P(l^* - z_n/n) - v(z_n/n), \qquad \forall n, \tag{23}$$

$$\pi_n \geq 0, \qquad \forall n, \tag{24}$$

where utility of unpenalized and penalized individuals is, respectively, written as $u_n^U = u_n + \pi_n P(l^* - z_n/n)$ and $u_n^P = u_n - (1-\pi_n)P(l^* - z_n/n)$. The final constraint assumes that the probability of monitoring cannot be smaller than zero. We assume that the cost of monitoring is sufficiently large to ensure that the constraint $\pi_n \leq 1$ is never binding.

The Hamiltonian function for this problem is given by:

$$\begin{aligned} \mathcal{H} \equiv\ & [(1-\pi_n)G(u_n + \pi_n P(l^* - z_n/n)) + \pi_n G(u_n - (1-\pi_n)P(l^* - z_n/n))] f(n) \\ & + \lambda [z_n + \pi_n P(l^* - z_n/n) - c_n - k(\pi_n)] f(n) \\ & - \frac{\theta_n z_n}{n^2} \left[v'(z_n/n) - \pi_n P'(l^* - z_n/n) \right] \\ & + \mu_n [u_n - c_n + v(z_n/n) + \pi_n P(l^* - z_n/n)] + \eta_n \pi_n, \end{aligned} \tag{25}$$

c_n, z_n and π_n are the control variables. u_n is the state variable with θ_n as its associated co-state variable. μ_n is the Lagrange multiplier for the definition of utility. λ is the Lagrange multiplier of the economy's resource constraint. η_n is the Kuhn-Tucker multiplier of the non-negativity constraint on π_n. The necessary first-order conditions are given by:

$$\frac{\partial \mathcal{H}}{\partial c_n} = 0 : -\lambda f(n) - \mu_n = 0, \qquad \forall n, \tag{26}$$

$$\begin{aligned} \frac{\partial \mathcal{H}}{\partial z_n} = 0 : & \left[(1-\pi_n)\pi_n \frac{P'(\cdot)}{n} (G'(u_n^P) - G'(u_n^U)) + \lambda \left(1 - \frac{\pi_n P'(\cdot)}{n} \right) \right] f(n) \\ & - \theta_n \left(\frac{v'(\cdot) + z_n v''(\cdot)/n - \pi_n (P'(\cdot) - z_n P''(\cdot)/n)}{n^2} \right) + \mu_n \left(\frac{v'(\cdot) - \pi_n P'(\cdot)}{n} \right) = 0, \qquad \forall n, \end{aligned} \tag{27}$$

$$\begin{aligned} \frac{\partial \mathcal{H}}{\partial \pi_n} = 0 : & \left[-G(u_n^U) + (1-\pi_n)P(\cdot)G'(u_n^U) + \pi_n P(\cdot) G'(u_n^P) + G(u_n^P) - \lambda \big(k'(\pi_n) - P(\cdot) \big) \right] f(n) \\ & + \frac{z_n \theta_n}{n^2} P'(\cdot) + \mu_n P(\cdot) + \eta_n = 0, \qquad \forall n, \end{aligned} \tag{28}$$

$$\frac{\partial \mathcal{H}}{\partial u_n} = \frac{\mathrm{d}\theta_n}{\mathrm{d}n} : \frac{\mathrm{d}\theta_n}{\mathrm{d}n} = \left[(1-\pi_n)G'(u_n^U) + \pi_n G'(u_n^P) \right] f(n) + \mu_n, \qquad \forall n, \tag{29}$$

$$\eta_n \pi_n = 0, \qquad \eta_n \geq 0, \qquad \pi_n \geq 0, \qquad \forall n, \tag{30}$$

$$\lim_{n \to \underline{n}} \theta_n = \lim_{n \to \bar{n}} \theta_n = 0. \tag{31}$$

Compared to the analysis of Mirrlees there are two new first-order conditions. Eq. (28) states the optimal monitoring condition, and eqs. (30) state the Kuhn-Tucker conditions for the non-negativity constraint on π_n.

4.1 Optimal wedge on labor

Proposition 2 gives the conditions for optimal income redistribution.

Proposition 2 *The optimal net marginal wedge on labor $\mathcal{W}_n$ at each ability level satisfies:*

$$\frac{\mathcal{W}_n}{1-\mathcal{W}_n} = A_n B_n C_n - D_n, \qquad \forall n, \tag{32}$$

where

$$A_n \equiv 1 + \frac{1}{\varepsilon_n} + \pi_n \frac{P'(\cdot)}{v'(\cdot)}\left(\varepsilon_n^P - 1\right), \tag{33}$$

$$B_n \equiv \frac{\int_n^{\overline{n}}(1-g_m)f(m)\mathrm{d}m}{1-F(n)}, \tag{34}$$

$$C_n \equiv \frac{1-F(n)}{nf(n)}, \tag{35}$$

$$D_n \equiv \frac{P'(\cdot)}{v'(\cdot)}\sigma_n, \tag{36}$$

$\sigma_n \equiv \frac{(1-\pi_n)\pi_n(G'(u_n^P)-G'(u_n^U))}{\lambda} > 0$ *is a measure for the welfare cost of inequality between penalized and unpenalized individuals at ability level n, $\varepsilon_n \equiv \left(\frac{l_n v''(l_n)}{v'(l_n)}\right)^{-1} > 0$ is the compensated wage elasticity of labor supply, and $g_n \equiv \frac{(1-\pi_n)G'(u_n^U)+\pi_n G'(u_n^P)}{\lambda} > 0$ is the average, marginal social value of income, expressed in money units, for individuals at ability level n.*

Proof. Integrate eq. (29) using a transversality condition from eq. (31). If follows that $\theta_n = \lambda \int_n^{\overline{n}}(1-g_m)f(m)\mathrm{d}m$. Substitute this result and eq. (26) in eq. (27), use eq. (4), and simplify to obtain the Proposition. ■

The A_n-term is related to the inverse of the efficiency cost of the labor wedge at income level z_n. The second term in A_n, $1/\varepsilon_n$, is the inverse of the labor-supply elasticity and it enters because the deadweight loss of the wedge increases in the labor-supply elasticity. The third term represents the efficiency gains of monitoring. As noted in before, penalties are useful in seperating high- and low-ability individuals if the elasticity of the penalty function ε^P is larger than 1. Penalties are more effective if the elasticity increases. The latter effect is stronger if the monitoring intensity π is larger. Finally, penalties are better at providing work incentives if the marginal penalty becomes relatively more important relative to the marginal disutility of labor, $\frac{P'}{v'}$. Hence, in comparison to the optimal wedge without monitoring (cf. Diamond, 1998; Saez, 2001) monitoring reduces the efficiency cost of taxation provided the elasticity of the penalty function is larger than 1.

The B_n-term measures the equity gain of an increase in the labor wedge at income level z_n. The first term, 1, captures the revenue gain of a larger marginal labor wedge at n, such that individuals with an income level above z_n pay one unit of extra income tax. The welfare loss of extracting one unit of income from the individuals above n is g_m for all individuals $m \geq n$. Therefore, $\int_n^{\overline{n}} (1 - g_m)\mathrm{d}F(m)$ measures the redistributional gain of the labor wedge at n. The B_n-term is not directly affected by monitoring. Since welfare weights g_n are always declining with income, B_n always rises with income, see also Diamond (1998).

C_n is the inverse relative hazard rate of the skill distribution. Its numerator is the fraction of the population whose net income is decreased by increasing the wedge and its denominator captures the size of the tax base that is distorted by the wedge. Hence, the numerator in C_n gives weights to average equity gains in B_n and the denominator to average efficiency losses in A_n – as in the Mirrlees model without monitoring. The numerator of C_n always declines with income; there are fewer individuals paying marginal taxes if the tax rate is increased at a higher income level. Hence, for a given B_n the total distributional benefits of raising the labor wedge fall as the income level rises. For a unimodal skill distribution the denominator of C_n always increases with income before the mode, since both n and $f(n)$ are rising. Thus, labor wedges always decrease with income before modal income. After the mode, $f(n)$ falls, although n continues to rise with income. Hence, it depends on the empirical distribution of n whether C_n rises or falls with income after modal income. For most empirical distributions, C_n appears to rise after the mode and converges to a constant at the top. See also Diamond (1998), Saez (2001) and Zoutman et al. (2013).

Finally, D_n measures the welfare loss associated with within-ability inequality. Earnings at n decrease if the labor wedge increases. Therefore, the penalty at n increases, which in turn increases inequality between monitored and unmonitored individuals. σ_n measures the marginal welfare cost of this within-ability inequality. The effect of a wedge on within-ability inequality is increasing in the relative importance of the penalty function with respect to the marginal disutility of labor (expressed in monetary units), $\frac{P'}{v'}$. D_n increases in the monitoring probability for $\pi_n < .5$ because the within-ability variance of monitoring is increasing in π_n for $\pi_n < .5$. Finally, D_n is increasing in the concavity of the welfare function, because the difference in welfare weights between penalized and unpenalized individuals, $\frac{G'(u_n^p) - G'(u_n^u)}{\lambda}$, is larger if the government is more inequality averse.

We can summarize the impact of monitoring on optimal labor wedges as follows. Monitoring decreases the efficiency cost of setting a higher labor wedge, but introduces within-ability inequality. Therefore, the total effect of monitoring on the optimal labor wedge is theoretically ambiguous. Our simulations below demonstrate that the efficiency gains of monitoring outweigh the distributional loss due to inequality between monitored and non-monitored individuals.

We can derive the non-linear tax function, which implements the second-best allocation as the outcome of decentralized decision making in a competitive labor market. Substituting eq. (3) into eq. (32) yields:

$$\frac{T'(z_n) + \pi'(z_n)P(l^* - z_n/n) - \pi(z_n)P'(l^* - z_n/n)/n}{1 - T'(z_n) - \pi'(z_n)P(l^* - z_n/n) + \pi(z_n)P'(l^* - z_n/n)n} = A_n B_n C_n - D_n, \quad \forall n. \tag{37}$$

Thus, when we know the optimal monitoring schedule $\pi(z_n)$, this equation implicitly defines the optimal non-linear income tax function $T(z_n)$.

4.2 Optimal monitoring

The next proposition derives the optimal monitoring schedule.

Proposition 3 *The optimal level of monitoring at each ability level follows from:*

$$k'(\pi_n) + \Delta_n - g_n P(\cdot) \geq \left(\frac{\frac{\mathcal{W}_n}{1-\mathcal{W}_n} + D_n}{A_n} \right) l_n P'(\cdot) \qquad \forall n, \tag{38}$$

where $\Delta_n \equiv \frac{G(u_n^U)-G(u_n^P)}{\lambda}$ is the welfare difference between a penalized and an unpenalized individual expressed in money units. If $\pi_n > 0$, the equation holds with equality.

Proof. Substitute eq. (26) into eq. (28), rearrange terms, employ the definitions for B_n and C_n, and use the fact that $\eta_n \geq 0$. Finally, substitute eq. (32) for $B_n C_n$ to obtain the expression. By eq. (30) η_n only equals zero if $\pi_n > 0$ and therefore the equation holds with equality if $\pi_n > 0$. ■

The first term on the left-hand side in condition (38) is the marginal cost of raising the monitoring intensity. The second and third terms on the left-hand side jointly represent the welfare effect of a compensated increase in the monitoring probability. That is, the welfare effect of an increase in the monitoring probability, while keeping expected utility at skill level n unchanged. The second term represents the uncompensated, direct welfare loss of an increase in the monitoring probability. If the monitoring probability increases, there will be more penalized and less unpenalized individuals. Therefore, the loss is equal to the welfare difference between penalized and unpenalized individuals. The third term represents the welfare gain associated with the compensation to keep expected utility unchanged if the monitoring probability is increased. The compensation at ability level n requires a transfer of P and its associated welfare effect is thus given by $g_n P$. In Lemma 2 we derive how the compensated welfare effect of monitoring changes with the monitoring probability for given levels of utility in monitored and unmonitored states.

Lemma 2 *The compensated welfare effect of the monitoring probability is decreasing in π_n, positive if $\pi_n = 0$ and negative if $\pi_n = 1$ for given levels of utility in penalized and unpenalized states.*

Proof. By a first-order Taylor expansion around u_n^U we can write Δ_n as:

$$\Delta_n = \frac{G(u_n^U) - G(u_n^P)}{\lambda} = \frac{G'(u_n^U)(u_n^U - u_n^P) + R(P)}{\lambda} = \frac{G'(u_n^U)P}{\lambda} + R(P). \tag{39}$$

where $R(P)$ is a second-order remainder term. Similarly, a first-order Taylor expansion around u^P yields:

$$\Delta_n = \frac{G'(u_n^P)P}{\lambda} - \hat{R}(P), \tag{40}$$

where $\hat{R}(P)$ is again a second-order remainder term. By concavity of G both remainder terms are positive for $P > 0$: $R(P), \hat{R}(P) > 0$. Now multiply eq. (39) with $(1 - \pi_n)$ and eq. (40) with

π_n and add them to find:

$$\Delta_n - g_n P = (1-\pi_n)R(P) - \pi_n \hat{R}(P). \tag{41}$$

The right-hand side gives the compensated welfare effect of the monitoring probability, which is decreasing in π_n, always positive if $\pi_n = 0$, and always negative if $\pi_n = 1$, ceteris paribus. ■

The right-hand side of eq. (38) represents the marginal benefits of monitoring. The benefits of monitoring increase in the marginal penalty $P'(\cdot)$, which can be interpreted as the power of the penalty function. In addition, the marginal benefits of monitoring increase if labor-supply distortions are larger, i.e. if the labor wedge $\frac{\mathcal{W}_n}{1-\mathcal{W}_n}$ is larger or if the efficiency cost of taxation is larger, as captured by $1/A_n$. The benefits of monitoring also increase in within-ability inequality D_n. Intuitively, as more monitoring leads to larger labor supply, the expected penalty decreases. Hence, monitoring helps to reduce within-ability inequality.

From Proposition 3 it follows that the government does not engage in monitoring if and only if (evaluated at a no-monitoring equilibrium with $\pi_n = 0$):

$$k'(0) + \Delta_n - g_n P(\cdot) \geq \left(\frac{\frac{\mathcal{W}_n}{1-\mathcal{W}_n} + D_n}{A_n} \right) l_n P'(\cdot), \qquad \forall n. \tag{42}$$

That is, if the marginal cost of monitoring are higher than the marginal benefits for all types. By evaluating eq. (32) at $\pi_n = 0$ it easily follows that the optimal allocation is the allocation derived in Mirrlees (1971). Mirrlees (1971) is thus a special case of our model where monitoring is prohibitively expensive.

4.3 Boundary results

In the next Proposition we derive the optimal wedge and monitoring probability at the bottom and the top of the ability distribution.[10]

Proposition 4 *If the income distribution is bounded at the top, $\overline{n} < \infty$, the optimal wedge and monitoring probabilities at the extremes are:*

$$\mathcal{W}_{\underline{n}} = \mathcal{W}_{\overline{n}} = \pi_{\underline{n}} = \pi_{\overline{n}} = 0. \tag{43}$$

If the penalties are zero at the first-best levels of earnings, marginal tax rates are also zero at the endpoints:

$$T'(z_{\underline{n}}) = T'(z_{\overline{n}}) = 0. \tag{44}$$

Proof. From eq. (32) it follows that $\left(\frac{\mathcal{W}_n}{1-\mathcal{W}_n} + D_n\right)/A_n = B_n C_n$. The transversality conditions (31) imply $B_{\underline{n}} C_{\underline{n}} = B_{\overline{n}} C_{\overline{n}} = 0$. At the extremes, the optimal monitoring condition (38), therefore simplifies to: $\Delta_n - g_n P + k'(\pi_n) \geq 0$. Evaluate this expression at $\pi = 0$:

$$\Delta_n - g_n P + k'(0) = R(P) + k'(0) \geq 0. \tag{45}$$

[10] Due to the absence of income effects in labor supply, bunching at zero labor earnings is not an issue in deriving the boundary results, see also Seade (1977).

where $R(P) > 0$ is a second-order remainder term, and the second step follows from Lemma 2. The condition is always satisfied at $\pi_n = 0$. Hence, $\pi_n = 0$ is optimal at the extremes. The optimal wedges in eq. (32) at the extremes are zero, because the product $B_n C_n$ is zero by the transversality conditions, and D_n is zero, since $\pi_n = 0$. If the penalties are zero when labor supply is at a first-best level, then $P(\cdot) = 0$ at the endpoints, since labor supply is undistorted if the wedges are zero. Using $\pi_n = P(\cdot) = 0$ in eq. (4) then demonstrates that $\mathcal{W}_{\underline{n}} = \mathcal{W}_{\overline{n}} = T'(z_{\underline{n}}) = T'(z_{\overline{n}}) = 0$. ■

Proposition 4 establishes that the optimal zero wedge at the bottom and top of the model without monitoring carries over to the model with monitoring (Sadka, 1976; Seade, 1977). Intuitively, the wedge at n redistributes income from individuals above n to the government, and, hence indirectly to individuals below n. There are no individuals above $\overline{n}$ and no individuals below $\underline{n}$. Therefore, there are no benefits associated to a positive wedge at these points of the ability distribution. However, the wedge does distort the labor-supply decision. Hence, the optimal wedge must be zero. Because the wedge is zero, there is no efficiency gain of monitoring. As a result, the optimal monitoring probability is also zero.

However, marginal tax rates at the endpoints do not necessarily need to be zero. This critically depends on the penalty function. In particular, if the marginal monitoring probability is non-zero at the end-points ($\pi'(z_n) \neq 0$) and the expected penalty is positive, marginal tax rates at the endpoints have to be non-zero in order to compensate for the distortion caused by the change in monitoring intensity. In particular, marginal tax rates at the endpoints should be positive (negative) if $\pi'(z_n)P(\cdot) < 0$ (> 0). Only if penalties are zero if earnings at the end-points correspond to the first-best levels of earnings, then marginal tax rates at the end-points are zero as well.

5 Simulations

In this section we use numerical simulations to establish the shape of the optimal tax and monitoring schedules. The simulations require four main ingredients: the ability distribution, the individual preferences, the social preferences and the monitoring technology. First, we use the skill distribution from Mankiw et al. (2009). The hourly wage is used as a proxy for earnings ability. We follow Mankiw et al. (2009) by assuming that wage rates follow a log-normal distribution, which is extended with a Pareto distribution for the top tail of the wage distribution. In addition, we assume that there is an exogenous fraction of 5 percent disabled individuals having zero earning ability ($\underline{n} = 0$), which is also based on Mankiw et al. (2009). The earnings distribution is estimated from March 2007 CPS data. This resulted in a mean log-ability of $m = 2.76$ and a standard deviation of log ability of $s = 0.56$. The Pareto tail starts at the top 1 percent of the earnings distribution and features a Pareto parameter of $\alpha = 2$. The latter is in accordance with estimates of Saez (2001).

Second, a description of individual preferences is needed. For the purpose of our simulations it is convenient if optimal labor supply is restricted between zero and one. In addition, we follow the literature in assuming a constant elasticity of taxable income (see, e.g., Saez, 2001). The following utility function abides both features:

$$u(c_n, l_n) = c_n - \frac{n}{1+1/\varepsilon} l_n^{1+1/\varepsilon}, \qquad \varepsilon > 0. \tag{46}$$

ε is the (un)compensated elasticity of taxable income. This utility function has been used in Brewer et al. (2010). We follow the empirical literature estimating the elasticity of taxable income (see, e.g., Saez et al., 2012) and set $\varepsilon = 0.25$.

The third ingredient is the social welfare function. We assume an Atkinson social-welfare function featuring a constant elasticity of relative inequality aversion β:

$$\begin{aligned} G(u_n) &= \frac{u_n^{1-\beta}}{1-\beta}, \qquad \beta \geq 0, \quad \beta \neq 1, \\ G(u_n) &= \ln(u_n), \qquad \beta = 1. \end{aligned} \tag{47}$$

The utilitarian objective is obtained by assuming $\beta = 0$. A Rawlsian social welfare function results if $\beta \to \infty$. The baseline assumes a moderately redistributive government with $\beta = 0.99 \approx 1$. In the robustness analysis we also consider less redistributive governments ($\beta = 0.5$) and more redistributive governments ($\beta = 1.5$).

Finally, we need to make specific assumptions on the monitoring technology and the penalty function. Unfortunately, no empirical evidence is available that guides us to calibrate these functions. However, our theoretical model provides some restrictions on the choice of the functions. Also, we perform robustness checks on the parameter choices we have made for these functions. The cost of monitoring should be increasing and convex in the monitoring intensity π. We assume that the cost of monitoring is quadratic:

$$k(\pi_n) = \frac{\kappa}{2}\pi_n^2, \qquad \kappa > 0, \tag{48}$$

where κ is a cost parameter indicating the marginal cost of a higher monitoring probability. In the baseline we assume $\kappa = 1$. In the robustness analysis we vary κ between 0.25 and 4. We provide economic justification for these parameter values by showing in the robustness analysis that the change in the monitoring probability induced by the different values of κ is relatively large. In addition, we show that in our calibration total monitoring cost is a small, but significant fraction of total income earned in the economy.

In our baseline simulations, we assume that the reference level of labor supply l^* equals:

$$l^* \equiv \begin{cases} 1 & \forall n > \underline{n} \\ 0 & \text{if} \quad \underline{n} = 0 \end{cases}. \tag{49}$$

Therefore, all working individuals, i.e. those with positive earning ability ($n > \underline{n}$), should supply the first-best level of hours to avoid being penalized. Individuals that cannot work ($n = \underline{n} = 0$) are not required to work. In the robustness analysis we analyze the case where hours required is only half of the first-best labor supply, i.e. $l_n^* = 0.5$ for $n > \underline{n}$. We will demonstrate that such changes result in significant changes in the optimal monitoring intensity.

Table 1: Calibration for simulations

Parameter	Description	Base value	High value	Low value
m	Mean log ability	2.76	N/A	N/A
s	Standard deviation log ability	0.56	N/A	N/A
α	Pareto parameter	2.00	N/A	N/A
d	Fraction of disabled individuals	0.05	N/A	N/A
ε	Compensated elasticity	0.25	N/A	N/A
r	Government revenue as fraction of GDP	0.10	N/A	N/A
κ	Cost of monitoring	1.00	0.25	4.00
p	Penalty parameter	3.00	5.00	1.00
l^*	Reference labor effort	1.00	N/A	0.50
β	Relative inequality aversion	1.00	1.50	0.50

We assume that the penalty function is quadratic in labor hours l_n and is given by:[11]

$$P = \frac{p}{2}(\max\{0, l^* - l_n\})^2, \quad p > 0, \tag{50}$$

where p is a parameter determining the severity of the penalty. The penalty is a function of the reference level of labor hours l^*. All individuals facing a positive labor wedge supply fewer hours work than socially desired. Therefore, they are subject to a penalty when monitored, and increasingly so if their hours worked deviate more from the reference level of hours. Consequently, monitoring will be effective in boosting labor supply at all income levels. In the baseline we set $p = 3$. In the robustness checks we employ values of $p = 1$ and $p = 5$.

The government-revenue requirement is exogenous and set to 10 percent of labor earnings in the baseline specification without monitoring, following Tuomala (1984) and Zoutman et al. (2013). The choices for all the parameters can be found in Table 1.

In the table, the first column on the right-hand side gives the base value of the parameter. In addition, we perform robustness checks with high and low parameter values for the welfare function, all parameters of the penalty function and all parameters of the monitoring technology.[12]

5.1 Results

Figure 2 gives the optimal wedge, tax and monitoring schedules as a function of yearly income in US dollars. The fat solid line represents the optimal tax schedule with monitoring. The dashed line is the optimal tax schedule without monitoring. The circled line is the optimal total labor wedge with monitoring. And, the thin solid line is the optimal monitoring schedule. Recall that the optimal tax schedule coincides with the optimal labor wedge if there is no monitoring.

As can be seen, the optimal labor wedge follows a U-shape both with and without monitoring.

[11] Note that with this specification of the penalty function the elasticity ε^P is not unambiguously larger than 1 so that violations of second-order conditions might occur, see Lemma 1. In none of our reported simulations this is the case, however.

[12] The numerical procedure we use to solve for the optimal allocation is a so-called shooting method. We solve the differential equations (13) and (29) numerically for given initial values $\theta_{\underline{n}}$, $u_{\underline{n}}$, and λ. Subsequently, we 'shoot' for initial values until we meet boundary conditions (12) and (31). The wedge, tax, and monitoring schedule can be found using eq. (37). A more detailed explanation of the numerical procedure can be found in the Appendix.

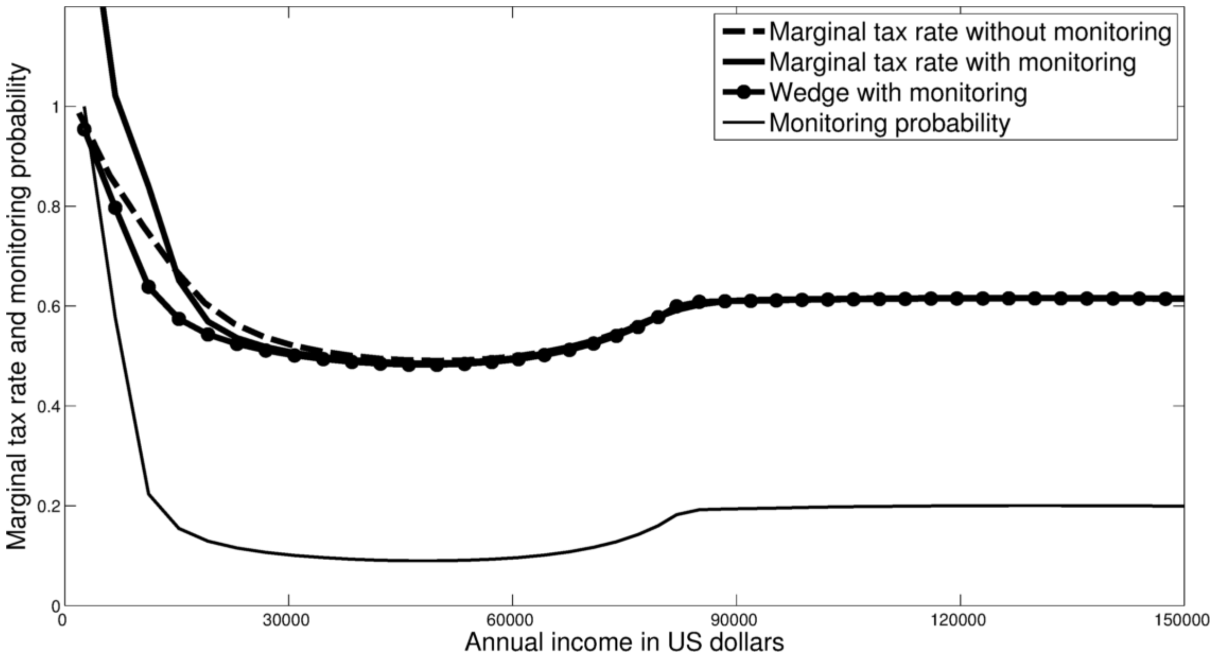

Figure 2: The optimal wedge, tax and monitoring schedules in the baseline scenario. Baseline parameter values of the model can be found in Table 1.

Marginal wedges are extremely large at the bottom of the labor market, relatively small for middle-income levels and somewhat higher at the top. The shape of these schedules is largely explained by the B_n and C_n terms in eq. (32). The B_n-term increases with income as the welfare loss of taxing away one unit of income unit from individuals above z_n decreases in z_n, see our previous discussion. The C_n-term follows a U-shape. At the bottom of the earnings distribution, the density of tax payers is small, and hence, efficiency costs of marginal taxes are low. In addition, the redistributional benefits of a higher marginal tax rate are large as it is paid by almost the entire population. Towards middle-income levels, the efficiency cost increases as the population density increases, whereas the redistributive benefits decrease as fewer individuals are paying a higher tax rate. After modal income marginal tax distortions decline more rapidly than distributional benefits of marginal taxes, hence marginal taxes increase. These results are entirely in line with previous simulations performed in e.g. Saez (2001), Brewer et al. (2010), and Zoutman et al. (2013).

Recall from the previous section, that the effect of monitoring on the labor wedge was theoretically ambiguous. However, in our simulations we see that the efficiency gain of monitoring in reducing labor distortions outweighs the distributional cost of raising within-skill group inequality. The optimal monitoring schedule also follows a U-shape. In eq. (38) the labor wedge determines the shape of the monitoring schedule, as the other elements of the monitoring schedule do not exhibit a very strong dependence on income. The monitoring intensity decreases very steeply at the bottom of the income distribution. This gives individuals a strong incentive to increase their labor supply. At middle-income levels the monitoring intensity is relatively low. The monitoring intensity increases slightly towards top-income levels. However, the effect of monitoring on the labor wedge and the tax schedule is very small at these high income levels.

The optimal tax schedule exhibits extremely large tax rates at the bottom of the earnings distribution. Indeed, the government can levy tax rates above 100 percent at the lowest income earners. The sharp decrease in the monitoring intensity works as an implicit subsidy on labor supply and partially offsets the high explicit tax on labor supply. The poverty trap found in many countries (see, e.g., Spadaro, 2005, Brewer et al., 2010 and OECD, 2011) can thus be optimal in the presence of monitoring. Indeed, there may not be a poverty trap if the monitoring schedule provides sufficient incentives, even if the tax-benefit system itself does not provide incentives to supply labor.

Note that the optimal wedge and monitoring probability at the top do not equal zero, as was derived in Proposition 4 for a bounded income distribution. Mirrlees (1971), Diamond (1998), and Saez (2001) show theoretically that the optimal wedge converges to a constant if the right tail of the ability distribution is Pareto distributed. In the Pareto tail of the earnings distribution, the ratio of marginal distributional benefits and marginal efficiency costs of taxes becomes constant, and the tax wedge converges to a constant. Our simulations confirm that this result holds as well in the model with monitoring. In addition, we find that the optimal monitoring probability also converges to a positive constant.

5.2 Sensitivity analysis

In this subsection we present the sensitivity analysis of the results obtained in the previous subsection. We especially explore the sensitivity of our simulation outcomes with respect to the monitoring technology and penalty function.

Figure 3 summarizes the simulations when the cost of monitoring is decreased ($\kappa = 0.25$) or increased ($\kappa = 4$). As expected, the monitoring schedule moves up if the monitoring cost decreases and down if the cost increases. However, the optimal tax schedule largely remains unaffected. From the optimal tax expression in eq. (37) we can infer that monitoring increases the optimal tax rate if the allocation remains unchanged. However, the allocation changes, since an increase in the monitoring probability increases revenue from taxation for any given tax rate. Therefore, the redistributive benefit of a marginal tax decreases at the same time. In our simulations, these two effects roughly cancel out and the optimal tax rates remain largely unaffected.

Figure 4 gives the optimal tax and monitoring schedules when the penalty parameter is decreased ($p = 1$) or increased ($p = 5$). As can be seen, for very low levels of income, both an increase and a decrease in the penalty parameter lead to a decrease in the monitoring intensity. This may seem counter-intuitive, but can be explained. An increase in the penalty parameter raises the effectiveness of monitoring, but it also increases within-skill level inequality. For low levels of income, the first effect dominates when penalties decrease, whereas the second effect dominates when penalties increase. However, beyond about 10,000 dollars of income, within-ability inequality becomes less relevant, and therefore, monitoring intensities always increase when the penalty parameter rises.

From the optimal tax formula in eq. (37) it follows that an increase in the penalty parameter affects the optimal tax rate through six channels. First, an increase in the marginal penalty raises the marginal tax rate for a given wedge. Second, an increase in the penalty itself may increase or decrease the optimal marginal tax rate for a given wedge depending on the sign

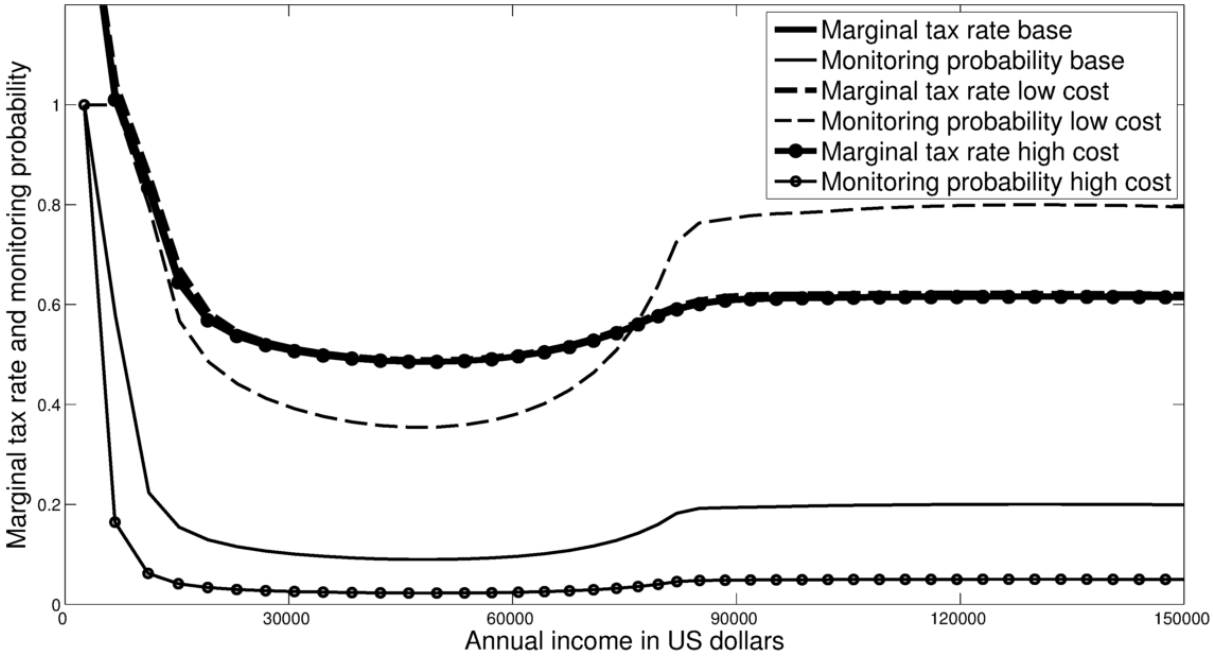

Figure 3: Optimal tax and monitoring schedules for high ($\kappa = 4$) and low ($\kappa = 0.25$) marginal cost of monitoring. All other parameters take baseline values, see Table 1.

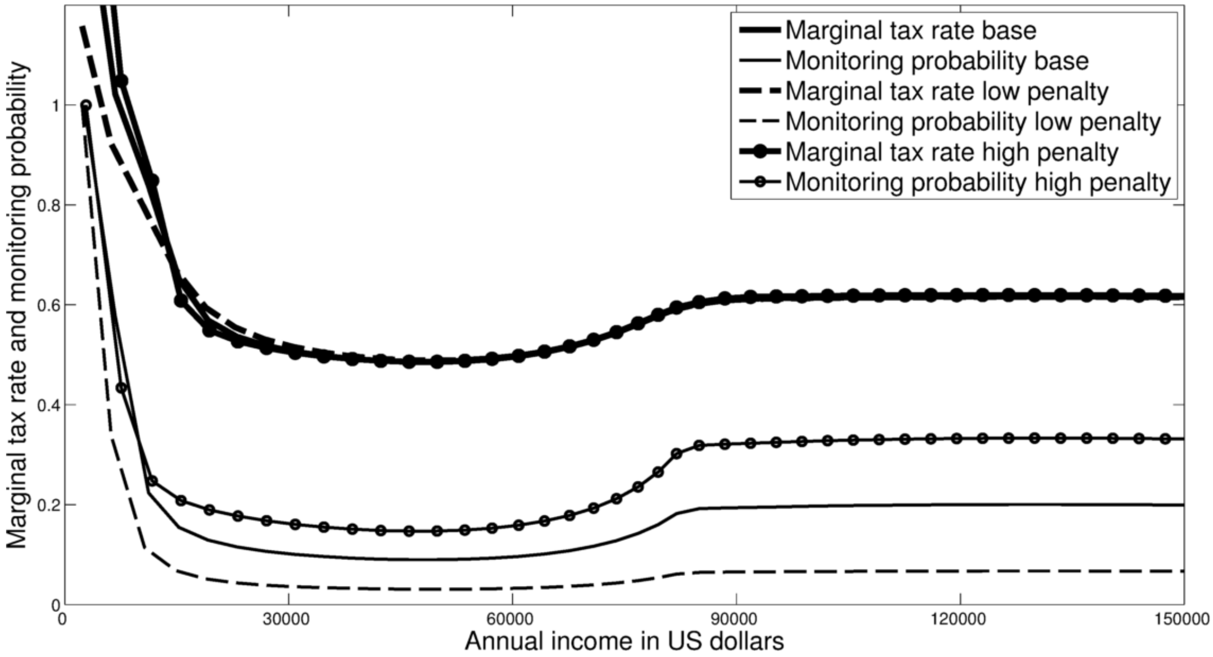

Figure 4: Optimal tax and monitoring schedules for strong ($p = 5$) and weak ($p = 1$) penalties. All other parameters take baseline values, see Table 1.

of $\pi'(z_n)$. Third, an increase in the convexity of the penalty function decreases the efficiency cost of a wedge. Fourth, the penalty affects the monitoring probability, although the effect is ambiguous. Fifth, an increase in the penalty increases within skill-level inequality, which decreases the optimal wedge. Finally, the allocation itself is affected, but it is a priori unclear whether higher penalties lead to more or less redistribution. The simulation outcomes confirm these theoretical ambiguities. The net effect is positive for very low income levels, negative for medium-income levels, and negligible for higher income-levels.

Figure 5 illustrates the effect of a decrease in the reference level of working hours ($l_n^* = 0.5$). As can be seen, the monitoring probability very quickly drops to zero, because all individuals find it in their best interest to work at least the reference amount of labor hours without monitoring. Surprisingly, the tax schedule remains virtually unaffected. This outcome demonstrates monitoring hours worked is most important at the bottom of the earnings distribution, where the labor wedge is highest. Still, marginal tax rates can be substantially above 100 percent at the bottom of the earnings distribution.

Finally, in Figure 6 we simulated the optimal tax and monitoring schedules for a higher degree of inequality aversion ($\beta = 1.5$) and a lower degree ($\beta = 0.5$) of inequality aversion. As can be seen, the optimal tax rate increases in inequality aversion as should be expected, although the difference at the bottom of the income distribution is small. Intuitively, monitoring decreases the distortion of a higher tax rate, but it also creates within skill-group inequality. The poorest individuals in society are the low-income individuals who are penalized. Hence, within-ability inequality is particularly costly if the government is strongly inequality-averse. For low levels of income, both an increase and a decrease of inequality aversion decrease the optimal monitoring intensity. At higher levels of income, within-skill group inequality aversion is less important, and the monitoring intensity unambiguously increases with inequality aversion as labor wedges are set higher when redistributive desires are stronger.

5.3 Allocations and welfare

Clearly, monitoring is part of the optimal redistributive tax-benefit system. But, how important is monitoring for the optimal allocation and welfare? Table 2 reports the average monitoring cost $\bar{k}/\bar{z}$, the average penalty $\bar{P}/\bar{z}$, the penalty for the lowest working individual, $P(\underline{n})/\bar{z}$, the transfer paid out to individuals having zero earnings, $-T(0)/\bar{z}$, and the change in average earnings, $\Delta\bar{z}/\bar{z}$. All table entries are in percentages of average earnings.

From the first column we can infer that the average monitoring cost $\bar{k}/\bar{z}$ is a relatively small percentage of average labor earnings: about 0.5 percent of average earnings in the baseline. An increase in the marginal cost of monitoring raises total monitoring costs very little, since the increase in marginal cost is accompanied by a decrease in the optimal monitoring intensity at the optimum. A change in the reference level of labor hours reduces the total monitoring cost to almost zero, because monitoring is only used at the bottom if the labor requirement is low. In addition, the cost of monitoring is sensitive to the severity of the penalty as monitoring outlays rise (fall) with a stronger (weaker) penalty. A government having access to a stronger penalty technology will on average rely more heavily on monitoring to provide work incentives. Monitoring costs also increase with inequality aversion, since a more inequality-averse government

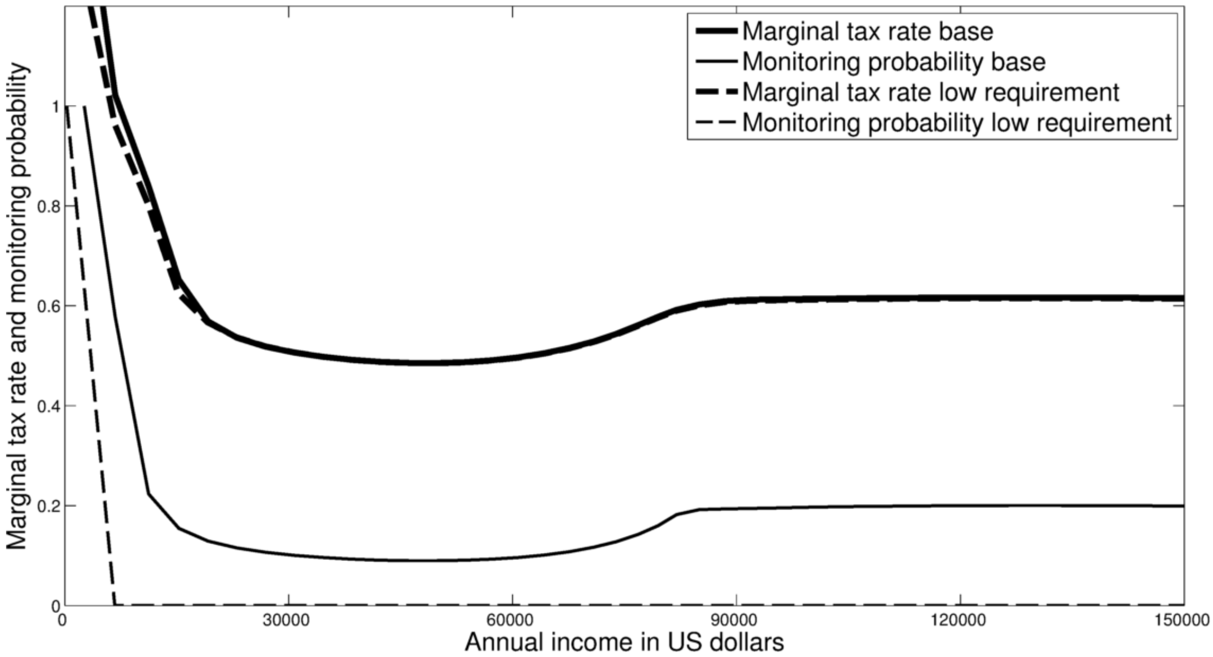

Figure 5: Optimal tax and monitoring schedules for a lower reference level of work effort ($l_n^* = 0.5$). All other parameters take baseline values, see Table 1.

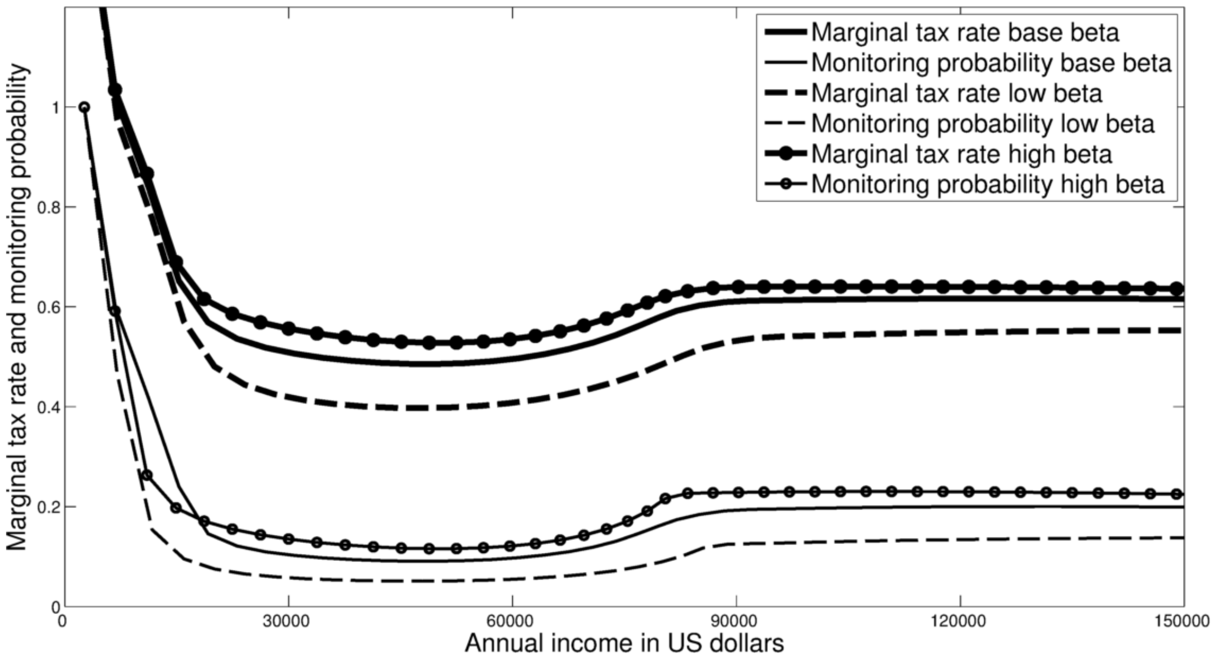

Figure 6: The tax and monitoring schedule for higher ($\beta = 1.5$) and a lower ($\beta = 0.5$) degree of of inequality aversion. All other parameters take baseline values, see Table 1.

relies on average more heavily on monitoring to alleviate the equity-efficiency trade-off.

Table 2: Change in allocation due to monitoring. (All numbers are in percentages of average earnings)

	$\frac{\bar{k}}{\bar{z}}$	$\frac{\bar{P}}{\bar{z}}$	$\frac{P(\underline{n})}{\bar{z}}$	$\frac{-T(0)}{\bar{z}}$	$\frac{\Delta\bar{z}}{\bar{z}}$
No Monitoring	0.00	0.00	0.00	29.46	0
Base scenario	0.49	0.35	7.83	33.65	1.35
Low monitoring cost	0.40	0.34	7.55	34.20	1.50
High monitoring cost	0.61	0.36	8.08	33.13	1.18
Low reference effort	0.03	0.02	5.63	34.85	1.07
Low penalty	0.24	0.14	3.53	29.67	0.43
High penalty	0.63	0.49	10.02	37.01	2.04
Low inequality aversion	0.32	0.23	8.20	30.25	5.55
High inequality aversion	0.61	0.42	7.62	35.23	−0.90

Note: $\bar{z}$ is per capita labor income in the specified calibration, $\bar{k}$ is the per capita monitoring cost, $\bar{P}$ is the average penalty over the monitored population, $P(\underline{n})$ is the penalty at the lowest skill level, $-T(0)$ is the transfer and $\Delta\bar{z}$ is the change in average labor earnings as compared to the model without monitoring.

The second column represents the average penalty given to monitored individuals as a percentage of average labor earnings $\bar{P}/\bar{z}$. As can be seen, penalties are relatively small. In the baseline, the average penalty equals 0.35 percent of average earnings. Penalties increase with the monitoring cost, because monitoring decreases with its marginal cost, and as a consequence, individuals work less and receive more severe penalties. The effects are very small, however. In addition, the average penalty falls strongly when the reference level of labor supply is lower. Similarly, the penalties increase (decrease) if the penalty parameter increases (decreases), as expected. The penalty also increases (decreases) with stronger (weaker) inequality aversion because a more (less) inequality-averse government sets higher (lower) wedges. The third column represents the average penalty at the bottom of the income distribution $P(\underline{n})/\bar{z}$. Penalties at the bottom are relatively large, because the wedge at the bottom is large. Comparative-static effects of the penalty at the bottom are roughly similar to the comparative statics of the average penalty.

The fourth column represents the transfer as a fraction of earnings, $-T(0)/\bar{z}$, and the fifth column is the change in average labor earnings as compared to optimal taxation without monitoring, $\Delta\bar{z}/\bar{z}$. In almost all simulations, both the transfer and average labor earnings increase, indicating an improvement in both equity and efficiency of the tax-transfer system. This effect is surprisingly unsensitive to a change in the monitoring cost and to a change in the reference level of labor hours. These outcomes can be explained by the fact that monitoring is most effective at the bottom of the skill distribution. At this point in the earnings distribution, monitoring costs are relatively unimportant as the density of monitored individuals is low. A lower labor requirement is also unimportant, since individuals are working far less than any of the work requirements we consider at the bottom end of the income scale. Results are more sensitive to the size of the penalty, since monitoring becomes less effective if the punishment technology is less effective. However, even if penalties are relatively low, the increase in both average labor earnings and the transfer is substantial. Finally, a change in the inequality aversion changes the emphasis given to either equity (higher transfers $T(0)$) or efficiency (higher average labor earnings $\bar{z}$). In our scenario with low inequality aversion both increase. However, in the scenario with high inequality aversion average labor earnings decreases slightly.

Table 3: Welfare effects of monitoring.

	Marginal dead weight loss	Welfare gain
No Monitoring	0.204	–
Base	0.203	1.421
Low monitoring cost	0.203	1.592
High monitoring cost	0.204	1.073
Low reference effort	0.203	0.969
Low penalty	0.204	0.267
High penalty	0.203	1.835
Low inequality aversion	0.170	1.015
High inequality aversion	0.214	1.76

Note: The marginal deadweight loss refers to the income-weighted average of the marginal deadweight loss of all households as a consequence of increasing the labor wedge on labor with one percent. Welfare gains are obtained by calculating the compensating variation as a percentage of average earnings in the specified simulation.

Finally, Table 3 reports the welfare effects of monitoring. The first column represents the income-weighted average of the marginal deadweight loss of increasing the marginal tax rate by one percent. As can be seen, monitoring decreases the marginal deadweight loss by about 0.5 percent in our baseline simulation from 0.204 to 0.203. This result is robust in our sensitivity analyses.

The last column reports the monetized welfare gain of monitoring. We compute the compensating variation by calculating the amount of resources that have to be injected into an economy without monitoring in order to attain the same social welfare as the economy with optimal monitoring. In our base scenario, the welfare gain is about 1.4 percent of average labor earnings, i.e. 1.4 percent of total output. The welfare gain increases if the cost of monitoring are lower and if penalties are higher. Interestingly, the welfare gain is almost unaffected by a lower reference level of labor supply. The reason is that the reference level of labor supply still generates positive penalties at the bottom of the earnings distribution, where the benefits of monitoring are highest. Also, an increase in inequality aversion increases the welfare gain of monitoring, because the efficiency gain of monitoring is increasing in the optimal labor wedges, which are larger when the government is more inequality averse. We find quantitatively substantial social welfare gains in all scenarios.

6 Conclusions

In this paper we demonstrate that redistributive governments should optimally monitor labor hours in order to redistribute income at the lowest efficiency cost. Monitoring of labor supply alleviates the equity-efficiency trade-off and raises equity, efficiency, or both. The reason is that distortions from redistribution derive from the informational problem that earning ability is private information. By using a monitoring technology this informational asymmetry is reduced. A first-best outcome cannot be reached, however, because monitoring is costly. Mirrlees (1971) is a special case of our model when monitoring is infinitely costly.

We demonstrated that monitoring labor supply works as an implicit subsidy on labor supply, which partially offsets the explicit tax on labor supply. We derived conditions on the desirability of monitoring and demonstrated that the optimal non-linear monitoring schedule generally follows the optimal labor wedge. Monitoring is more desirable when redistributive taxation creates larger distortions in labor supply. Moreover, optimal labor taxes can optimally be above 100 percent when monitoring is allowed for. At the endpoints of the earnings distribution labor wedges – including taxes and the implicit subsidy on work due to monitoring – are zero in the absence of bunching and with a finite skill level.

Simulations confirmed that the optimal monitoring intensity features a U-shaped pattern with income; very high at the lower end of the earnings distribution, declining towards the middle-income groups, increasing again towards the high-income groups, and becoming constant at the top-income groups. Our simulations demonstrated that marginal tax rates will be higher if the government monitors labor supply, while the labor wedges – including the explicit tax and implicit subsidy of monitoring – decreases. Indeed, monitoring is very effective to alleviate the equity efficiency trade-off.

In practice, monitoring is not infinitely costly as in Mirrlees (1971). By allowing for a monitoring technology we can explain our why work-dependent tax credits for low-income earners, that are employed in the UK, Ireland and New Zealand, are part of an optimal redistributive tax policy. Our findings also show that sanctions for welfare recipients, bonuses for low-income workers, and extensive monitoring of labor effort or working ability of low-earning individuals are especially desirable in more generous welfare states. Moreover, we can also explain why (large) penalties on hours worked supply (or high bonuses on hours worked) are more desirable when the government desires to redistribute more income. Finally, we find that marginal tax rates larger than 100 percent at the lower end of the earnings distribution, as commonly observed in many countries, can be optimal in the presence of monitoring of labor supply.

References

Allingham, Micheal G., and Agmar Sandmo (1972) 'Income tax evasion: A theoretical analysis.' *Journal of Public Economics* 1(3-4), 323–338

Armenter, Roc, and Thomas M. Mertens (2013) 'Fraud deterrence in dynamic Mirrleesian economies.' *Journal of Monetary Economics* 60(2), 139–151

Bassetto, Marco, and Christopher Phelan (2008) 'Tax riots.' *Review of Economic Studies* 75(3), 649–669

Boadway, Robin, and Katherine Cuff (1999) 'Monitoring job search as an instrument for targeting transfers.' *International Tax and Public Finance* 6(3), 317–337

Boone, Jan, and Jan C. Van Ours (2006) 'Modeling financial incentives to get the unemployed back to work.' *Journal of Institutional and Theoretical Economics* 162(2), 227–252

Boone, Jan, Peter Fredriksson, Bertil Holmlund, and Jan C. Van Ours (2007) 'Optimal unemployment insurance with monitoring and sanctions.' *Economic Journal* 117(518), 399–421

Brewer, Mike, Emmanuel Saez, and Andrew Shephard (2010) 'Means-testing and tax rates on earnings.' In *The Mirrlees Review – Dimensions of Tax Design,* ed. James A. Mirrlees, Stuart Adam, Timothy J. Besley, Richard Blundell, Steven Bond, Robert Chote, Malcolm Gammie, Paul Johnson, Gareth D. Myles, and James M. Poterba (Oxford: Oxford University Press) chapter 3, pp. 202–274

Chander, Parkash, and Louis L. Wilde (1998) 'A general characterization of optimal income tax enforcement.' *Review of Economic Studies* 65(1), 165–183

Cremer, Helmuth, and Firouz Gahvari (1994) 'Tax evasion, concealment and the optimal linear income tax.' *Scandinavian Journal of Economics* 2(96), 219–239

_ (1996) 'Tax evasion and the optimum general income tax.' *Journal of Public Economics* 60(2), 235–249

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

Diamond, Peter, and Eytan Sheshinski (1995) 'Economic aspects of optimal disability benefits.' *Journal of Public Economics* 57(1), 1–23

Ebert, Udo (1992) 'A reexamination of the optimal non-linear income tax.' *Journal of Public Economics* 49(1), 47–73

Fredriksson, Peter, and Bertil Holmlund (2006) 'Improving incentives in unemployment insurance: A review of recent research.' *Journal of Economic Surveys* 20(3), 357–386

Harris, Milton, and Robert M. Townsend (1981) 'Resource allocation under asymmetric information.' *Econometrica* 49(1), 33–64

Holmstrom, Bengt (1979) 'Moral hazard and observability.' *Bell Journal of Economics* 10(1), 74–91

Immervoll, Herwig (2004) 'Average and marginal effective tax rates facing workers in the EU: A micro-level analysis of levels, distributions and driving factors.' OECD Social, Employment and Migration Working Papers, No. 19 Paris: OECD Publishing

Jacquet, Laurence (2014) 'Tagging and redistributive taxation with imperfect disability monitoring.' *Social Choice and Welfare* 42(2), 403–435.

Kleven, Henrik J., Martin B. Knudsen, Claus T. Kreiner, Soren Pedersen, and Emmanuel Saez (2011) 'Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark.' *Econometrica* 79(3), 651–592

Kocherlakota, Narayana R. (2006) 'Advances in dynamic optimal taxation.' In *Econometric Society Monographs,* ed. Richard Blundell, Whitney K. Newey, and Torsten Persson (Cambridge: Cambridge University Press) chapter 7, pp. 269–297

Ljungqvist, Lars, and Thomas J. Sargent (1995a) 'The Swedish unemployment experience.' *European Economic Review* 39(5), 1043–1070

_ (1995b) 'Welfare states and unemployment.' *Economic Theory* 6(1), 143–160

Mankiw, N. Gregory, Matthew Weinzierl, and Danny Yagan (2009) 'Optimal taxation in theory and practice.' *Journal of Economic Perspectives* 23(4), 147–174

Mirrlees, James A. (1971) 'An exploration in the theory of optimum income taxation.' *Review of Economic Studies* 38(2), 175–208

_ (1976) 'Optimal tax theory: A synthesis.' *Journal of Public Economics* 6(4), 327–358

_ (1997) 'Information and incentives: The economics of carrots and sticks.' *Economic Journal* 107(444), 1311–1329

_ (1999) 'The theory of moral hazard and unobservable behaviour: Part I.' *Review of Economic Studies* 66(1), 3–21

Mookherjee, Dilip, and Ivan Png (1989) 'Optimal auditing, insurance, and redistribution.' *Quarterly Journal of Economics* 104(2), 399–415

Myerson, Roger B. (1979) 'Incentive compatibility and the bargaining problem.' *Econometrica* 47(1), 61–73

OECD (2011) 'Taxation and employment.' OECD Tax Policy Studies, No.21 Paris

Sadka, Efraim (1976) 'On income distribution, incentive effects and optimal income taxation.' *Review of Economic Studies* 43(2), 261–267

Saez, Emmanuel (2001) 'Using elasticities to derive optimal income tax rates.' *Review of Economic Studies* 68(1), 205–229

Saez, Emmanuel, Joel B. Slemrod, and Seth 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

Sandmo, Agnar (1981) 'Income tax evasion, labour supply, and the equity–efficiency tradeoff.' *Journal of Public Economics* 16(3), 265–288

Schroyen, Fred (1997) ‘Pareto efficient income taxation under costly monitoring.’ *Journal of Public Economics* 65(3), 343–366

Seade, Jesus K. (1977) ‘On the shape of optimal tax schedules.’ *Journal of Public Economics* 7(2), 203–235

Slemrod, Joel (1994) ‘Fixing the leak in Okun’s bucket: Optimal tax progressivity when avoidance can be controlled.’ *Journal of Public Economics* 55(1), 41–51

Slemrod, Joel, and Shlomo Yitzhaki (2002) ‘Tax avoidance, evasion, and administration.’ In *Handbook of Public Economics Volume 3,* ed. Alan J. Auerbach and Martin Feldstein (Amsterdam: Elsevier) chapter 22, pp. 1423–1470

Slemrod, Joel, and Wojciech Kopczuk (2002) ‘The optimal elasticity of taxable income.’ *Journal of Public Economics* 84(1), 91–112

Spadaro, Amedeo (2005) ‘Micro-simulation and Normative Policy Evaluation: An Application to Some EU Tax Benefits Systems.’ *Journal of Public Economic Theory* 7(4), 593–622

Stern, Nicholas (1982) ‘Optimum taxation with errors in administration.’ *Journal of Public Economics* 17(2), 181–211

Townsend, Robert M. (1979) ‘Optimal contracts and competitive markets with costly state verification.’ *Journal of Economic Theory* 21(2), 265–293

Tuomala, Matti (1984) ‘On the optimal income taxation: Some further numerical results.’ *Journal of Public Economics* 23, 351–366

Zoutman, Floris T., and Bas Jacobs (2014) ‘Optimal redistribution and monitoring of ability.’ CES-ifo Working Paper No. 4646, Munich: CESifo.

Zoutman, Floris T., Bas Jacobs, and Egbert L.W. Jongen (2013) ‘Optimal redistributive taxes and redistributive preferences in the Netherlands.’ Rotterdam. Mimeo: Erasmus University Rotterdam/CPB Netherlands Bureau for Economic Policy Analysis

A Simulation algorithm

The algorithm we use to solve for the optimal allocation consists of two steps. First, we find the optimal allocation using a shooting method. Second, we calculate the implied wedge, tax, and monitoring schedules.[13]

A.1 Finding the optimal allocation

We find the optimal allocation through four nested loops:

[13] All Matlab programs used in the computations are available from the authors upon request.

1. The outer loop solves the resource constraint (12) for λ. A higher value of λ implies a higher shadow value of resources, and thus, a lower resource deficit, and vice versa. Therefore, we can satisfy the resource constraint arbitrarily by altering the value of λ.

2. The second loop solves the transversality condition at the top (31) for a given utility level at the bottom $u_{\underline{n}}$, and λ. The most important determinant in $u_{\underline{n}}$ is the transfer implied by $T(0)$. Therefore, one can think of this procedure as finding the intercept of the tax function $T(0)$. If the intercept is too low, the distortion at the top has to be positive to finance the transfer, and vice versa if the intercept is set too high. As a consequence, by varying the transfer $T(0)$ we can satisfy the transversality condition arbitrarily closely.

3. The third loop solves the differential equations (13) and (29) for given $u_{\underline{n}}$, λ, and $\theta_{\underline{n}}$ using a Runge-Kutta method to integrate over n.

4. The inner loop maximizes the Hamiltonian (25) with respect to π_n and z_n for a given state u_n and costate variable θ_n at each n.

The above algorithm is known as a shooting method because it shoots for the initial values of the differential equations that satisfy the boundary condition.

A.2 Finding the optimal wedge, tax, and monitoring schedules

The above algorithm gives us a numerical approximation of the allocation $\{u_n, \theta_n, z_n, \pi_n\}$ at each n. $\pi'(z_n)$ can be approximated by taking the first difference:

$$\pi'(z_n) \approx \frac{\Delta \pi_n}{\Delta z_n}. \tag{51}$$

With $\pi'(z_n)$ we have all the information we need to find the optimal tax schedule using eq. (37).