

# MORAL HAZARD, INCOME TAXATION, AND PROSPECT THEORY<sup>\*</sup>

by

**Ravi Kanbur**

Cornell University

**Jukka Pirttilä<sup>#</sup>**

Bank of Finland

and

**Matti Tuomala**

University of Tampere

This version: 17 March 2004

---

<sup>\*</sup> We are grateful to Hamish Low and Tuomas Takalo for helpful discussions.

<sup>#</sup> Corresponding author, email: [Jukka.Pirttila@bof.fi](mailto:Jukka.Pirttila@bof.fi)

**Abstract:** The standard theory of optimal income taxation under uncertainty has been developed under the assumption that individuals maximize expected utility. However, prospect theory has now been established as an alternative model of individual behaviour, with empirical support. This paper explores the theory of optimal income taxation under uncertainty when individuals behave according to the tenets of prospect theory. It is seen that many of the standard results are either overturned, or modified in interesting ways. The validity of the First Order Approach requires new conditions that are developed in the paper. And when these conditions are valid, it is shown that optimal marginal tax rates on low incomes will tend to be lower under prospect theory than under expected utility theory.

**Key words:** redistributive taxation, income uncertainty, moral hazard, prospect theory, loss aversion

**JEL classification:** D81, H21.

## 1 Introduction

In principal-agent models with moral hazard, the agent's income and utility depends randomly on effort. In one of his Nobel prize winning contributions, Mirrlees (1974) characterizes optimal income taxation where the government is the principal and ex ante identical individuals are the agents<sup>1</sup>. In this contribution, as in most principal-agent analysis, expected utility theory is used as a description of agents' behaviour under uncertainty. In his Nobel Lecture, Mirrlees (1997, p. 1324) calls for closer scrutiny of this approach:

'Problems of this kind are usually analysed with the assumption that people try to maximise their expected utility. There are good reasons for thinking that may be a mistake. At least the consequences of alternative theories of decisions under uncertainty for these situations should be explored.'

One such alternative theory of decision making under uncertainty has in turn led to another Nobel prize. Prospect theory, developed by Kahneman and Tversky (1979), has garnered significant empirical support (see Kahneman's Nobel lecture, 2003, and Camerer and Lowenstein, 2003). In prospect theory, an individual's utility depends on how the outcome deviates from some reference point, rather than directly on the absolute value of the outcome. Individuals are loss averse, in other words, a loss leads to a larger change in welfare than a gain of a similar size. Finally, individuals may misperceive probabilities underlying the decision problem.

The purpose of this paper is to confront the Mirrlees project of characterizing optimal income taxation with moral hazard under uncertainty, with the Kahneman project of developing alternatives to standard expected utility theory. In the original Mirrlees' (1974) formulation of the income tax model, workers and the government maximise workers' expected utility over income and effort. Our purpose is to introduce elements of prospect theory into individual behaviour, in keeping with the emerging empirical consensus. However, the preferences used by the social planner remains an open question. Should the government maximize individual welfare as the individual sees it, ie should it be "welfarist"? Or, as is suggested in some recent

---

<sup>1</sup> Of course the case with no uncertainty but where individuals differ in their productivities, was also introduced and explored by Mirrlees (1971) in the more famous of his Nobel prize winning contributions. We will refer to this "adverse selection" case from time to time, but our focus is on the "moral hazard" case where there is uncertainty but no ex ante differentiation among individuals.

behavioural economics literature, should it be “non-welfarist” and use expected utility theory to evaluate outcomes, even though individuals use prospect theory? The latter approach is relatively common in conventional public economics, and has been used recently in the behavioural public economics literature as well.<sup>2</sup> In this paper we develop a general approach that encompasses both welfarist and non-welfarist perspectives, and then specialise to draw implications for each case.

When prospect theory is used as a description of individual behaviour, another interesting connection to earlier tax analysis emerges. What determines the reference income level which individuals use in assessing losses and gains in prospect theory? One plausible specification is that individuals compare their ex post outcome relative to the mean of the outcome for other individuals. This comes very close to models of optimal taxation with utility interdependence (or envy), explored in the conventional optimal taxation setting without income uncertainty (e.g. Oswald, 1983 and Tuomala, 1990). Thus, parts of this paper can also be seen as conducting an analysis of taxation with utility interdependence under income uncertainty. Another possibility is that the reference point could be a past consumption level. Then the analysis bears resemblance to habit formation models, such as Ljungqvist and Uhlig (2000) or Carroll, Overland and Weil (2000).

The paper first reviews the standard, benchmark, model of optimal taxation with moral hazard under income uncertainty, in Section 2. As is common in models of this kind, we mostly focus on results based on the so-called “first-order approach”. This section therefore also examines the exact conditions under which this approach is valid, an issue that will be seen to be relevant when prospect theory is introduced. Section 3 develops a solution to the tax problem under uncertainty with moral hazard in the general setting where individuals use one set of preferences to evaluate outcomes, but the government uses a different set of preferences. Section 4 then interprets these general results when individual behaviour is described by prospect theory, but the government uses expected utility theory to evaluate outcomes. The case where individuals and the government both use prospect theory is discussed in the Appendix. Section 5 extends the analysis by considering the case where the reference income of prospect theory is endogenised. Section 6 concludes.

---

<sup>2</sup> An example of the former is Kanbur, Keen and Tuomala (1994), while O’Donoghue and Rabin (2003) is an example of the latter.

## 2. The standard model

Consider an economy, as in Mirrlees (1974), where the worker-consumer does not know what income he or she will receive for each possible level of effort. Thus the worker's gross income,  $z$ , depends randomly on effort,  $y$ . There is a single worker or, alternatively, all workers are ex ante identical. Thus income differences are not due to differences in innate skills (as in conventional optimal income taxation model under certainty, Mirrlees, 1971), but due to luck (for any given level of effort). Let  $f(z, y)$  and  $F(z, y)$  denote the continuous density and distribution functions of income  $z$  given that effort  $y$  is undertaken by the worker; it is assumed that they are continuously differentiable for all  $z$  and  $y$ . The worker-consumer chooses effort  $y$  to maximize expected utility

$$(1) \int u(x, y) f(z, y) dz,$$

where  $x = z - T(z)$  is the after tax income / consumption. As in much of the literature, we concentrate on an additively separable specification of the utility function, written as  $u(x, y) = v(x) - y$ . The consumer is risk averse, hence  $v' > 0, v'' < 0$ . The first-order condition for the maximisation of (1) is

$$(2) \int v(x) f_y dz - 1 = 0$$

The government is utilitarian and maximises (1) subject to the individual optimisation constraint (2) and the budget constraint which, for large identical population with independent and identically distributed states of nature, can be written in the form

$$(3) \int [z - x] f(z, y) = 0$$

Taking multipliers  $\alpha$  and  $\lambda$  for the constraints (2) and (3) respectively, the Lagrangean and the first-order condition wrt  $x$  (pointwise optimisation) are as follows:

$$(4) L = \int \{ [v(x) + \lambda(z - x)] f(z, y) + \alpha v(x) f_y \} dz - \alpha - y$$

$$(5) \quad 1 + \alpha g = \frac{\lambda}{v'},$$

where  $g = f_y/f$  is the likelihood ratio. This approach, where incentive compatibility is modelled using equation (2), is the so-called first-order approach (FOA). Mirrlees (1975, 1999) was the first to point out that FOA is not necessarily a valid procedure in a potentially large number of cases, because it might lead to a local instead of a global optimum. Mirrlees (1976), Rogerson (1985), Jewitt (1988) and Alvi (1997) have explored conditions under which the FOA provides necessary and sufficient conditions for the optimisation. When the utility function is separable as in our set-up, sufficient conditions are the so-called monotone likelihood ratio condition (MLRC) and convex distribution function condition (CDFC). We demonstrate this below to provide a comparison to the novel cases in later sections.

First, as an intermediate step, it must be checked that consumption is increasing in income / effort, i.e.  $x'(z)$  is positive. The right-hand side of the first-order condition in (5) is increasing in  $z(y)$ , since  $v'' < 0$ . The left-hand side of (5) is increasing with  $z(y)$ , provided that  $\alpha > 0$ , since  $g' > 0$  when we assume that the MLRC (monotone likelihood ratio condition) holds.

Following Jewitt (1988) and Laffont and Martimort (2002) we now show that  $\alpha$  is indeed positive. Dividing (5) by  $\lambda$ , multiplying it with  $f$  and integrating over the support  $[\underline{z}, \bar{z}]$  yields

$$(6) \quad \frac{1}{\lambda} = \int \frac{1}{v'} f dz$$

since  $\int f_y dz = 0$ . Using (5) again gives

$$(7) \quad \frac{\alpha}{\lambda} g = \frac{1}{v'} - \int \frac{1}{v'} f dz$$

Multiply both sides by  $v'f$  and integrate over the support  $[\underline{z}, \bar{z}]$  to get

$$(8) \quad \frac{\alpha}{\lambda} \int v f_y dz = \text{cov}(v, \frac{1}{v'})$$

$$\frac{\alpha}{\lambda} = \text{cov}(v, \frac{1}{v'})$$

where  $\text{cov}()$  denotes the covariance operator. The second line follows from the incentive constraint (2). Since  $v$  and  $v'$  covary in opposite, we necessarily have  $\alpha \geq 0$ . However, the only case where the covariance is zero is when  $x$  is constant irrespective of income. But then the worker has no incentives to provide positive effort. Therefore, to induce effort,  $\alpha > 0$ .

Finally, it remains to be shown that expected utility is concave. For this, rewrite the expected utility as follows:

$$\int v(x) f(z, y) dz - y$$

$$(9) = [vF(z, y)]_{\underline{z}}^{\bar{z}} - \int v' x' F(z, y) dz - y$$

$$= v(x(\bar{z})) - \int v' x' F(z, y) dz - y$$

where the second line follows from integration by parts and the third from using the property that  $F(\underline{z}, y) = 0$  and  $F(\bar{z}, y) = 1$ . From the last row, since  $v'$  and  $x'$  are positive, the expected utility is concave in  $y$  if  $F_{yy} > 0$ . This property is called the convexity of the distribution function condition (CDFC). Therefore, the FOA is a valid strategy given that MLRP and CDFC hold.

We can now turn to the properties of the solution. Differentiate (5) again with respect to  $z$  and reorganise to obtain the shape of  $x'(z)$ :

$$(10) \quad x' = -\frac{\alpha(v')^2 g'}{\lambda v''}$$

Denote the coefficient of absolute risk aversion as  $\delta = -(v'')/(v')$ . Based on (10), the marginal tax rate ( $MTR = T'(z)$ ) is therefore given by

$$(11) \quad MTR = 1 - x' = 1 - \frac{\alpha v' g'}{\lambda \delta}$$

This shows that the optimal marginal tax rate is a compromise between risk aversion and providing incentives. If the consumers become more risk averse, the marginal tax rate increases, *ceteris paribus*. On the other hand, if effort is more tightly connected with income ( $g'$  goes up), workers' effort can be more reliably tracked, and the optimal marginal tax rate is reduced.

### 3. Moral hazard and non-welfarism

The standard solution presented above was based on “expected utility welfarism” – the government used the same objective function as the individuals in assessing social welfare, namely maximising expected utility. In non-welfarist welfare economics, the individuals' and the government's objective functions differ. Seade (1980) derives optimal tax rules under a general non-welfarist objective function for the adverse selection case. Kanbur et al (1994) and Pirttilä and Tuomala (2004) consider the implications of poverty alleviation as a policy objective on optimal income taxation and the combination of income commodity taxation rules, respectively. All these papers deal with the Mirrlees (1971) formulation of optimal tax policy, where individuals' innate skill levels differ. Our purpose is here to provide a general non-welfarist tax analysis in the moral hazard situation. The tax rules are then interpreted from the prospect theory viewpoint in Section 4.

Consider a situation where the government uses a utility function  $u(x, y) = v(x) - y$ , whereas the individuals themselves maximise their welfare using another utility function, say,  $\tilde{u}(x, y)$ . For simplicity, let us concentrate on the case where the difference is only related to the utility derived from consumption. Then we define  $\tilde{u}(x, y) \equiv e(x) - y$ . The individual maximises this with respect to effort, giving rise to the first-order condition

$$(12) \quad \int e(x) f_y dz - 1 = 0$$

The government maximises its utility function subject to (12) and the budget constraint (3). The Lagrangean and the first-order condition are now:

$$(13) L = \int \{ [v(x) + \lambda(z - x)]f(z, y) + \alpha e(x)f_y \} dz - \alpha - y$$

$$(14) v'f + \alpha e'f_y - \lambda f = 0$$

Differentiation of (14) yields

$$(15) x' = -\frac{\alpha e'g'}{v'' + \alpha e''g}$$

Let us first revisit the conditions when the FOA is a valid solution procedure for this optimisation problem. Again, as an intermediate step, consumption  $x$  should be increasing in income  $z$ . The numerator of (15) is positive, assuming that  $\alpha > 0$  and  $e' > 0$ , and that MLRC holds. Given that the numerator is positive, the denominator must be negative, i.e.  $v'' + \alpha e''g < 0$ . If both the government's and the individuals' utility function exhibit decreasing marginal utility, this is always satisfied. Alternatively, one function, say  $e$ , may not be concave, but its impact is offset by the concavity of  $v$ .

The next step is to show that  $\alpha > 0$ . Rearrange (14) to get

$$(16) \frac{1}{\lambda} + \frac{\alpha e'}{\lambda v'} g = \frac{1}{v'}$$

Integrating over the support  $[z, \bar{z}]$  yields

$$(17) \frac{1}{\lambda} = \int \frac{1}{v'} f dz$$

and the first-order condition can be written as:

$$(18) \frac{\alpha}{\lambda} \frac{e'}{v'} g = \frac{1}{v'} - \int \frac{1}{v'} f dz$$

By multiplying with  $v$  and integrating over the support  $[\underline{z}, \bar{z}]$  one obtains

$$(19) \frac{\alpha}{\lambda} \int \frac{e'v}{v'} g = \text{cov}(v, \frac{1}{v'})$$

The right-hand side is always non-negative. So is the term multiplying  $\alpha$  on the left. Therefore we necessarily have  $\alpha \geq 0$ . Again, to induce a positive effort level,  $\alpha$  cannot be equal to zero. Hence,  $\alpha$  is positive.

It remains to show that the objective function is concave. This has actually been already shown above in equation (9). To summarise,

***Proposition 1.*** *The first-order approach is a valid solution procedure if the monotone likelihood ratio and convexity of the distribution function properties hold and a novel condition is valid, namely a combination of  $v''$  and  $e''$  is sufficiently negative.*

We now turn to the interpretation of the tax rule. Note first that in the standard case of the previous section, the marginal tax rate, given in (11), can be rewritten as  $MTR = 1 + \frac{\alpha v' g'}{(1 + \alpha g)v''}$  by (5). By adding and subtracting  $\alpha v' g'$  to the right-hand side of (15), dividing it by  $(1 + \alpha g)v''$ , and rearranging, the marginal tax rate of this section can be written as

$$(20) MTR = 1 - x' = 1 + \frac{1 + \alpha g}{1 + \alpha g \frac{e''}{v''}} \left[ \frac{\alpha v' g'}{(1 + \alpha g)v''} - \frac{\alpha g'(v' - e')}{(1 + \alpha g)v''} \right].$$

The first term in the brackets on the right-hand side of (20) is similar to the tax rule in the welfarist setting. It is, however, multiplied by a novel term that depends on the relative concavity of  $v$  and  $e$ . In addition, there is another new term, the second-term within the

brackets, arising from the non-welfarist objective. It measures the difference between the welfare assessments of the individual and the government. Depending on the magnitude of  $e$  and  $v$ , this term may either increase or decrease the marginal tax rate. The following proposition summarises:

***Proposition 2.** The marginal tax rate in the non-welfarist moral hazard problem is a combination of the standard marginal tax rule and a new term measuring the deviation between individual and social preferences.*

The structure of the non-welfarist tax rule in (20) for the moral hazard case is similar in spirit to the non-welfarist rule calculated by Seade (1980) for the adverse selection case. Note finally that if  $e = v$ , the rule collapses to the standard welfarist rule, expressed in equation (11) in the previous section.

#### **4. Prospect theory and moral hazard**

This section interprets the general analysis in the previous section by assuming that individual behaviour is described by prospect theory, while government behaviour is described by expected utility theory. The case where both principal and agent use prospect theory is discussed in the appendix.

In prospect theory, the utility function is replaced by a value function. The key assumptions about the value function are that it ‘is (i) defined on deviations from the reference point; (ii) generally concave for gains and convex for losses; (iii) steeper for losses than for gains’ (Kahneman and Tversky 1979). The two latter properties capture the idea that individuals are loss averse. Hence, the value function takes the S-shaped value as in Figure 1.

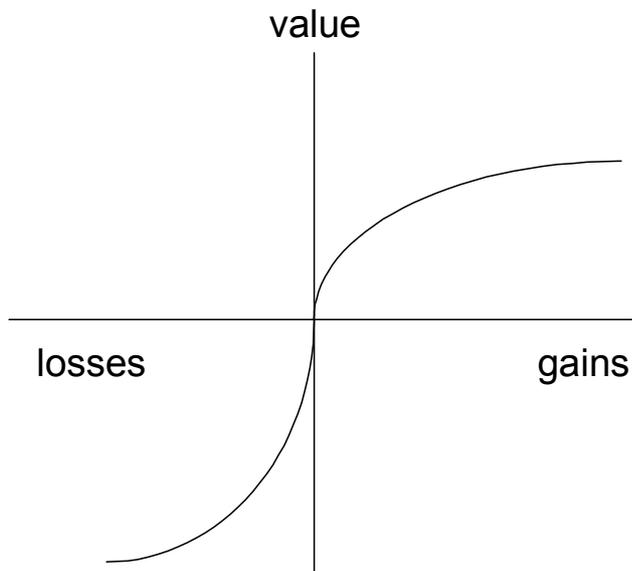


Figure 1: An example of a value function.

Individuals now maximise the expectation of the value function

$$(21) \int e(x - \bar{x}) f(z, y) dz - y$$

which is again additively separable between effort and income. The reference income, which is the basis of comparison for the individual, is depicted by  $\bar{x}$ . The reference income is for the moment assumed to be exogenous; it is endogenised in Section 5.

To capture the shape of the value function, we make the following assumptions about the properties of the utility function:<sup>3</sup>

---

<sup>3</sup> In prospect theory, there is a kink in the value function at the reference income. Here it is assumed that the second derivative smoothly changes from positive to zero and to negative to guarantee differentiability. Differentiability is crucial for the “first order approach” to principal-agent problems. Analysis of the non-differentiable case is a topic for future research.

$$(22) \quad \begin{aligned} &e' > 0, \\ &e'' > 0 \text{ for } x < \bar{x}, e'' < 0 \text{ for } x > \bar{x}, e'' = 0 \text{ for } x = \bar{x} \end{aligned}$$

We can now utilise the results of the previous section by assuming that the government is non-welfarist, using  $v$ , and the individuals'  $e$  has its shape from prospect theory. For the moment we also assume that the government and the individual have the same perception of probabilities—the misperception issue will be dealt with presently. It will be helpful to rewrite  $x'$  in (15), using (14) as

$$(23) \quad x' = \frac{\alpha e' g'}{v' \delta_v - \delta_e (\lambda - v')},$$

where  $\delta_v = -v''/v'$  is the coefficient of absolute risk aversion and  $\delta_e = e''/e'$  is the degree of loss aversion. Note that both are defined to be positive for  $x < \bar{x}$ . The term  $\lambda - v'$  is positive because of (14). The marginal tax rate is

$$(24) \quad MTR = 1 - x' = 1 - \frac{\alpha e' g'}{v' \delta_v - \delta_e (\lambda - v')},$$

For the first-order approach to be a valid solution procedure,  $x$  should be increasing in  $z$ . For  $x \geq \bar{x}$ , this is always the case because then  $\delta_e \leq 0$ , and the denominator in (23) and the term at the right of (23) are positive.

However, if  $x < \bar{x}$ , the denominator in (23) may be either positive or negative. Then the FOA remains valid if the coefficient of loss aversion is sufficiently smaller than the coefficient of risk aversion, i.e.  $v' \delta_v > \delta_e (\lambda - v')$ . This is equivalent to the novel condition found in Proposition 1. However, with sufficiently strong loss aversion,  $v' \delta_v < \delta_e (\lambda - v')$ , the denominator and the right-hand side of (23) are negative, i.e. consumption would be decreasing in income (effort). This violates the validity of the first-order approach. Therefore, the following proposition holds.

**Proposition 3.** *With sufficiently strong loss aversion, the first-order approach is not a valid solution procedure for the entire range of realised income, when the individuals' decisions are based on prospect theory.*

The optimal solution is therefore non-continuous. For consumption above the reference level, the marginal tax rate is given by (24). For low consumption and strong loss aversion, a randomised schedule is optimal. To see this, note that loss aversion, the convexity of the value function, implies that the consumer is in fact risk loving, if income is below the reference point. It is therefore conceivable why conditions needed for an *insurance* scheme are then not valid. In contrast, a risk-loving consumer by definition prefers a randomised schedule  $(x_0, \bar{x})$ , where  $x_0$  is the smallest possible value of income, to a certain combination of  $x_0$  and  $\bar{x}$ . This suggests that the following position holds.

**Proposition 4.** *In a non-welfarist moral hazard tax problem, when the individuals' decision making is based on prospect theory, the optimal incentive structure for incomes below the reference point ( $x < \bar{x}$ ) is a randomised schedule  $(x_0, \bar{x})$  if individuals' loss-aversion sufficiently outweighs the government's risk aversion. If the government's risk aversion sufficiently outweighs individuals' loss-aversion or income is above the reference point ( $x > \bar{x}$ ), the FOA is valid and the marginal tax rate is given by (24).*

Then the overall tax schedule is a combination of the randomisation for  $x \leq \bar{x}$  and a typical tax function with some redistribution for  $x > \bar{x}$ . The point that randomisation may be desirable in moral hazard context is not new. Holmström (1979) and Arnott and Stiglitz (1988) have shown, however, that randomisation is never optimal for a standard, concave, moral hazard problem as in Section 2. Rather, it can become optimal in more complicated situations (Arnott and Stiglitz 1988).<sup>4</sup>

It is interesting to examine how the continuous part of the solution, the marginal tax rate given in (24), changes when loss aversion changes. Since  $\lambda - v'$  is positive, an increase in (positive)  $\delta_e$  tends to reduce the marginal tax rate. *Ceteris paribus*, the marginal tax rate of the

---

<sup>4</sup> Of course, the question remains how randomisation can be implemented in real world tax policy. One of the ideas presented in this context is lax control of tax evasion.

non-welfarist solution with prospect theory is therefore *smaller* than one given by the standard solution.<sup>5</sup> It may be surprising that the fact that individuals dislike losses – in the present framework ending up with smaller income than the average – results in a smaller marginal tax rate than in the standard case. However, it follows from a desire by the government to reduce the possibility that an individual ends up with a small income. Then it is in the individuals' and the government's interest to reduce the marginal tax rate to induce more effort. A reduction in the marginal tax rate is also understandable due to the fact that that for a range of income, individuals are risk loving, i.e.  $e'' > 0$ . Then the need for insurance is diminished. When income is above the reference level,  $e$  has a standard risk aversion feature, i.e.  $\delta_e = \frac{e''}{e'} < 0$ . Then the income tax offers insurance.

Of course, this kind of comparative static analysis treats the Lagrangean multipliers as fixed when changing the parameters and the functions of the model.<sup>6</sup> It is clear that the results then are at best an approximation. To gain better understanding of the form of optimal schedule would require numerical simulations. They are challenging in the present case with a partly convex objective function. This is a topic for further research.

In this class of models, the sign of the change in the marginal tax rate – i.e. whether the marginal tax rate increases or decreases with income – cannot be solved in general. However, one can find a formulation that helps gain intuition of the terms affecting the marginal tax. The derivative of the marginal tax rate can be written as follows:

$$(25) \quad \frac{\partial MTR}{\partial z} = \frac{\alpha}{\xi} (e''x'g' + e'g'') - \frac{\alpha e'g'}{\xi^2} [v'''x' + \alpha(e'''x'g + e''g')],$$

where  $\xi = v'' + \alpha e''g$ . The sign of (25) remains ambiguous in general. However, one can notice, following Low and Maldoom (2004), that the marginal tax rate is the less progressive, the higher is  $v'''$ . The third derivative measures the importance of precautionary incentives ('prudence'). Other things equal, an increase in prudential behaviour reduces the progressivity of the marginal tax rate. In the present context, the progressivity also depends on  $e'''$ . Unlike

<sup>5</sup> This effect can also be found from the formulation (20) of the marginal tax rate by considering a small increase in  $e''$  on the MTR, and noting that the first term in the brackets is negative.

<sup>6</sup> The problem is similar in standard analysis of moral hazard, such as in Varian (1980).

for prudence, similar intuition has not been developed, to our knowledge, for the third derivative of the value function contained in the prospect theory. One can note that if  $e'''$  is positive, loss aversion decreases with income, and the marginal income tax rate becomes less progressive.

There is one element of prospect theory which we have not addressed yet. This is the possibility that the government and individuals perceive the distribution of  $z$  in a different manner. Indeed prospect theory also maintains that individuals do not assess probabilities with the true density function, say  $f$ , but with some transformation  $\pi(f)$ , which underweighs high probabilities and overweighs small probabilities (Kahneman and Tversky 1979).

This has potentially two problems. First, the weighting schemes do not always satisfy stochastic dominance. Second, weighting schemes may not necessarily be easily extended to a continuum of outcomes. These problems are addressed in the cumulative version of prospect theory (Tversky and Kahneman 1992). In this context, one must assume that the distribution function used by the individuals have some of the standard properties for the first-order approach to be valid (monotone likelihood ratio and convexity of the distribution function). In principle, there is no particular reason why these could not hold. While the level of density function differs between the true  $f$  and  $\pi(f)$ , it does mean that the properties of the distribution function should change.

Introducing  $\pi(f)$  to the optimisation problem is straightforward.<sup>7</sup> Then in the optimal tax rule the government corrects ‘wrong’ perceptions by the individuals in its marginal tax rate, (24), through newly defined  $g = \frac{\pi f_y}{\pi}$ . This observation bears interesting resemblance to the merit good analysis of Sandmo (1983). He builds on the idea that individuals may misperceive probabilities of uncertain outcomes. The ex post economic outcome, resulting from misinformed choices, is therefore inefficient. Instead of evaluating welfare based on individuals’ expected utility, he proposes to specify the social welfare function in terms of individuals’ ex post or realised utility levels. Such a formulation means that consumer sovereignty is only disregarded ex ante. The case for government intervention arises because

---

<sup>7</sup> For brevity, these calculations are not presented here. They are available from the authors upon request.

consumer demands do not adequately represent their tastes due to distorted information. Similar reasoning applies here. The marginal tax rate differs from one in the standard model by correcting the way misperceived probabilities affect the choice for effort.

Finally, one can also consider a case where both the government and the individuals apply prospect theory as the basis of decision making. This situation can be called “prospect theory welfarism”, as the government accepts the individuals' valuation for evaluation of social welfare. This problem can be analysed as a special case of the model in Section 3, when the objective function for the government and for the individual has the properties stipulated in (22). The analytics of this case are presented in full in the Appendix, but the results remain broadly the same. The first-order approach is not valid for the full range of income, randomisation becomes optimal for incomes below the reference point, and above the preference point, the marginal tax rate provides some, less than full, insurance. A key difference is that the formula for the marginal tax rate for income above the reference point is similar to the standard solution of Section 2, and does not include a term correcting for the difference between private and social valuation (as they are the same).

## **5. Endogenous reference point**

Prospect theory brings to center stage the reference level of income around which the value function is defined. But what determines this reference point? If we assume that the reference income is somehow related to the overall economy then we are in a terrain familiar to optimal tax theory. Oswald (1983) and Tuomala (1990), for example, consider the implication of utility interdependence (or 'envy') – the situation in which individual's utility is negatively affected by others' income – on optimal income taxation. But in these models differences are due to differences in ex ante skill levels. In the Mirrlees moral hazard model, differences arise due to luck.

An additional problem related to utility interdependence is that it is not clear whether it should be allowed to enter the social welfare function: is envy a trait one wants to honour? The non-welfarist approach, such as ours, does not suffer from this criticism. Utility interdependence affects the way people behave, which the government must take into account

as a constraint when designing tax schedules, but allowing for envy need not be included in the government objective function.

We endogenise the reference income  $\bar{x}$  in the following way. As in utility interdependence models, suppose it is a function of aggregate consumption as follows

$$(26) \quad \bar{x} = \int \phi(z)x(z) f dz$$

where  $\phi$  are the weights given for each individual's consumption. Simple examples are where the sum of  $\phi$ :s is equal to one, when  $\bar{x}$  is equal to aggregate consumption, or when  $\phi = \frac{1}{n}$  where  $n$  is the number of individuals, when  $\bar{x}$  is equal to average consumption. With no loss in generality, here we concentrate on the former.

Denoting the Lagrange multiplier related to the constraint in (26) as  $\mu$ , the Lagrangian with endogenous  $\bar{x}$  looks like:

$$(27) \quad L = \int \{ [v(x) + \lambda(z-x) + \mu x] f(z, y) + \alpha [e(x - \bar{x})] f_y \} dz - \alpha - y - \mu \bar{x}$$

The first-order conditions for  $x$  and  $\bar{x}$  are

$$(28) \quad v'f + \alpha e'f_y - \lambda f + \mu f = 0$$

$$(29) \quad -\alpha e'f_y - \mu = 0$$

From (29), one notices that  $\mu = -\alpha e'f_y < 0$ . Rewrite (28) as

$$(30) \quad 1 + \alpha \frac{e'}{v'} g = (\lambda - \mu) \frac{1}{v'}$$

Differentiation of (30) yields

$$(31) \quad x' \left[ \alpha g \frac{v''e' - e''v'}{v'} + (\mu - \lambda) \frac{v''}{v'} \right] = \alpha e' g'$$

Rearranging of (31) gives

$$(32) \quad x' [(\lambda - \mu - \alpha g e') \delta_v - \alpha g e' \delta_e] = \alpha e' g'$$

Utilising the definitions  $\delta_v = -v''/v'$  and  $\delta_e = e''/e'$ , and the property that  $\lambda - \mu - \alpha e' g = v'$  (from equation 28), one can write

$$(33) \quad x' = \frac{\alpha e' g'}{v' \delta_v - \alpha e' g \delta_e}$$

and

$$(34) \quad MTR = 1 - x' = 1 - \frac{\alpha e' g'}{v' \delta_v - \alpha e' g \delta_e}$$

Now for the FOA to be valid,  $x$  must again be increasing in  $z$ . Assuming that  $\alpha > 0$ , the numerator of (33) is positive. Using (29), the denominator is positive if  $v' \delta_v > \delta_e (\lambda - v' - \mu)$ , i.e if risk aversion is sufficiently higher than loss aversion. Because of  $\mu < 0$ , the condition is more stringent than in the case, derived in the previous section, where the reference income was exogenous. The intuition is that an increase in an individual's income now has a smaller social value, as it reduces the perceived utility by others.

With some manipulation, one can then show that  $\alpha$  is indeed positive in this setting as well,<sup>8</sup> and the rest of the conditions stay the same. Therefore, the conditions for the FOA to be valid are the monotone like ratio property, the convexity of the distribution function, and that risk aversion is higher than a new threshold level that is dependent on loss aversion and the magnitude of envy. If the last condition is not satisfied, randomisation again occurs.

---

<sup>8</sup> A crucial step for showing this is to demonstrate that a counterpart of equation (17) can be derived in the presence of  $\mu$  in the first-order condition (28). This is possible by substituting for  $\mu$  from (29) in (28).

For the marginal tax rate, the following result emerges:

**Proposition 5.** *When the reference income is a positive function of aggregate income, the optimal incentive structure for incomes below the reference point ( $x < \bar{x}$ ) is a randomised schedule  $(x_0, \bar{x})$  if individuals' loss aversion outweighs the government's risk aversion. If the government's risk aversion outweighs individuals' loss-aversion or income is above the reference point ( $x > \bar{x}$ ), the FOA is valid and marginal tax rate tends to rise in comparison to the case with exogenous reference level.*

The first part of this follows from the invalidity of the FOA. The second part is due to the presence of  $\mu < 0$  in (34). By (29), It increases  $v'$ , and therefore increases the marginal tax rate compared to the case where  $\bar{x}$  exogenous. The intuition is that it is optimal to reduce the incentives for exerting effort as higher effort and higher income have a negative externality by increasing the reference income of other individuals in the economy. The presence of  $\mu$  therefore introduces a corrective (or externality-internalising) element to the marginal tax rate.

A similar effect has been earlier found in utility interdependence models in the conventional income tax framework. The intuition obtained there therefore carries over to the present moral hazard case. It is interesting that while the general tax rules in the moral hazard and the adverse selection cases have little in common, the influence of utility interdependence on tax rates in both cases is similar. Note finally that the presence of utility interdependence does not change the incentive structure below the reference income, as the solution is not continuous.

## 6. Conclusion

This study analysed a model of optimal non-linear income taxation under income uncertainty, along the lines of Mirrlees (1974), from a non-welfarist point of view where the government's and the individuals' objective functions differ. It then interpreted the optimal income tax rule derived under non-welfarism assuming that the individuals' decision making follows prospect theory as developed by Kahneman and Tversky (1979).

As does most of the literature in the area, we focused on the so-called first-order approach for solving the optimisation problem. The conditions for the validity of this approach include a novel requirement in the non-welfarist case. And the marginal income tax rate was shown to be a combination of the standard, welfarist, rule, and a new term that corrects for the differences between individuals' and government's assessment of welfare.

The non-welfarist tax rule becomes particularly relevant when agent's behaviour is interpreted using prospect theory instead of expected utility maximisation. It appeared that changing the underlying assumption of agent behaviour in moral hazard models has a surprisingly drastic influence on optimal compensation structure / marginal tax rates.

In particular, one can show that the solution is non-monotonic under prospect theory. This follows from loss aversion, in other words convexity of the value function, implied by prospect theory, when the consumers are in fact risk-loving for a certain range of low incomes. If consumers' loss aversion is sufficiently strong and it offsets the government's risk aversion, it becomes optimal to offer a randomised tax schedule for low incomes. A corollary of this finding is that the solution procedure commonly applied in this sort of models, the first-order approach, is not valid for the full range of income. If government's loss aversion outweighs individuals' risk aversion, the optimal solution is a function of both risk aversion and loss aversion. In this case the optimal marginal tax rate under prospect theory is likely to be smaller than in the conventional moral hazard models for small incomes. By this policy, the government attempts to enhance effort to reduce the possibility that individuals incur losses from low income.

These findings have potentially important implications. For income taxation models, some of the results indicate a more likely case for progressive income taxation than one derived from conventional models. For economic research in general, the results of this paper show that it may become worthwhile assessing the robustness of other results, derived using expected utility maximisation, to assumptions that are more in line with findings in behavioural economics.

## References

- Alvi, E. (1997) 'First-Order Approach to the Principal-Agent Problems: A Generalization', *The Geneva Papers on Risk and Insurance Theory* 22, 59-65.
- Arnott, R. and J.E. Stiglitz (1988) 'Randomization with asymmetric information', *RAND Journal of Economics* 19, 344-362.
- Camerer, C.F. and G. Loewenstein (2003) 'Behavioral economics: Past, Present, Future' in C.F. Camerer, G. Loewenstein and M. Rabin (eds.) *Advances in Behavioral Economics*. Princeton: University Press.
- Carroll, C.D., J. Overland and D.N. Weil (2000) 'Saving and growth with habit formation', *American Economic Review* 90, 341-355.
- Holmström, B. (1979) 'Moral hazard and observability', *Bell Journal of Economics* 10, 74-91.
- Jewitt, I. (1988), Justifying the first order approach to principal-agent problems, *Econometrica* 56, 1177-90.
- Kahneman, D. (2003) 'Maps of bounded rationality: Psychology for behavioral economics', *American Economic Review* 93, 1449-1475.
- Kahneman, D. and A. Tversky (1979) 'Prospect theory: An analysis of decision under risk' *Econometrica* 47, 263-281.
- Kanbur, R., Keen, M. and Tuomala, M. (1994) 'Labor supply and targeting in poverty alleviation programs', *The World Bank Economic Review* 8, 191-211.
- Laffont, J.-J. and D. Martimort (2002) *The theory of incentives. The principal-agent model*. Princeton University Press.
- Ljunqvist, L. and H. Uhlig (2000) 'Tax policy and aggregate demand management under catching up with the Joneses', *American Economic Review* 90, 356-366.
- Low, H. and Maldoom, D. (2004) 'Optimal taxation, prudence and risk-sharing', *Journal of Public Economics* 88, 443-464.
- Mirrlees, J.A. (1974), 'Notes on welfare economics, information and uncertainty', in Balch, McFadden and Wu (Eds.), *Essays on Economic Behaviour under Uncertainty*, Amsterdam: North Holland.

- Mirrlees, J.A (1975,1999) ‘The Theory of Moral Hazard and Unobservable Behaviour. Part I’, *Review of Economic Studies* 66, 3-22.
- Mirrlees, J.A. (1976) ‘The optimal structure of authority and incentives within an organization’, *Bell Journal of Economics* 7, 105-31.
- Mirrlees, J.A. (1997) ‘Information and incentives: The economics of carrot and sticks’, *The Economic Journal* 107, 1311-1329.
- O’Donoghue, T. and M. Rabin (2003) ‘Studying optimal paternalism, illustrated by a model of sin taxes’, *American Economic Review* 93, 186-191.
- Oswald, A. (1983) ‘Altruism, jealousy and the theory of optimal nonlinear taxation’, *Journal of Public Economics* 20, 77-87.
- Pirttilä, J. and M. Tuomala (2004) ‘Poverty alleviation and tax policy’, *European Economic Review*, forthcoming.
- Rogerson, W. (1985) ‘The First-Order Approach to Principal-Agent Problems’, *Econometrica* 53,1357-67.
- Sandmo, A. (1983) ‘Ex post welfare economics and the theory of merit goods’, *Economica* 50, 19-33.
- Seade, J. (1980) ‘Optimal non-linear policies for non-utilitarian motives’, in D. Collard, R. Lecomber and M. Slater (eds.) *Income distribution: the limits to redistribution*, Bristol: Scientechnica.
- Tuomala, M. (1990) *Optimal income tax and redistribution*, Oxford: Clarendon Press.
- Tuomala, M. (1984) ‘Optimal degree of progressivity under income uncertainty’, *Scandinavian Journal of Economics*,87,184-93.
- Tversky, A. and D. Kahneman (1992) ‘Advances in prospect theory: Cumulative representation of uncertainty’, *Journal of Risk and Uncertainty* 5, 297-323.
- Varian, H. (1980), ‘Redistributive taxation as social insurance’, *Journal of Public Economics* 14, 49-68.

## Appendix: Prospect theory welfarism

The government's objective function is now the same as the individuals, i.e.,  $e$ . The Lagrangean and the first-order condition with respect to  $x$  (pointwise optimisation) are as follows:

$$(A.1) \quad L = \int \{ [e(x - \bar{x}) + \lambda(z - x)] f(z, y) + \alpha e(x - \bar{x}) f_y \} dz - \alpha - y$$

$$(A.2) \quad 1 + \alpha g = \frac{\lambda}{e'}$$

Differentiate (A.2) again with respect to  $z$  and reorganise to obtain the shape of  $x'(z)$ :

$$(A.3) \quad x' = -\frac{\alpha(e')^2 g'}{\lambda e''}$$

To provide incentives for exerting effort,  $x$  should be increasing in  $z$ . Depending on whether realised income is above or below the reference income,  $\bar{x}$ , three cases emerge.

1) For income above the reference income,  $x > \bar{x}$ , consumption  $x$  is indeed increasing in income  $z$ , since the value function has similar properties to the standard case of Section 2. In other words,  $e = v$ . The first-order approach is valid (following the arguments of equations (5)-(9), and the marginal tax rate is given by (11).

2) If  $x = \bar{x}$ , the right-hand side of (A.3) is not defined.

3) If the income is below the reference income,  $x < \bar{x}$ , the right-hand side of (A.3), i.e.  $x'(z)$ , is non-negative only if  $\alpha < 0$ , since  $e'' > 0$ . Using a similar procedure as in Section 2, one can determine the sign of  $\alpha$  from

$$(A.4) \quad \frac{\alpha}{\lambda} \int e f_y dz = \text{cov}(e, \frac{1}{e'})$$

$$\frac{\alpha}{\lambda} = \text{cov}(e, \frac{1}{e'})$$

Equation (A.4) is a counterpart of earlier equation (8). Now  $e$  and  $e'$  covary in the *same* directions for  $x < \bar{x}$ , we necessarily have  $\alpha \leq 0$ . However, the only case where the covariance is zero is when  $x$  is constant irrespective of income. But then the worker has no incentives to provide positive effort. Therefore, to induce effort,  $\alpha < 0$ .

However, if  $\alpha < 0$ , a relaxation in the incentive constraint reduces the social welfare determined by (A.1). For a meaningful government optimisation problem this cannot hold. Therefore, one must conclude that the FOA is not valid for income below the reference level. As in prospect theory non-welfarism, the first-order approach is not a valid solution procedure for the entire range of realised income.

The optimal incentive solution is again non-continuous. For incomes at or below the reference point ( $x \leq \bar{x}$ ), it is a randomised schedule  $(x_0, \bar{x})$ . For income above the reference point ( $x > \bar{x}$ ), the FOA is valid and the standard solution, given in equation (11), applies.