



**Optimal Income Taxation  
and Public-Goods Provision  
with Preference and  
Productivity Shocks**

Felix Bierbrauer





# **Optimal Income Taxation and Public-Goods Provision with Preference and Productivity Shocks**

Felix Bierbrauer

May 2010

# Optimal Income Taxation and Public-Goods Provision with Preference and Productivity Shocks

Felix Bierbrauer\*

Max Planck Institute, Bonn, Germany

May 6, 2010

## Abstract

We study how an optimal income tax and an optimal public-goods provision rule respond to preference and productivity shocks. A conventional Mirrleesian treatment is shown to provoke manipulations of the policy mechanism by individuals with similar interests. We therefore extend the Mirrleesian model so as to include a requirement of coalition-proofness. The main results are the following: first, the possibility of preference shocks yields a new set of collective incentive constraints. Productivity shocks have no such implication. Second, the optimal policy gives rise to a positive correlation between the public-goods provision level, the extent of redistribution and marginal tax rates.

*Keywords:* Optimal Taxation, Public goods, Mechanism Design.

*JEL:* D71, D82, H21, H41

---

\*I am indebted to Martin Hellwig for numerous discussions about the material in this paper. This paper builds on our joint work on “Public-Good Provision in a Large Economy.” I also benefited from conversations with Sophie Bade, Marco Bassetto, Pierre Boyer, Georges Casamatta, Christoph Engel, Mike Golosov, Stephan Lauer mann, Marco Sahm, Joel Slemrod, Christian Traxler, Aleh Tsyvinski, Georg von Heusinger and John Weymark. I thank participants at the 2009 Decentralization Conference in St. Louis, seminar participants at the University of Heidelberg, WZB in Berlin, Tor Vergata in Rome, Toulouse School of Economics, University of Michigan and the Chicago Fed. I am also grateful for the hospitality of MIT where part of this research was conducted.

# 1 Introduction

A society that wants to provide public goods and redistribute income and therefore has to tax individuals faces a number of information problems. Given that taxes paid and transfers received should reflect an individual's ability to generate income, each individual's earning ability has to be determined. In addition, information on preferences for public goods has to be acquired because an optimal public expenditure policy requires an assessment of the social costs and benefits of public spending.

The theory of optimal taxation in the tradition of Mirrlees (1971) focusses on the problem to tax individuals according to their earning ability. The optimal policy is therefore the solution of a screening problem, i.e., for any one individual the problem is to determine this individual's characteristics so that the individual can be taxed accordingly. In this literature, problems of information aggregation do not arise; e.g., there is no issue of having to acquire the information on how many individuals have a high earning ability. Also, for extended versions of this model that include a decision on public-goods provision, there is no need to acquire the information on how many individuals value a public good highly.<sup>1</sup> These aggregates are taken to be known quantities.

The theory of public-goods provision in the tradition of Clarke (1971) and Groves (1973), by contrast, focusses on problems of information aggregation. In this literature, information on the public-goods preferences of any one individual has to be acquired because it is an essential input for the determination of the social benefits from public-goods provision. This literature, however, disregards the production side of the economy and the tax system as an alternative source of public-goods finance. Also, it does not include distributive considerations which are based on individual differences in productive abilities.

This paper provides a unified approach to these issues so that we can simultaneously analyze problems of optimal taxation and problems of information aggregation. This makes it possible to provide answers to the following questions: should the tax system become more redistributive if the average worker becomes more productive? What are the implications of such a productivity shock for public-goods provision? Should public spending expand if the demand for public goods goes up? If so, what are the implications of such a preference shock for the shape of the tax system?

The formal analysis is based on a large economy model with endogenous production, as is the theory of optimal taxation, and uses a mechanism design approach.<sup>2</sup> The economy is populated by high-skilled and by low-skilled individuals, who either have a high or a low preference for public goods. A *state* of the economy is identified with a cross-section distribution of those characteristics; that is, a state is a triplet consisting of the population share of high-skilled

---

<sup>1</sup>See, for example, Boadway and Keen (1993), Gahvari (2006), or Kreiner and Verdelin (2010).

<sup>2</sup>The paper thus contributes to a recent literature in public economics which uses a mechanism design approach in order to characterize optimal insurance contracts or tax systems; see, for example, Golosov et al. (2003), or Kocherlakota (2005). Predecessors are Hammond (1979) and Guesnerie (1995). The work that is most closely related to this paper is by Bassetto and Phelan (2008) and Kocherlakota and Phelan (2009), who are also concerned with the characterization of optimal policies in large economies with aggregate uncertainty. However, none of these papers includes an analysis of public-goods provision.

individuals, the fraction of high-skilled individuals with a high taste for public goods, and the fraction of low-skilled individuals with a high taste for public goods. A mechanism specifies how the income tax schedule and the public-goods provision level vary with the state of the economy; that is, how fiscal policy responds to preference and productivity shocks. The aim of the paper is to characterize the *optimal* mechanism, or, equivalently, the optimal response to preference and productivity shocks.

Our mechanism design approach invokes a requirement of robustness with respect to the specification of the individuals' probabilistic beliefs.<sup>3</sup> As has been shown in Bierbrauer (2009), robust mechanism design in a large economy yields a framework that is equivalent to a Mirrleesian model of income taxation and public-goods provision. This contribution of the present paper is to introduce, in addition, a requirement of *coalition-proofness*. This requirement is motivated by the observation that the Mirrleesian model is vulnerable to collective manipulations by like-minded individuals. We show that *any* mechanism which implements the Mirrleesian outcome has alternative equilibria with at least one the following properties:

(i) *Individuals benefit from lies about their public-goods preferences.* High-skilled individuals may have an incentive to exaggerate their public-goods preferences and low-skilled individuals may have an incentive to understate theirs. The reason is that low-skilled individuals suffer more from the need to pay for the public good because they have a harder time generating income. Since a utilitarian policy maker does not give full weight to the utility loss of the low-skilled, these individuals have an incentive to lie about their public-goods preferences so as to convince the policy maker that the provision level should be reduced. (A symmetric argument explains why the high-skilled have an incentive to exaggerate their preferences.)

(ii) *Individuals benefit from lies about their productive abilities.* Suppose, for the sake of the argument, that an optimal income tax is more redistributive if the economy is “rich”, in the sense that there are many high-skilled individuals. Consequently, if enough high-skilled individuals lie about their productive abilities, this causes the policy maker to believe that the economy is poorer than it actually is, so that redistribution is reduced and the high-skilled are better off.

The problems in (i) and (ii) have a common cause: the implementability of the Mirrleesian outcome rests on the assumption that individuals behave truthfully simply because, in a large economy, a unilateral change of behavior would neither make a difference for the public-goods provision level, nor, for the income tax schedule. However, individuals are not indifferent regarding the policy that is implemented. Hence, from a theoretical perspective, if individuals can coordinate on an alternative equilibrium that is more attractive to them, it is unconvincing to assume that the truthful equilibrium will be played. Also, from an empirical point of view, if one thinks about political parties and special interest groups, it is plausible that individuals with common interests manage to induce policies that are favorable to them.

To address this concern, this paper uses the notion of *coalition-proof implementation in a large economy*, which has been developed in Bierbrauer and Hellwig (2010). In this approach, individuals are given the possibility to coordinate their communication with the policy-maker

---

<sup>3</sup>This notion of robustness has been introduced by Bergemann and Morris (2005) in an attempt to reduce the reliance on specific common prior assumptions in the theory of mechanism design.

so as to take advantage of the possibility that, if sufficiently many individuals lie about their characteristics, this affects the policy maker's perception of the state of the economy and hence the policy that is ultimately chosen. Coalition-proofness fails if there is an alternative equilibrium in which a group of individuals lies about their characteristics, and, moreover, benefits from the change in the policy that is induced by this deviation.<sup>4</sup>

Bierbrauer and Hellwig (2010) study the provision of an indivisible public good which is either provided or not. Moreover, individuals differ only in their public-goods preferences. The present paper extends this analysis in various directions. Individuals now differ both in public-goods preferences and in productive abilities. Moreover, the distribution of public-goods preferences and the distribution of productive abilities are a priori unknown. The public-goods provision level can be continuously adjusted, and, finally, production is endogenous. In particular, these extensions make it possible to study the interdependence of optimal tax and expenditure policies.

The main part of the analysis is concerned with the characterization of an optimal rule for income taxation and public goods-provision that is both robust and coalition-proof. This yields two main results.

First, there is a fundamental difference between preference and productivity shocks: while the possibility of productivity shocks has essentially no bearing on the set of admissible policies, the possibility of preference shocks leads to a new set of *collective incentive constraints*. Lies about productive abilities can be deterred by minor adjustments of the income tax schedule which ensure that a truthful communication of abilities is each individual's unique best response. Lies about public-goods preferences cannot be addressed in this way. Since the economy is large so that no individual is pivotal for how much of a public good is provided, it is impossible to provide incentives for a truthful communication of preferences. Hence, the policy rule has to deter these lies by making their consequences unattractive. This gives rise to an additional set of constraints.

Second, we characterize the optimal mechanism that satisfies these collective incentive constraints. We show that the optimal mechanism displays a *complementarity* between public-goods provision, redistribution, and marginal tax rates; i.e., deviations from the conventional Mirrleesian model take one of the following forms:

*Upward distortions.* The public-goods provision level is higher than stipulated by the Samuelson rule, there is more redistribution, and marginal tax rates are higher.

*Downward distortions.* The public-goods provision level is lower than stipulated by the Samuelson rule, there is less redistribution, and marginal tax rates are lower.

Moreover, upward distortions are associated with states in which many individuals have a high preference for the public good, whereas downward distortions are associated with states in which many individuals have low preference for the public good. At an empirical level, these results imply that we should observe a positive correlation between the public-goods provision level, the

---

<sup>4</sup>This approach has been inspired by the work of Laffont and Martimort (1997, 2000) who treat the formation of a deviating coalition as a mechanism design problem with its own set of incentive and participation constraints.

level of redistribution, and marginal tax rates. We would not predict such a correlation on the basis of a conventional Mirrleesian model that does not include collective incentive constraints.

The remainder of the paper is organized as follows: Section 2 describes the economic environment. As a benchmark, Section 3 reviews Mirrleesian approach to income taxation and public goods provision and relates it to a model of robust mechanism design. In Section 4, we introduce the solution concept of a robust and coalition-proof Bayes-Nash equilibrium. The optimal robust and coalition-proof mechanism for income taxation and public-goods provision is characterized in Section 5. The last section contains concluding remarks. All proofs are in the Appendix.

## 2 The Environment

### Payoffs and social choice functions

There is a continuum of individuals identified with the unit interval  $I = [0, 1]$ . Individual  $i$ 's utility function is given by

$$U(q, c, y, w^i, \theta^i) = \theta^i q + u(c) - \frac{y}{w^i},$$

where  $q$  is the amount of a public good,  $c$  is the individual's consumption of a private good and  $y$  is the individual's contribution to the economy's output. Individual  $i$ 's utility from the public good depends on a taste parameter  $\theta^i$  which either takes a high or a low value; for all  $i$ ,  $\theta^i \in \Theta = \{\theta_L, \theta_H\}$ , where  $0 < \theta_L < \theta_H$ . The function  $u$  gives utility from private-goods consumption and is assumed to be strictly increasing and strictly concave. The disutility from productive effort depends on a skill parameter  $w^i$ , which, again, takes either a high or a low value; for all  $i$ ,  $w^i \in W = \{w_L, w_H\}$ , where  $0 < w_L < w_H$ . Individuals are privately informed about their taste parameter and about their skill level. To simplify the exposition, we assume that  $\theta_L = w_L$  and that  $\theta_H = w_H$ .

A *state* of the economy is identified with a cross-section distribution of productivity and preference parameters. Formally, a state  $s$  of the economy is a triple  $s = (f_H, p_H, p_L)$ , where  $f_H$  is the population share of individuals with a high taste parameter,  $p_H$  is the fraction of high-skilled individuals with a high taste parameter, and  $p_L$  is the fraction of low-skilled individuals with a high taste parameter. The set of states is in the following denoted by  $S = [0, 1]^3$ .

A social choice function formalizes the dependence of outcomes on the state of the economy. It consists of a provision rule for the public good  $q : S \mapsto \mathbb{R}_+$  that specifies for each state how much of the public good is provided. It also specifies an individual's private-goods consumption and output requirement as a function of the state of the economy and the individual's characteristics. Private-goods consumption is determined by the function  $c : S \times W \times \Theta \mapsto \mathbb{R}_+$ , and the output requirement is determined by  $y : S \times W \times \Theta \mapsto \mathbb{R}_+$ .

A social choice function is said to be feasible, if, for every  $s$ ,

$$\begin{aligned} & f_H \left( p_H(y(s, w_H, \theta_H) - c(s, w_H, \theta_H)) + (1 - p_H)(y(s, w_H, \theta_L) - c(s, w_H, \theta_L)) \right) \\ & + (1 - f_H) \left( p_L(y(s, w_L, \theta_H) - c(s, w_L, \theta_H)) + (1 - p_L)(y(s, w_L, \theta_L) - c(s, w_L, \theta_L)) \right) \\ & \geq r(q(s)), \end{aligned} \quad (1)$$

where  $r$  is a strictly increasing and strictly convex cost function which captures the resource requirement of public-good provision.

## Types and beliefs

The analysis below focusses on social choice functions that are robustly implementable in the sense that their implementability does not rely on assumptions about the individuals' probabilistic beliefs. This notion of robustness is formally defined below. As a preliminary step, we introduce the notion of a *type space*, which we borrow from Bergemann and Morris (2005). This makes it possible to view an individual's type as a two-dimensional object, consisting of a *payoff type* affecting the individuals' preferences, and a *belief type*.

More formally, let  $(T, \mathcal{T})$  be a measurable space,  $\pi = (w, \theta)$  be a measurable map from  $T$  into  $W \times \Theta$ , and  $\beta$  a measurable map from  $T$  into the space  $\mathcal{M}(\mathcal{M}(T))$  of probability distributions over measures on  $T$ . We interpret  $t^i \in T$  as the abstract type of agent  $i$ ,  $\pi(t^i) = (w(t^i), \theta(t^i))$  as the payoff type of agent  $i$  and  $\beta(t^i)$  as the belief type of agent  $i$ . We assume throughout that the function  $\pi$  is surjective.

From an individual's perspective, the cross-section distribution of types, henceforth denoted by  $\delta$ , is a random variable. The belief type  $\beta(t^i)$  indicates the agent's probabilistic beliefs about  $\delta$ . Thus, for any  $X \subset \mathcal{M}(T)$ ,  $\beta(X | t^i)$  is the probability that agent  $i$  assigns to the event  $\delta \in X$ . We refer to the map  $\beta : T \rightarrow \mathcal{M}(\mathcal{M}(T))$  as the *belief system* of the economy.

A given belief system specifies, in particular, an individual's beliefs about the payoff types of other individuals. To see this, note that each  $\delta \in \mathcal{M}(T)$  induces a cross-section distribution of payoff types  $s(\delta) := \delta \circ \pi^{-1}$ .

We assume that the measures  $\beta(t)$ ,  $t \in T$ , are mutually absolutely continuous, i.e., that they all have the same null sets. We refer to this property by saying that the belief system is *moderately uninformative*. If the belief system is moderately uninformative, observation of the event  $t^i = t$  does not permit agent  $i$  to rule out any event that has positive probability with some other specification of beliefs.<sup>5</sup>

## Mechanisms and robust implementation

We seek to implement a social choice function by means of an allocation mechanism  $M = [(A, \mathcal{A}), Q, C, Y]$ , where  $(A, \mathcal{A})$  is a measurable space, and  $A$  is the set of actions that individuals can take.<sup>6</sup> The function  $Q : \mathcal{M}(A) \rightarrow \mathbb{R}_+$  gives the public-good provision level as a

<sup>5</sup>We can leave open whether or not these beliefs are derived from a common prior. For a discussion of moderately uninformative belief systems under a common prior assumption, see Bierbrauer and Hellwig (2010).

<sup>6</sup>We do not (yet) restrict attention to direct mechanism and to truth-telling equilibria because, for the coalition-proof Bayes-Nash equilibria that will be studied below, the revelation principle does not generally hold.

function of the cross-sectional distribution of actions, and the functions  $C : \mathcal{M}(A) \times A \rightarrow \mathbb{R}_+$  and  $Y : \mathcal{M}(A) \times A \rightarrow \mathbb{R}_+$  specify a consumption level  $C$  and an output requirement  $Y$ , respectively, as a function of an individual's message and of the cross-section distribution of messages.

The mechanisms that we consider are anonymous in the sense that the decision on public-goods provision depends only on the cross-section distribution of actions that the mechanism designer receives. Also, an individual's consumption level and output requirement depend on the own action and, again, the distribution of actions. Since the economy is large, a single individual cannot affect the distribution of actions that the mechanism designer receives. In particular, this implies that no single individual can influence the public-goods provision level, or the consumption levels and output requirements of other individuals.

A mechanism induces a game. In the following section, as a benchmark, we focus on Bayes-Nash equilibria. (The additional requirement of coalition-proofness will be introduced in Section 4.) With this solution concept, a social choice function is said to be implementable on a given type space if, for this type space, there exists a mechanism  $M$ , and a Bayes-Nash equilibrium so that the equilibrium outcome is equal to the outcome stipulated by the social choice function. It is *robustly implementable* if, for every  $(T, \mathcal{T})$ , and  $\pi : T \rightarrow W \times \Theta$ , there exists a mechanism that implements it on the type space  $[(T, \mathcal{T}), \pi, \beta]$ , for every moderately uninformative belief system  $\beta$ .<sup>7</sup>

### 3 A Mirrleesian approach

Prior to introduction of coalition-proofness in the subsequent section, as a benchmark, we characterize the social choice functions that are robustly implementable as a Bayes-Nash equilibrium. In particular, we show that the problem of choosing an optimal robust social choice function is equivalent to a Mirrleesian problem of optimal income taxation and public-goods provision. Finally, we will demonstrate that the optimal Mirrleesian policy provokes manipulations of the policy outcome by like-minded individuals.

#### Implications of robust implementability as a Bayes-Nash equilibrium

**Proposition 1** *The following statements are equivalent.*

- (a) *A social choice function  $(q, c, y)$  is robustly implementable as a Bayes-Nash equilibrium.*
- (b) *A social choice function  $(q, c, y)$  satisfies the following individual incentive compatibility constraints: For every  $s \in S$ , every  $(w, \theta) \in W \times \Theta$ , and every  $(\hat{w}, \hat{\theta}) \in W \times \Theta$ ,*

$$\theta q(s) + u(c(s, w, \theta)) - \frac{y(s, w, \theta)}{w} \geq \theta q(s) + u(c(s, \hat{w}, \hat{\theta})) - \frac{y(s, \hat{w}, \hat{\theta})}{w}. \quad (2)$$

---

<sup>7</sup>Our notion of robustness is slightly stronger than that of Bergemann and Morris (2005). Following Ledyard (1978), we require that the same mechanism is used whatever the belief system is. In contrast, Bergemann and Morris allow mechanisms to depend on the belief system. While this has no bearing on the set of robustly implementable social choice functions, it matters for the formulation of a robust mechanism design problem.

Proposition 1 adapts arguments by Ledyard (1978) and Bergemann and Morris (2005) to the given large economy setup. The individual incentive compatibility constraints can be interpreted as follows: a truthful revelation of types must be an ex post equilibrium; i.e., once the state of the economy has been revealed, no individual regrets having reported his characteristics truthfully to the mechanism designer.

The incentive compatibility constraints in (2) can be equivalently written as follows: for every  $s \in S$  and every  $(w, \theta) \in W \times \Theta$ ,

$$u(c(s, w, \theta)) - \frac{y(s, w, \theta)}{w} \geq u(c(s, \hat{w}, \hat{\theta})) - \frac{y(s, \hat{w}, \hat{\theta})}{w}, \quad (3)$$

for all  $(\hat{w}, \hat{\theta})$ . The utility that individuals derive from public goods does not matter for incentive compatibility because (i) the economy is large, and (ii) the utility function is separable so that an individual's marginal rate of substitution between consumption  $c$  and output  $y$  does not depend on the supply of public goods.

The inequalities in (3) imply that, for every  $s$ , for every given  $w$  and every pair  $\theta$  and  $\hat{\theta}$ ,

$$u(c(s, w, \theta)) - \frac{y(s, w, \theta)}{w} = u(c(s, w, \hat{\theta})) - \frac{y(s, w, \hat{\theta})}{w}, \quad (4)$$

so that two individuals who differ only in their taste parameter derive the same utility from their respective  $(c, y)$  combination, in every state  $s$ . Given condition (4), it is without loss of generality to assume that also  $c(s, w, \theta) = c(s, w, \hat{\theta})$  and  $y(s, w, \theta) = y(s, w, \hat{\theta})$ , for every  $s$ ,  $w$ , and every pair  $(\theta, \hat{\theta})$ .<sup>8</sup> In the following, we may hence drop the dependence of consumption levels and output requirements on taste parameters and write simply  $c(s, w)$  and  $y(s, w)$ , respectively.

With this notation, we can write the individual incentive compatibility constraints as follows: for every  $s$ , every  $w$ , and every  $\hat{w}$ ,

$$u(c(s, w)) - \frac{y(s, w)}{w} \geq u(c(s, \hat{w})) - \frac{y(s, \hat{w})}{w}. \quad (5)$$

The economy's resource constraint in (1) can now be written as follows: For all  $s = (f_H, p_H, p_L)$ ,

$$f_H(y(s, w_H) - c(s, w_H)) + (1 - f_H)(y(s, w_L) - c(s, w_L)) \geq r(q(s)). \quad (6)$$

It has become common practice to use a mechanism design approach for the analysis of the Mirrleesian income tax problem; that is, instead of assuming that individuals are confronted with an income tax schedule  $T$  that relates their pre-tax-income,  $y$ , to their after-tax-income,  $c$ , and then choose  $y$  and  $c$  in a utility-maximizing way, one looks directly at the social choice functions that permit a decentralization via some income tax schedule.<sup>9</sup> This yields implementability conditions that, for a given  $s$ , coincide with the constraints in (6) and (5). Hence, finding an optimal robustly implementable social choice function is equivalent to the Mirrleesian problem of optimal income taxation.

<sup>8</sup>Any welfare-maximizing social choice function is such that individual utility levels are generated at a minimal resource cost. Hence it must be true that  $y(s, w, \theta) - c(s, w, \theta) = y(s, w, \theta') - c(s, w, \theta')$ . This equality in conjunction with the fact that indifference curves in a  $y - c$  diagram are strictly increasing and strictly convex, yields  $c(s, w, \theta) = c(s, w, \theta')$  and  $y(s, w, \theta) = y(s, w, \theta')$ .

<sup>9</sup>Examples are Stiglitz (1982), Boadway and Keen (1993), Gahvari (2006), or Hellwig (2007).

## The optimal robust social choice function

An optimal utilitarian social choice function solves the following maximization problem: choose  $q : S \rightarrow \mathbb{R}_+$ ,  $c : S \times W \rightarrow \mathbb{R}_+$  and  $y : S \times W \rightarrow \mathbb{R}_+$  in order to maximize expected utilitarian welfare  $E[W(s)]$ , where  $W(s)$  is utilitarian welfare in state  $s$ , subject to the constraints in (6) and (5). However, since there is no constraint that links the outcomes for different states, we may assume, without loss of generality, that each state  $s$  gives rise to its own optimization problem, without repercussions for the outcomes in other states.

Formally, for every  $s$ ,  $q(s)$ ,  $c(s, w_L)$ ,  $y(s, w_L)$ ,  $c(s, w_H)$  and  $y(s, w_H)$  are chosen in order to maximize

$$W(s) = \bar{\theta}(s)q(s) + f_H \left( u(c(s, w_H)) - \frac{y(s, w_H)}{w_H} \right) + (1 - f_H) \left( u(c(s, w_H)) - \frac{y(s, w_H)}{w_H} \right),$$

where

$$\bar{\theta}(s) = (f_H p_H + (1 - f_H) p_L) \theta_H + (f_H (1 - p_H) + (1 - f_H) (1 - p_L)) \theta_L$$

is the population average of the taste parameter in state  $s$ .

As is well-known,<sup>10</sup> the solution to this problem is such that the incentive constraint for the high-skilled individuals is binding,

$$u(c(s, w_H)) - \frac{y(s, w_H)}{w_H} = u(c(s, w_L)) - \frac{y(s, w_L)}{w_H}, \quad (7)$$

and the incentive constraint of the low-skilled individuals is slack,

$$u(c(s, w_L)) - \frac{y(s, w_L)}{w_L} > u(c(s, w_H)) - \frac{y(s, w_H)}{w_L}.$$

Intuitively, the reason is that the utilitarian mechanism designer wants to allocate the same consumption to high-skilled and low-skilled individuals so as to equate their marginal utilities of consumption. At the same time, he wants to have as much output as possible generated by the high-skilled because their marginal effort cost is smaller. Hence, unless the high-skilled individuals' incentive constraint is binding,  $W(s)$  can be increased by lowering  $y(s, w_L)$  and increasing  $y(s, w_H)$ , so that aggregate output remains unchanged.

The resource constraint is also binding,

$$f_H(y(s, w_H) - c(s, w_H)) + (1 - f_H)(y(s, w_L) - c(s, w_L)) = r(q(s)). \quad (8)$$

Otherwise  $y(s, w_L)$  and  $y(s, w_H)$  could both be decreased in a way that maintains incentive compatibility.

Knowing that these constraints are binding, we can use a Lagrangean approach to characterize the optimal choices of  $q(s)$ ,  $c(s, w_L)$ ,  $y(s, w_L)$ ,  $c(s, w_H)$  and  $y(s, w_H)$ . The results from this exercise are summarized in the following proposition which we state without proof.

**Proposition 2** *For every  $s$ , the values of  $q(s)$ ,  $c(s, w_L)$ ,  $y(s, w_L)$ ,  $c(s, w_H)$  and  $y(s, w_H)$  which maximize  $W(s)$  subject to the constraints in (7) and (8) are characterized by the following system of equations:*

---

<sup>10</sup>A formal proof can be found in Weymark (1986) or Hellwig (2007).

i) The optimal consumption levels satisfy

$$u'(c^*(s, w_H)) = \frac{1}{w_H} \quad \text{and} \quad u'(c^*(s, w_L)) = \frac{1}{w_L} \frac{1 - f_H \frac{w_H - w_L}{w_H}}{1 - f_H \frac{w_H - w_L}{w_L}}.$$

The implicit marginal tax rates, which are a measure of how distortionary the income tax system is, are given by

$$\tau^*(s, w_H) := 1 - \frac{1}{w_H u'(c^*(s, w_H))} = 0 \quad \text{and} \quad \tau^*(s, w_L) := 1 - \frac{1}{w_L u'(c^*(s, w_L))} > 0.$$

ii) The optimal public-goods provision level satisfies the Samuleson rule,  $\bar{\theta}(s) = \lambda(s)r'(q^*(s))$ , where  $\lambda(s) := \frac{f_H}{w_H} + \frac{1-f_H}{w_H}$  gives the marginal costs of public funds in state  $s$ .

iii) The optimal output requirements satisfy

$$\begin{aligned} y^*(s, w_H) &= e^*(s) + (1 - f_H)w_H(u(c^*(s, w_H)) - u(c^*(s, w_L))) \quad \text{and} \\ y^*(s, w_L) &= e^*(s) - f_H w_H(u(c^*(s, w_H)) - u(c^*(s, w_L))), \end{aligned}$$

where  $e^*(s) := f_H c^*(s, w_H) + (1 - f_H)c^*(s, w_L) + r(q^*(s))$  denotes aggregate expenditures on public and private goods in state  $s$ .

Proposition 2 makes it possible to analyze how a change in the distribution of productivity or preference parameters affects the optimal policy. This comparative statics exercise is facilitated by the assumptions that preferences are additively separable and that the individuals' effort costs are linear.

For instance, if the economy as a whole becomes “richer” in the sense that the average worker’s productivity increases, i.e., if  $f_H$  goes up, this implies that  $\lambda$  goes down so that there is more public-goods provision. It also leaves the consumption of the high-skilled unaffected, whereas the consumption of the low-skilled goes down. This also implies that the income tax system becomes more distortionary, as reflected by an increase of  $\tau(s, w_L)$ . Hence, an increase in  $f_H$  can be viewed as generating a further deviation from a laissez-faire outcome without redistribution and without distortionary taxation.

We can also analyze how a change in the distribution of public-goods preferences among the high skilled (a change in  $p_H$ ), or among the low-skilled (a change in  $p_L$ ) affects the optimal policy. An increase of  $p_H$  or  $p_L$  generates an increase of the average valuation of the public good,  $\bar{\theta}(s)$ , and hence leads to a higher provision level of the public good. It has neither an impact on the consumption of private goods, nor on marginal tax rates. However, all individuals have to produce more output. More specifically, an increase of the revenue requirement  $r$  by some  $\Delta > 0$  implies that all individuals in the economy have to increase their output by  $\Delta$ .

### Problems with the optimal robust social choice function

In the following, we discuss two examples in order to demonstrate that the implementability of the social choice function in Proposition 2 is questionable because individuals may have an incentive to coordinate their behavior in such a way that the optimal policy is manipulated.

**Example 1: Public-goods preferences.** An increase of the revenue requirement  $r$  by some  $\Delta > 0$  implies that all individuals in the economy have to increase their output by  $\Delta$ , while their consumption levels remain unaffected. For a low-skilled individual this implies a utility loss of  $\frac{\Delta}{w_L}$ , whereas the utility loss for the high-skilled individual equals only  $\frac{\Delta}{w_H}$ . Ceteris paribus, this implies that public-goods provision is less attractive from the perspective of the low-skilled. This may give them an incentive to understate their public-goods preferences collectively. (A symmetric example can be constructed in which high-skilled individuals exaggerate their public-goods preferences.)

To articulate this concern more formally, we find it useful to define the indirect utility function  $V^* : S \times W \times \Theta \rightarrow \mathbb{R}$ , with

$$V^*(s, w, \theta) = \theta q^*(s) + u(c^*(s, w, \theta)) - \frac{y^*(s, w, \theta)}{w},$$

where  $(q^*, c^*, y^*)$  is the social choice function characterized in Proposition 2. One easily derives that, for every  $s, w, \theta$ , and  $k \in \{L, H\}$ ,

$$\frac{\partial V^*(s, w, \theta)}{\partial p_k} = (\theta w - r'(q^*(s))) \frac{1}{w} \frac{\partial q^*(s)}{\partial p_k} = \left( \theta w - \frac{\bar{\theta}(s)}{\lambda(s)} \right) \frac{1}{w} \frac{\partial q^*(s)}{\partial p_k}. \quad (9)$$

Now consider a low-skilled individual with a high taste for the public good; i.e.,  $w = w_L$  and  $\theta = \theta_H$ . Also, for the sake of the argument, suppose that the type space under consideration is such that this individual's beliefs assign a lot of probability mass to states  $s$  such that both  $p_H$  and  $p_L$  are high; i.e., the individual believes that most other individuals have a high taste parameter. This implies that  $\bar{\theta}(s)$  is close to  $\theta_H$  so that  $\frac{\partial V^*(s, w_L, \theta_H)}{\partial p_H}$  is close to

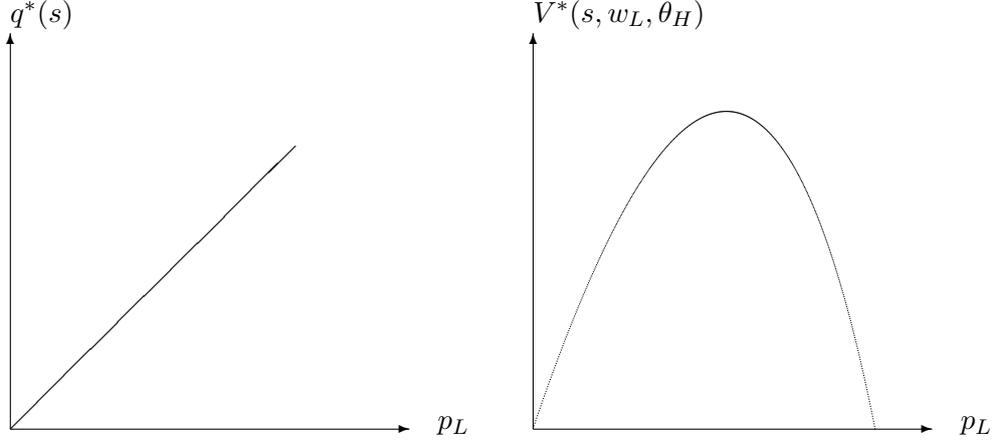
$$\theta_H \left( w_L - \frac{1}{\lambda(s)} \right) \frac{1}{w_L} \frac{\partial q^*(s)}{\partial p_L}.$$

Since, for all  $s$ ,  $w_L < \frac{1}{\lambda(s)}$ , and  $\frac{\partial q^*(s)}{\partial p_L} > 0$ , this implies that

$$\frac{\partial V^*(s, w_L, \theta_H)}{\partial p_L} < 0.$$

This situation is illustrated in Figure 1. Assuming a quadratic cost function, the provision level  $q^*(s) = q^*(f_H, p_H, p_L)$  is, given  $f_H$  and  $p_H$ , a linearly increasing function of the fraction of low-skilled individuals with a high taste parameter,  $p_L$ . The indirect utility function of these individuals  $V^*(s, w_L, \theta_H) = V^*(f_H, p_H, p_L, w_L, \theta_H)$  is, however, increasing in  $p_L$  only if  $p_L$  is low and is decreasing if  $p_L$  is high. Hence, if these individuals think that it is likely that they will find themselves on the downward-sloping part of their indirect utility function they would be happy if they could make the mechanism designer believe that  $p_L$  was lower than it actually is. Moreover, if many of them falsely communicated a low taste parameter, the mechanism designer would perceive  $p_L$  to be lower than it actually is and reduce the public-goods provision level. Finally, note that such lies are perfectly in line with the incentives at the individual level: It is an implication of individual incentive-compatibility (recall equation (4)) that neither an individual's consumption level  $c$  nor his productive effort  $y$  depend on the preference parameter. Consequently, lying is a best response.

Figure 1: State-dependent public-goods provision

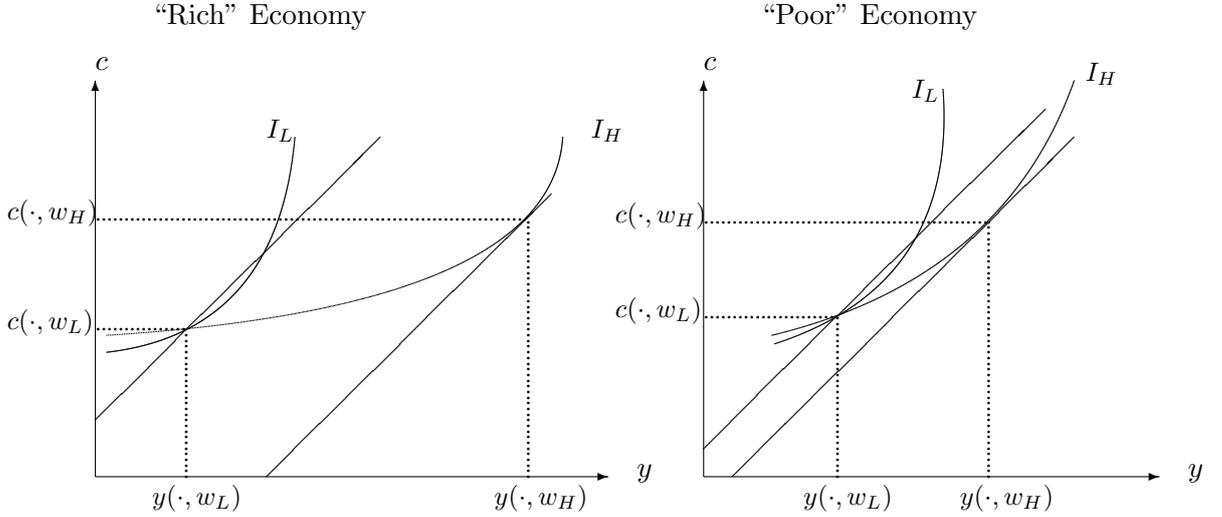


**Example 2: Productive Abilities.** We can also question whether information on the fraction of high-skilled individuals,  $f_H$ , can be acquired if the social choice function in Proposition 2 is used. To demonstrate this in an easy way, consider a simplified version of our model without public goods. Suppose that we seek to implement a social choice function with the following properties: for all states, there is a binding incentive compatibility constraint so that high-skilled individuals are indifferent between the bundles  $z(s, w_L) := (c(s, w_L), y(s, w_L))$  and  $z(s, w_H) = (c(s, w_H), y(s, w_H))$ , and there is redistribution from the high-skilled to the low-skilled,  $y(s, w_H) - c(s, w_H) > 0$  and  $y(s, w_L) - c(s, w_L) < 0$ . Moreover, suppose that the level of redistribution varies across states; it is large whenever the economy is “rich”, in the sense that most workers are high-skilled ( $f_H > \frac{1}{2}$ ), and it is small otherwise. This is illustrated by Figure 2. In this Figure,  $I_L$  is the relevant indifference curve of the low-skilled, and  $I_H$  is the one of the high-skilled individuals.

Does it make sense to assume that high-skilled individuals are communicating their skill level truthfully to the mechanism designer? The “rich states” involve more redistribution than the “poor states” so that the high-skilled individuals are better off in the latter. Moreover, for every  $s$ , the incentive constraint of the high-skilled is binding, so that the high-skilled are giving a best response if they lie about their skill level. These individuals could therefore be inclined to lie about their skill level so as to convince the mechanism designer that there are only few high-skilled individuals in the population and that it is therefore optimal to have only a moderate level of redistribution.

The implementability of the social choice function in Proposition 2 is based on the assumption that individuals do not lie about their characteristics, because, in a large economy, they cannot affect the outcome anyway. We consider this way of breaking the individual’s indifference in favor of truth-telling to be unconvincing. If all like-minded individuals – e.g., all individuals with a low skill level and a high preference parameter in Example 1, or all high-skilled individu-

Figure 2: State-dependent redistribution



als in the example in Example 2 – coordinated their behavior, they could affect the outcome in a way that makes all of them strictly better off, without violating the postulate that each individual’s action is a best response to the actions chosen by all other individuals. To articulate this concern more formally, we will introduce a notion of coalition-proofness in the following section.

## 4 Robust and coalition-proof social choice functions

In this section, we develop the notion of a robust and coalition-proof social choice function and state necessary and sufficient conditions that characterize such a social choice function. A main result of this section will be that preference and productivity shocks have very different implications: the possibility of preference shocks indeed gives rise to an additional set of collective incentive constraints that a social choice function has to fulfill. By contrast, productivity shocks do not give rise to such constraints. Hence, a mechanism designer has to provide appropriate incentives in order to learn  $p_H$  and  $p_L$ , while he gets the information on  $f_H$  for free.

As a first step, however, we define formally what it means that the game induced by a mechanism  $M = [(A, \mathcal{A}), Q, C, Y]$  has a coalition-proof equilibrium.<sup>11</sup> We will then introduce the requirement of robustness, and provide a characterization of robust and coalition-proof social choice functions.

<sup>11</sup>The definition below is a simplified version of the notion of a coalition-proof Nash equilibrium due to Bernheim et al. (1986). In particular, we also take a non-cooperative approach to coalition formation, and we also require that a coalition is subcoalition-proof, i.e., that the formation of a coalition cannot be undermined by the further deviation of a subcoalition. For reasons of tractability, however, we do not model a possibly infinite chain of successive formations of subcoalitions.

## 4.1 Coalition-proof Bayes-Nash equilibrium

A (mixed) strategy in the game induced by  $M$  is a function  $\sigma : T \rightarrow \mathcal{M}(A)$  that specifies a probability distribution over actions for each type of individual. Put differently, the action chosen by individual  $i$  is a random variable  $a(t^i)$ . The probability, conditional on the event  $t^i = t$ , that  $a(t^i)$  takes values in subset  $A'$  of  $A$  is in the following denoted by  $\sigma(A' | t)$ .

We find it convenient to introduce the following notation: suppose that, for the game induced by mechanism  $M$ , individuals follow a strategy  $\sigma : T \rightarrow \mathcal{M}(A)$ , then the expected payoff of a type  $t$  individual from behaving according to  $\chi \in \mathcal{M}(A)$  is given by

$$\tilde{U}_M(\sigma, \chi, t) := \int_{\mathcal{M}(T)} \int_A \tilde{u}_M(\alpha(\delta, \sigma), a, w(t), \theta(t)) d\chi(a) d\beta(\delta | t),$$

where, for any  $\alpha, a, \theta$  and  $w$  we define

$$\tilde{u}_M(\alpha, a, w, \theta) := \theta Q(\alpha) + u(C(\alpha, a)) - \frac{Y(\alpha, a)}{w},$$

and  $\alpha(\delta, \sigma) = \delta \circ \sigma^{-1}$  is the cross-section distribution of actions induced by strategy  $\sigma$  if the cross-section distribution of types is given by  $\delta$ . We assume that a law of large numbers for large economies holds so that we can interpret  $\sigma(A' | t)$  both as the probability that the action chosen by a type  $t$  individual belongs to a subset  $A'$  of  $A$  and as the fraction of type  $t$  individuals who choose an action in  $A'$ .<sup>12</sup> Consequently, for a given  $\delta$ , we can treat  $\alpha(\delta, \sigma)$  as a non-random quantity.

**Definition 1** *Given a mechanism  $M$  and a type space  $[(T, \mathcal{T}), \pi, \beta]$ , a strategy  $\sigma^* : T \rightarrow \mathcal{M}(A)$  is said to be a coalition-proof Bayes-Nash equilibrium if it is a Bayes-Nash equilibrium, and there is no set of types  $T' \subseteq T$  who can deviate to a strategy  $\sigma'_{T'} : T' \rightarrow \mathcal{M}(A)$  so that the following conditions are fulfilled:*

- (a) *The strategy profile  $(\sigma^*_{T \setminus T'}, \sigma'_{T'})$ , where  $\sigma^*_{T \setminus T'}$  is the restriction of  $\sigma^*$  to types not in  $T'$ , is a Bayes-Nash equilibrium.*
- (b) *Deviators are made better off: the outcome that is induced if all types in  $T \setminus T'$  play according to  $\sigma^*_{T \setminus T'}$ , and all types in  $T'$  play according to  $\sigma'_{T'}$ , is preferred by all individuals with types in  $T'$ ; i.e., for all  $t \in T'$ ,*

$$\tilde{U}_M((\sigma^*_{T \setminus T'}, \sigma'_{T'}), \sigma'_{T'}(t), t) > \tilde{U}_M(\sigma^*, \sigma^*(t), t). \quad (10)$$

- (c) *The deviation is subcoalition-proof: there is no strict subset  $T''$  of  $T'$  – i.e., a subset  $T''$  of  $T'$  so that there are  $t' \in T'$  and  $t'' \in T''$  with  $w(t') \neq w(t'')$ , or  $\theta(t') \neq \theta(t'')$ , or  $\beta(t') \neq \beta(t'')$  – with a strategy  $\sigma''_{T''} : T'' \rightarrow \mathcal{M}(A)$  so that  $(\sigma^*_{T \setminus T'}, \sigma'_{T' \setminus T''}, \sigma''_{T''})$  is a Bayes-Nash equilibrium, and, for all  $t \in T''$ ,*

$$\tilde{U}_M((\sigma^*_{T \setminus T'}, \sigma'_{T' \setminus T''}, \sigma''_{T''}), \sigma''_{T''}(t), t) \geq \tilde{U}_M((\sigma^*_{T \setminus T'}, \sigma'_{T'}), \sigma'_{T'}(t), t). \quad (11)$$

<sup>12</sup>For a discussion of the law of large numbers in large economies, see Sun (2006), Al-Najjar (2004) or Judd (1985).

An equilibrium  $\sigma^*$  is coalition-proof only if it does not leave incentives for a subset of individuals to coordinate their behavior in such a way that they induce an outcome that makes all of them better off. Our definition is very demanding with respect to the consistency requirements that such a deviation from an equilibrium strategy  $\sigma^*$  has to satisfy: the behavior that is prescribed by the deviation must induce a new Bayes-Nash equilibrium, i.e., playing according to  $(\sigma_{T \setminus T'}^*, \sigma'_{T'})$  must be a best response, both for the deviating types as well as for the non-deviating types. Also, the outcome that is induced by the deviation must be beneficial for all deviating types. Finally, we require that a deviation must itself be coalition-proof; that is, it must not trigger a further deviation by a subcoalition of the deviators.

With this definition, we may think of the collective deviation as resulting from an own mechanism design problem that the deviating agents face. Condition (a) can be interpreted as a incentive compatibility constraint, so that behaving according the strategy profile  $(\sigma_{T \setminus T'}^*, \sigma'_{T'})$  is indeed a best response. Condition (b) is a participation constraint which ensures that the deviators are made better off. Finally, condition (c) requires that the mechanism on which the collective deviation is based, must also be coalition-proof. A similar approach to coalition formation has previously been introduced by Laffont and Martimort (1997, 2000), and has been extended to a large economy model by Bierbrauer and Hellwig (2010). These papers explicitly model the formation of a coalition as an extensive form game, so that first an overall mechanism is announced, then a coalition organizer may propose a collusive side mechanism to a set of deviating agents, and ultimately a subcoalition organizer may propose a further side mechanism to a subset of the deviators. A mechanism is then said to be coalition-proof if it does not provoke the formation of a collusive side mechanism.

The approach taken here is different in that we define the notion of a coalition-proof equilibrium with reference to a given normal form game. The reason for this approach is that it makes the exposition easier, without affecting the conclusions. Indeed, the constraints on social choice functions that are derived below resemble those identified by Bierbrauer and Hellwig (2010), albeit in a somewhat different model.

## 4.2 Robust and coalition-proof implementation

For a given type space, a social choice function  $(q, c, y)$  is said to be implementable as a coalition-proof Bayes-Nash equilibrium, if there is a mechanism  $M$  and a strategy  $\sigma^*$  such that (i)  $\sigma^*$  is a coalition-proof Bayes-Nash equilibrium, and (ii) the equilibrium allocation coincides with the prescription of the social choice function for every  $\delta$ ; i.e., we have that, for every  $\delta$ ,

$$Q(\alpha(\delta, \sigma^*)) = q(s(\delta)) \tag{12}$$

and, for each  $\delta$  and  $t$ ,

$$C(\alpha(\delta, \sigma^*), a(t)) = c(s(\delta), w(t), \theta(t)) \text{ and } Y(\alpha(\delta, \sigma^*), a(t)) = y(s(\delta), w(t), \theta(t)) , \tag{13}$$

$\sigma^*(t)$ -almost surely.

We say that a social choice function is robustly implementable and coalition-proof, if, given  $(T, \mathcal{T})$  and  $\pi$ , there is a mechanism  $M$  and a strategy  $\sigma^*$  such that requirements (i) and (ii) are

fulfilled, for every belief system  $\beta$ .

Subsequently, we derive necessary and sufficient conditions for robustness and coalition-proofness. However, before going into details, we use Example 2 to illustrate the interplay of robustness and coalition-proofness. The example is meant to illustrate the following: if our analysis was based on specific assumptions about beliefs, the mechanism designer could typically exploit her knowledge of the belief system so that the requirement of coalition-proofness would not have a lot of bite. If instead we insist on the use of robust mechanisms, coalition-proofness becomes a substantive constraint. The example also illustrates why, with the solution concept of a coalition-proof Bayes-Nash equilibrium, the revelation principle does not hold. It shows that non-direct mechanisms may eliminate failures of coalition-proofness that cannot be avoided if direct mechanisms are used.<sup>13</sup>

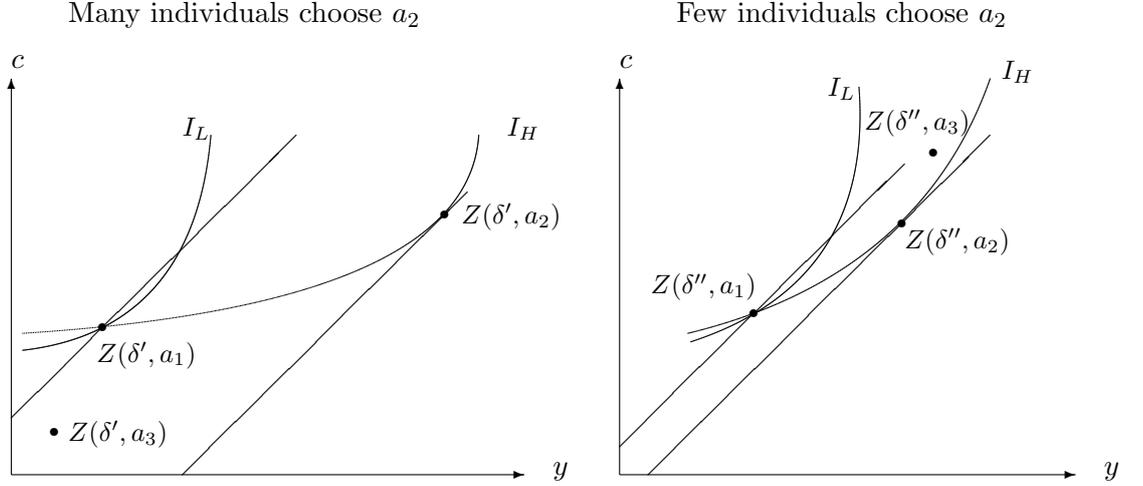
**Example 2 revisited.** We argued before that the social choice function in Figure 2 cannot be implemented in a coalition-proof way, under the assumption that attention is restricted to direct mechanisms and to truth-telling equilibria: a collective lie so that some high-skilled individuals declare a low-skill level, induces a new equilibrium in which the high-skilled individuals are better off. In the following, we show that a mechanism designer who knows the belief system  $\beta$  can generally eliminate the scope for such a collective deviation by using a non-direct mechanism. We will also argue that the scope of such non-direct mechanisms is reduced if we insist on robustness.

For the sake of concreteness, suppose first that all high-skilled individuals believe that rich and poor states are equally likely, and, in addition, that this is known by the mechanism designer. He may then use a non-direct mechanism with an action set  $A = \{a_1, a_2, a_3\}$  and an outcome function  $Z := (C, Y)$  which works as follows: in any state, an individual who chooses action  $a_1$  receives the consumption-income bundle dedicated to the low-skilled individuals; i.e., taking the action  $a_1$  is interpreted as saying “I am a low-skilled individual”. Likewise, an individual who chooses action  $a_2$  receives the bundle intended for the high-skilled individuals. The action  $a_3$  gives a very unattractive bundle if many individuals choose action  $a_2$ . This is illustrated in the left part of Figure 2, where  $\delta'$  indicates a distribution of actions so that more than half of the population chooses  $a_2$ . However, if only few individuals choose action  $a_2$ , then action  $a_3$  is very attractive for the high-skilled individuals. This is illustrated by the right part of Figure 2, where  $\delta''$  indicates a distribution of actions so that less than half of the population chooses  $a_2$ . For the given specification of beliefs, the non-direct mechanism implements the social choice function in a coalition-proof way, because there is no longer an equilibrium in which high-skilled individuals communicate having a low-skill level. Conditional on many high-skilled individuals choosing action  $a_1$ , it is a best response for a high-skilled individual to choose action  $a_3$ . Moreover, the equilibrium in which all high-skilled individuals choose action  $a_2$  and thereby communicate their skill level truthfully to the mechanism designer remains intact: if the situations in Figure 3 arise

---

<sup>13</sup>It is well-known that the objective to implement a social choice function as the *unique* equilibrium of some mechanism, generally makes the use of non-direct mechanisms necessary (see, e.g., Jackson (2001) for an overview). Here, the situation is similar in that our notion of coalition-proofness postulates that there must not exist a second equilibrium with certain properties.

Figure 3: The non-direct mechanism



with equal probability, then choosing  $a_2$  is a best response for the high-skilled because action  $a_3$  yields a very bad outcome with probability  $\frac{1}{2}$ .

However, the non-direct mechanism does no longer work if we insist on robustness. To see this, suppose that all individuals believe that poor states occur with probability 1. Then, all individuals believe that the payoffs associated with actions  $a_1$ ,  $a_2$ , and  $a_3$  are as in the right part of Figure 2. Consequently, all individuals with a high-skill level will choose action  $a_3$  instead of action  $a_2$ , so that the social choice function is no longer reached.

#### 4.2.1 Necessary conditions for robustness and coalition-proofness

For a given social choice function  $(q, c, y)$ , define the associated indirect utility function  $V$  by

$$V(s, w, \theta) = \theta q(s) + u(c(s, w)) - \frac{y(s, w)}{w}.$$

**Proposition 3** *If  $(q, c, y)$  is robust and coalition-proof, then it must be true that: (i) For any given pair  $(f_H, p_L)$ ,  $V(s, w_H, \theta_L)$  is a non-increasing function of  $p_H$ , and  $V(s, w_H, \theta_H)$  is a non-decreasing function of  $p_H$ , and (ii) for any given pair  $(f_H, p_H)$ ,  $V(s, w_L, \theta_L)$  is a non-increasing function of  $p_L$ , and  $V(s, w_L, \theta_H)$  is a non-decreasing function of  $p_L$ .*

The logic of the proof is straightforward.<sup>14</sup> If, say, the constraint that  $V(s, w_L, \theta_H)$  is a non-decreasing function of  $p_L$  is violated, this implies that there exist  $p_L$  and  $p'_L$  with  $p'_L > p_L$  so that  $V((f_H, p_L, p_H), w_L, \theta_H) > V((f_H, p'_L, p_H), w_L, \theta_H)$ . If we now consider a type space, so that

<sup>14</sup>A difficulty for the proof of Proposition 3 is that we cannot rely on the revelation principle. However, given some equilibrium  $\sigma^*$ , a false communication of, say, the taste parameter by a type  $t$  individual can still be defined in a meaningful way: it takes the form of behaving according to  $\sigma^*(\hat{t})$ , for some type  $\hat{t} \neq t$  with  $\theta(\hat{t}) \neq \theta(t)$ .

all individuals assign mass 1 to a distribution of types  $\delta$  with  $s(\delta) = (f_H, p'_L, p_H)$ , individuals with a low skill level and a high taste parameter have an incentive to lie. If they communicate a low as opposed to a high taste parameter to the mechanism designer – more specifically, if they, falsely, announce a low taste parameter with probability  $1 - \frac{p_L}{p'_L}$ , and, truthfully, announce a high taste parameter with probability  $\frac{p_L}{p'_L}$  – they will receive the outcome intended for the case that  $s = (f_H, p_L, p_H)$ , and are thereby made better off, i.e., requirement (b) in Definition 1 is fulfilled. Since the lie involves only a false communication of taste parameters and, by individual incentive compatibility, an individual's  $(c, y)$ -bundle does not depend on the taste parameter, every individual is giving a best response. Hence, the deviation satisfies property (a). Finally, all these individuals have the same preferences, and the same beliefs so that there exists no strict subset of types. This implies that the deviation is subcoalition-proof, i.e., property (c) is also satisfied.

To see the significance of the constraints in Proposition 3, it is instructive to check which of these constraints are satisfied and which ones are violated by the Mirrleesian social choice function in Proposition 2. The constraint that low-skilled individuals with a high taste parameter do not benefit from understating their public-goods preferences is satisfied provided that  $\frac{\partial V^*(s, w_L, \theta_H)}{\partial p_L} \geq 0$ , where  $V^*$  is the indirect utility function that was defined in Section 3. By equation (9), this condition is violated whenever

$$\theta_H w_L < \frac{\bar{\theta}(s)}{\lambda(s)}. \quad (14)$$

The constraint that high-skilled individuals with a low taste parameter do not benefit from exaggerating their public-goods preferences is satisfied provided that  $\frac{\partial V^*(s, w_H, \theta_L)}{\partial p_H} \leq 0$ . Again, by equation (9), this condition is violated whenever

$$\theta_L w_H > \frac{\bar{\theta}(s)}{\lambda(s)}. \quad (15)$$

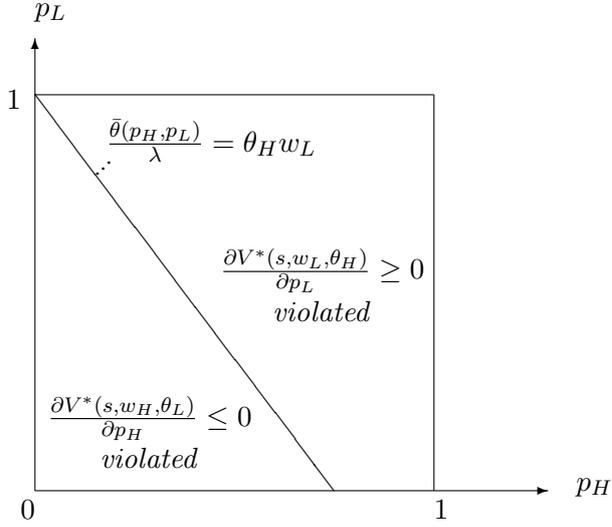
Since  $\theta_L w_H = \theta_H w_L$ , equations (14) and (15) imply that the Mirrleesian social choice function violates coalition-proofness, whenever  $\theta_H w_L \neq \frac{\bar{\theta}(s)}{\lambda(s)}$ . In essence this implies that there is no state  $s$  of the economy so that the Mirrleesian social choice function is coalition-proof. This is illustrated by Figure 4. In this Figure, we treat  $f_H$  as a given parameter, and vary only  $p_H$  and  $p_L$ . The line in the figure is the locus along which  $\theta_L w_H = \frac{\bar{\theta}(s)}{\lambda(s)}$ , or, equivalently, along which  $p_L = 1 - \frac{f_H}{1-f_H} p_H$ . (The graph is drawn under the assumption that  $f_H > \frac{1}{2}$ .)

#### 4.2.2 Sufficient conditions for robustness and coalition-proofness

The following Proposition states a sufficient condition for coalition-proofness. More specifically, it states that all social choice functions in a set  $\Omega(\epsilon)$  are robust and coalition-proof. This set is defined as the set of social choice functions with the following properties: (i) For every  $s$ , the necessary conditions in Proposition 3 are satisfied, and at most one of these conditions is binding, (ii) for every  $s$ , the resource constraint in (1) holds, and (iii) for some given  $\epsilon > 0$ ,

$$u(c(s, w)) - \frac{y(s, w)}{w} \geq u(c(s, \hat{w})) - \frac{y(s, \hat{w})}{w} + \epsilon, \quad (16)$$

Figure 4: Violation of coalition-proofness



for every  $s$ ,  $w$ , and  $\hat{w}$ . These constraints require that, for every  $s$ , an individual with skill level  $w$  prefers the “own” consumption-output bundle  $(c(s, w), y(s, w))$  strictly over any alternative bundle  $(c(s, \hat{w}), y(s, \hat{w}))$ , where the parameter  $\epsilon$  is the minimal intensity of this strict preference; that is, the constraints in (16) require that there is a little bit of slack in the incentive compatibility constraints in (5).

**Proposition 4** *A social choice function in  $\Omega(\epsilon)$  is robustly implementable as a coalition-proof Bayes-Nash equilibrium.*

The proof is based on a direct mechanism that reaches the given social choice function in a truth-telling equilibrium.<sup>15</sup> We verify that this equilibrium is coalition-proof, whatever the belief system  $\beta$  is. As a first step, we observe that there is no collective deviation that involves a false communication of productive abilities. For social choice functions in  $\Omega(\epsilon)$ , all individual incentive compatibility constraints hold as strict inequalities which implies that there is no equilibrium in which individuals declare false productivity levels. Hence, any such deviation would violate condition (a) in Definition 1.

Next, consider a deviation that involves only lies about taste parameters. Any such deviation induces a new equilibrium because  $(c, y)$ -bundles do not depend on taste parameters. Suppose first that the types who participate all have the same payoff type, i.e., that they all have the same skill level and the same taste parameter. For instance, suppose that they all have the skill level  $w_H$  and the taste parameter  $\theta_L$ . If these individuals lie about their taste parameter, this implies that the mechanism designer ends up with the perception that  $p_H$  is higher than

<sup>15</sup>Hence, our proof implies, in particular, that if the sufficient conditions in Proposition 4 are satisfied, then we may without loss of generality focus on direct mechanisms and truth-telling equilibria.

it actually is. By the constraints in Proposition 3,  $V(s, w_H, \theta_L)$  is a non-increasing function of  $p_H$ , so that this deviation does not make the participating individuals better off, i.e., it violates condition (b) in Definition 1.

Now suppose that the types who collectively lie about their taste parameters have diverse payoff types. The assumption that, for every  $s$ , at most one of the necessary conditions in Proposition 3 binds, implies that there is always a set of types who would like to “withdraw” their contribution to the deviation, thereby free-riding on the contribution of others. For instance, suppose that individuals with payoff type  $(w_H, \theta_L)$  and individuals with payoff type  $(w_H, \theta_H)$  lie about their taste parameters. From an ex post perspective, either the  $(w_H, \theta_L)$ -individuals think that  $p_H$ , as perceived by the mechanism designer, is too high, or the  $(w_H, \theta_H)$ -individuals think that  $p_H$  is too low. Consequently, ex interim, individuals with payoff type  $(w_H, \theta_L)$  understand that, taking the lie of individuals with payoff type  $(w_H, \theta_H)$  as given, they are weakly better off if they communicate their characteristics truthfully. Likewise, the  $(w_H, \theta_H)$ -individuals are weakly better off if they refuse to lie about their public-goods preferences. Moreover, if the deviation affects the implemented policy with positive probability (which is necessary in order to satisfy condition (b) in Definition 1), then one of these groups is in fact strictly better off if it communicates truthfully, which implies that the deviation does not satisfy condition (c) in Definition 1.

#### 4.2.3 Why are these conditions useful for finding an optimal social choice function?

Propositions 3 and 4 make it possible to solve for the optimal social choice function via the following procedure: First, characterize the optimal social choice function among those that are individually incentive-compatible, resource-feasible and satisfy the necessary conditions in Proposition 3. Second, verify that the optimal social choice function is indeed such that, for every  $s$ , at most one of the monotonicity constraints holds as an equality, more formally that it belongs to the set  $\Omega(0)$ . This procedure will be applied in the following Section.

The social choice functions in  $\Omega(0)$  are not generally coalition-proof, as we explain below. However, under a mild technical assumption, every such social choice function can be approximated by a social choice function that is coalition-proof. This is stated more formally in the following Corollary.

**Corollary 1** *Suppose that there is some  $\bar{\epsilon} > 0$  so that the set  $\bigcup_{0 \leq \epsilon \leq \bar{\epsilon}} \Omega(\epsilon)$  is compact. Then, for every social choice function  $(q, c, y) \in \Omega(0)$ , and for every  $\tilde{\epsilon} > 0$ , there is a social choice function  $(q', c', y')$  that is robustly implementable as a coalition-proof Bayes-Nash equilibrium, and satisfies*

$$|U(q'(s), c'(s, w), y'(s, w), w, \theta) - U(q(s), c(s, w), y(s, w), w, \theta)| \leq \tilde{\epsilon},$$

for every  $s$ , and every  $(\theta, w)$ .

**Example 2 revisited.** To illustrate why the social choice functions in  $\Omega(0)$  are not necessarily coalition-proof themselves but can be approximated by coalition-proof social choice functions,

it is instructive to look once more at the example in Section 3, where the social choice function that is illustrated in Figure 2 cannot be implemented as a robust and coalition-proof equilibrium because the high-skilled have an incentive to lie. We will now argue that there is, however, a social choice function which is arbitrarily close and does not face this problem.

Suppose that the social choice function in Figure 2 is modified as follows: in both graphs, the bundle for high-skilled individuals is moved to a slightly higher indifference curve.<sup>16</sup> This implies that truth-telling is the unique best response of the high-skilled, for every state  $s$ . A deviation that involves lies about skill levels is therefore no longer consistent with equilibrium behavior. This illustrates the general insight in Corollary 1. Once we introduce a tiny amount of slack into the incentive compatibility constraints, deviations that involve lies about skill levels are no longer viable. The example also shows why the slack is needed. If incentive compatibility constraints are binding, lies that involve skill levels are a concern.

### 4.3 On the separability of individual and collective incentive problems

The reasoning in section 4.2 translates the requirement of coalition-proofness into a simple set of inequality constraints: there must not exist a group of individuals who could benefit from the policy change that is induced by a false communication of public-goods preferences. This simple characterization is available because as far as coalition-proofness is concerned, we may, without loss of generality, assume that productive abilities are communicated truthfully: if we introduce a tiny amount of slack into individual incentive compatibility constraints, any lie that involves a false communication of productive abilities is effectively deterred.

A first major insight of the paper is therefore that preference and productivity shocks have very different implications for the set of robust and coalition-proof social choice functions: while appropriately calibrated incentives at the individual level make a manipulative communication of productive abilities unviable, the communication of public-goods preferences cannot be addressed in this way. As we have seen in Section 3, individual incentive compatibility implies that individuals who differ only in their public-goods preferences need to be treated equally in terms of their consumption level  $c$  and their output requirement  $y$ . Consequently, individuals are willing to lie about their public-goods preferences if this has positive consequences at an aggregate level. A social choice function therefore has to be such that those lies are unattractive.

## 5 The optimal robust and coalition-proof social choice function

In this section, we characterize the social choice function which maximizes expected utilitarian welfare  $E[W(s)]$  subject to the requirements of individual incentive compatibility, resource feasibility and coalition-proofness.

We assume that the mechanism designer has subjective beliefs about the possible realizations of  $s$ . For simplicity, we assume that she has an agnostic prior in the following sense: she views  $f_H$ ,  $p_H$  and  $p_L$  as independent random variables, which are uniformly distributed over the unit

---

<sup>16</sup>To preserve feasibility, we may simultaneously have to move the low-skilled individuals to a slightly lower indifference curve.

interval.<sup>17</sup>

We proceed in two steps. We first characterize the optimal social choice function that satisfies the necessary conditions identified in Proposition 3; i.e., we study the following optimization problem: choose  $q : s \mapsto q(s)$ ,  $y : (s, w) \mapsto y(s, w)$  and  $c : (s, w) \mapsto c(s, w)$  in order to maximize  $E[W(s)]$  subject to the requirements that, for any  $k \in \{L, H\}$ ,  $V(s, w_k, \theta_L)$  is a non-increasing function of  $p_k$ , and  $V(s, w_k, \theta_H)$  is a non-decreasing function of  $p_k$ , and that, for every  $s$ , the resource constraints in (6), and the individual incentive compatibility constraints in (5) are satisfied. Subsequently, we show that the resulting social choice function is such that, for every  $s$ , at most one of these necessary conditions is binding, i.e., that the social choice function in question belongs to the set  $\Omega(0)$ . With an appeal to Corollary 1, this implies that the social choice function is approximately coalition-proof.

Our strategy for solving the optimization problem is to decompose it into a number of subproblems, each of which take only some of necessary conditions in Proposition 3 into account. (We will verify ex post that the neglected constraints are indeed satisfied.) More specifically, for each possible value of the parameter  $f_H \in [0, 1]$  we study the following set of subproblems:

- i) **Problems of the  $\mathcal{P}_L(p_H, f_H)$ -type:** Fix  $p_H$  and  $f_H$  and consider the values of  $p_L$  so that  $\frac{\bar{\theta}(p_H, p_L)}{\lambda} \geq \theta_H w_L$ . Choose  $c(s, w_L)$ ,  $y(s, w_L)$ ,  $c(s, w_H)$ ,  $y(s, w_H)$  and  $q(s)$  in order to maximize  $E \left[ W(s) \mid \frac{\bar{\theta}(p_H, p_L)}{\lambda} \geq \theta_H w_L, f_H, p_H \right]$  subject to the incentive constraints in (5), the resource constraints in (6), and the constraint  $\frac{\partial V(s, w_L, \theta_H)}{\partial p_L} \geq 0$ .
- ii) **Problems of the  $\mathcal{P}_H(p_L, f_H)$ -type:** Fix  $p_L$  and  $f_H$  and consider the values of  $p_H$  so that  $\frac{\bar{\theta}(p_H, p_L)}{\lambda} \leq \theta_H w_L$ . Choose  $c(s, w_L)$ ,  $y(s, w_L)$ ,  $c(s, w_H)$ ,  $y(s, w_H)$  and  $q(s)$  in order to maximize  $E \left[ W(s) \mid \frac{\bar{\theta}(p_H, p_L)}{\lambda} \leq \theta_H w_L, f_H, p_L \right]$  subject to the incentive constraints in (5), the resource constraints in (6), and the constraint  $\frac{\partial V(s, w_H, \theta_L)}{\partial p_H} \leq 0$ .

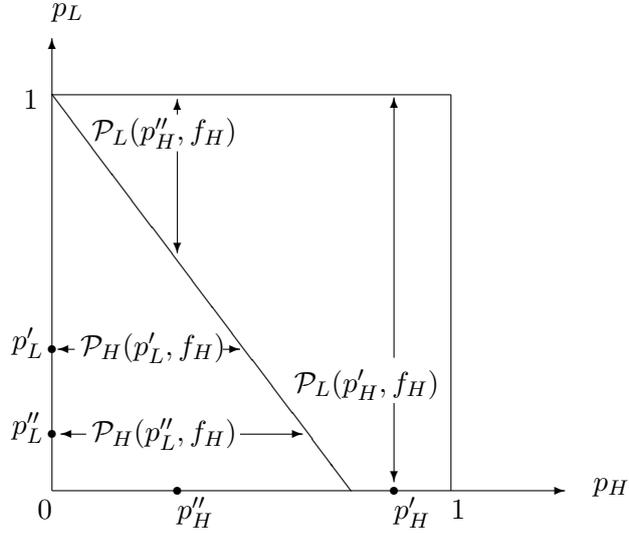
More intuitively, with a  $\mathcal{P}_L(p_H, f_H)$ -problem, one has to deviate from the Mirrleesian outcome in Proposition 2 because the low-skilled individuals need to be prevented from understating their public-goods preferences. With a problem of the  $\mathcal{P}_H(p_L, f_H)$ -type one has to deviate because the high-skilled need to be prevented from exaggerating theirs. Figure 5 illustrates how these problems relate to each other. Along the downward sloping line we have, as in Figure 4, that  $\frac{\bar{\theta}(p_H, p_L)}{\lambda} = \theta_H w_L$ .

Observe that, if we determine for every  $f_H$ , the solutions to all problems of the  $\mathcal{P}_L(p_H, f_H)$ -type and to all problems of the  $\mathcal{P}_H(p_L, f_H)$ -type we characterize a social choice function.<sup>18</sup> For every  $s$ ,  $q(s)$ ,  $c(s, w_L)$ ,  $y(s, w_L)$ ,  $c(s, w_H)$ , and  $y(s, w_H)$  are then given by the solution to the relevant subproblem, in the following denoted by  $c^{**}(s, w_L)$ ,  $y^{**}(s, w_L)$ ,  $c^{**}(s, w_H)$ ,  $y^{**}(s, w_H)$  and  $q^{**}(s)$ . For later use, we denote by  $v_k^{**}(s) = u(c^{**}(s, w_L)) - \frac{y^{**}(s, w_k)}{w_k}$ , the utility derived by individuals with skill level  $w_k$  from their  $(c, y)$  bundle in state  $s$ . Analogously, we denote by

<sup>17</sup>These beliefs affect the way in which the mechanism designer is making trade-offs between welfare levels in different states of the economy. The assumption of an agnostic prior simplifies the exposition. However, the logic of the analysis would remain the same with alternative assumptions about the mechanism designer's beliefs.

<sup>18</sup>Along the  $\frac{\bar{\theta}(p_H, p_L)}{\lambda} = \theta_H w_L$  line, the outcome is determined by two such problems. However, as follows from Propositions 5 and 6 below, these solutions are identical.

Figure 5: Relaxed optimization problems



$v_k^*(s)$  the value that results from the Mirrleesian social choice function in Proposition 2. Finally, we denote by  $\tau^{**}(s, w_k) = 1 - \frac{1}{w_k u'(c^{**}(s, w_k))}$  the implicit marginal tax rate for individuals with skill level  $w_k$ .

Propositions 5 and 6 below respectively characterize the solutions of a problem of the  $\mathcal{P}_L(p_H, f_H)$ -type, and a problem of the  $\mathcal{P}_H(p_L, f_H)$ . For the sake of concreteness, both consider the case  $f_H < \frac{1}{2}$ .<sup>19</sup>

### 5.1 Preventing the low-skilled from understating their preferences

**Proposition 5** *Let  $f_H < \frac{1}{2}$ . The solution to Problem  $\mathcal{P}_L(p_H, f_H)$  has the following properties. There is a cutoff value  $\hat{p}_L$  such that:*

- (a) *For  $p_L \in [\max\{0, 1 - \frac{f_H}{1-f_H}p_H\}, \hat{p}_L]$  redistribution and public-goods provision are distorted downwards,  $v_L^{**}(s) < v_L^*(s)$ ,  $v_H^{**}(s) > v_H^*(s)$ , and  $q^{**}(s) < q^*(s)$ . Also, the implicit marginal tax rates are lower,  $\tau^{**}(s, w_L) < \tau^*(s, w_L)$ , and  $\tau^{**}(s, w_H) \leq \tau^*(s, w_H)$ .*
- (b) *For  $p_L = \hat{p}_L$  the allocation is undistorted. Also, if  $1 - \frac{f_H}{1-f_H}p_H \geq 0$ , then the allocation is undistorted for  $p_L = 1 - \frac{f_H}{1-f_H}p_H$ .*
- (c) *For  $p_L \in (\hat{p}_L, \bar{p}_L]$  redistribution and public-goods provision are distorted upwards,  $v_L^{**}(s) > v_L^*(s)$ ,  $v_H^{**}(s) < v_H^*(s)$ , and  $q^{**}(s) > q^*(s)$ . Also, the implicit marginal tax rates are higher,  $\tau^{**}(s, w_L) > \tau^*(s, w_L)$  and  $\tau^{**}(s, w_H) = \tau^*(s, w_H)$ .*

The Proposition says that, in comparison to the Mirrleesian social choice function, the public-goods provision level, the extent of redistribution (measured by the utility levels  $v_L$  and  $v_H$  of

<sup>19</sup>Alternative assumptions about  $f_H$  would only affect how the line along which  $\frac{\bar{\theta}(p_H, \eta(p_H))}{\lambda} = \theta_H w_L$  is located, without affecting the results otherwise.

high-skilled and low-skilled individuals), and marginal income tax rates are distorted in the same direction: if there are only few low-skilled individuals who value the public good highly, then there is underprovision of the public good and deficient redistribution, accompanied by lower marginal tax rates. By contrast, if many low-skilled individuals value the public good highly, then there is overprovision of the public good, excessive redistribution, and marginal income tax rates are higher.

Below, we give a sketch of the formal proof.<sup>20</sup> An intuitive understanding may be obtained as follows: first, to see why public goods provision and redistribution are always distorted in the same direction, note that the constraint  $\frac{\partial V(s, w_L, \theta_H)}{\partial p_L} \geq 0$  is binding throughout. Hence, given  $f_H$  and  $p_H$ , different values of  $p_L$  give rise to the same utility level  $\theta_H q(s) + v_L(s)$  for the low-skilled. Now suppose that, for some  $p_L$ , we had underprovision of the public good, in the sense that  $q(s)$  was lower than stipulated by the Samuelson rule, and excessive redistribution, in the sense that utilitarian welfare would go up if we reduced  $v_L(s)$  and increased  $v_H(s)$ . Then we could increase the public-goods provision level,  $q(s)$ , and simultaneously decrease the low-skilled individuals' utility from their  $(c, y)$ -bundle,  $v_L(s)$ , so that  $\theta_H q(s) + v_L(s)$  would remain constant. Utilitarian welfare, by contrast, would go up. Hence, it cannot be optimal to have underprovision of public goods, and at the same time and excessive redistribution. A symmetric argument rules out overprovision and deficient redistribution.

Second, why do we have downward distortions of public-goods provision and redistribution if  $p_L$  is low, and upward distortions if  $p_L$  is high? The reason is that, the higher  $p_L$ , the more attractive is public-goods provision from a welfare perspective. Upward distortions therefore appear if  $p_L$  is high and downward distortions if  $p_L$  is low. Moreover, the higher the public-goods provision level, the more accompanying redistribution is needed to compensate the low-skilled and thereby prevent them from understating their preferences.

Third, why are downward distortions associated with lower marginal tax rates and upward distortions with higher marginal tax rates? The general intuition is that more redistribution from high-skilled individuals to low-skilled individuals goes together with more distortionary taxation. The validity of this intuition is formally confirmed in Bierbrauer and Boyer (2010) who show that the marginal tax rates of low-skilled and high-skilled individuals, respectively, are non-decreasing functions of  $v_L$ .<sup>21</sup>

**Sketch of the Proof of Proposition 5.** We solve problem  $\mathcal{P}_L(p_H, f_H)$  using a two-step-procedure: first, for given  $p_L$ , we treat the public-goods provision level,  $q(p_L)$ , and the utility that low-skilled individuals realize from their  $(c, y)$ -bundle,  $v_L(p_L)$ , as given. (Note that since we treat  $f_H$  and  $p_H$  as given,  $q$  and  $v_L$  can be written as functions of  $p_L$ , i.e., we may suppress

<sup>20</sup>A rigorous proof is in the Appendix.

<sup>21</sup>More specifically, if, starting from the optimal utilitarian outcome,  $v_L$  is increased further, the marginal tax rate of the high-skilled individuals does not change, i.e., we still have “no distortion at the top”. At the same time, the “downward distortion” of low-skilled labor supply becomes more severe; that is, the marginal income tax rate for the low-skilled go up. Similarly, a small reduction in  $v_L$  will again leave the high-skilled individuals' marginal tax rate unchanged, but reduce the downward distortions for the low-skilled. Larger reductions, by contrast, may drive all marginal tax rates down to zero, or even be associated with upward distortions in the supply of high-skilled labor (negative marginal tax rates).

the dependence on the whole vector  $s = (f_H, p_H, p_L)$ .) A solution to problem  $\mathcal{P}_L(p_H, f_H)$  has to be such that, given these variables, the utility of the high-skilled is chosen optimally subject to the individual incentive compatibility and resource constraints; i.e.,

$$v_H(p_L) = V_H(v_L(p_L), r(q(p_L))),$$

where, for any pair  $(v_L, \rho)$ ,

$$\begin{aligned} V_H(v_L, \rho) &:= \max u(c_H) - \frac{y_H}{w_H} \\ \text{s.t.} \quad &u(c_H) - \frac{y_H}{w_H} \geq u(c_L) - \frac{y_L}{w_H}, u(c_L) - \frac{y_L}{w_L} \geq u(c_H) - \frac{y_H}{w_L}, \\ &f_H(y_H - c_H) + (1 - f_H)(y_L - c_L) = \rho, u(c_L) - \frac{y_L}{w_L} = v_L. \end{aligned}$$

The function  $V_H$  can be interpreted as the Pareto-frontier of a Mirrleesian income tax problem with no public goods, but an exogenous revenue requirement  $\rho$ .<sup>22</sup>

Given that  $v_H(p_L) = V_H(v_L(p_L), r(q(p_L)))$ , we can, in a second step, determine the optimal values of  $q(p_L)$  and  $v_L(p_L)$ . For this purpose, we use optimal control theory to determine the solution to the following optimization problem: choose the functions  $q : p_L \mapsto q(p_L)$  and  $v_L : p_L \mapsto v_L(p_L)$  in order to maximize

$$\int_{\kappa(p_H)}^1 \{\bar{\theta}(p_L)q(p_L) + f_H V_H(v_L(p_L), r(q(p_L))) + (1 - f_H)v_L(p_L)\} dp_L$$

subject to the constraint, that for all  $p_L \in [\kappa(p_H), 1]$ , with  $\kappa(p_H) := \max\{0, 1 - \frac{f_H}{1-f_H}p_H\}$ ,

$$\theta_H q'(p_L) + v'_L(p_L) \geq 0.$$

The proof of Proposition 5 relies mainly on two optimality conditions. The first one is as follows:

$$\frac{1}{\theta_H} (\bar{\theta} + f_H V_{H2} r'(q)) = f_H V_{H1} + 1 - f_H. \quad (17)$$

This equation requires that the marginal social benefit from increased public-goods provision  $\bar{\theta} + f_H V_{H2} r'(q)$  is proportional to the marginal social benefit from increased redistribution  $f_H V_{H1} + 1 - f_H$ . The social choice function in Proposition 2 also satisfies this equation, since it prescribes optimal utilitarian redistribution and optimal utilitarian public-goods provision,

$$f_H V_{H1} + 1 - f_H = 0 \quad \text{and} \quad \bar{\theta} + f_H V_{H2} r'(q) = 0.$$

However, as we have argued before (see Figure 4), it violates the constraint  $\theta_H q'(p_L) + v'_L(p_L) \geq 0$ . Now, Equation (17) implies a *complementarity* between redistribution and public-goods provision: if we have excessive redistribution, so that the utility level of the low-skilled is higher than optimal, which implies  $f_H V_{H1} + 1 - f_H < 0$ , then it has to be the case that public-good provision is also higher than optimal,  $f_H V_{H2} r'(q) > 0$ , and vice versa. Moreover, since the function  $V_H$  is decreasing in  $v_L$ , if the low-skilled are better off than otherwise, this implies in turn that there is no longer room to make the high-skilled as well off, i.e.,  $v_H^{**}$  falls short of  $v_H^*$  whenever  $v_L^{**}$  exceeds  $v_L^*$ . Likewise, insufficient redistribution goes together with insufficient public-goods provision, and with higher utility for the high-skilled.

<sup>22</sup>For a complete characterization of the function  $V_H$ , see Bierbrauer and Boyer (2010).

A further insight of Proposition 5 concerns the optimal allocation of the distortions that are due to the binding monotonicity constraint. The Proposition stipulates that if  $p_L$  is high public-goods provision and redistribution should be distorted upwards, and that they should be distorted downwards otherwise. This follows from an optimality condition which stipulates that the “average distortion” must be zero, i.e., a solution has to be such that

$$\int_{\kappa(p_H)}^1 \{\bar{\theta} + f_H V_{H2} r'(q)\} dp_L = \int_{\kappa(p_H)}^1 \{f_H V_{H1} + 1 - f_H\} dp_L = 0. \quad (18)$$

This condition says that any upward distortion of public-goods provision and redistribution that occurs over some subinterval of  $[\kappa(p_H), 1]$  has to be balanced by a downward distortion over some other subinterval. Given that this “budget condition” holds, it is optimal to have the upward distortions of public-goods supply concentrated in the region where it contributes most to welfare, i.e. where  $\bar{\theta}$  is particularly high. This is the case for high values of  $p_L$ .

## 5.2 Preventing the high-skilled from exaggerating their preferences

The following Proposition gives the characterization of the solution to problem  $\mathcal{P}_H(p_L, f_H)$ . It is the mirror image of Proposition 5, i.e., it follows from exactly the same reasoning, so that we can simply state it without need of further elaboration.

**Proposition 6** *Let  $f_H < \frac{1}{2}$ . The solution to Problem  $\mathcal{P}_H(p_L, f_H)$  has the following properties. There is a cutoff value  $\hat{p}_H$  such that:*

- (a) *For  $p_H < \hat{p}_H$  redistribution and public-goods provision are distorted downwards,  $v_L^{**}(s) < v_L^*(s)$ ,  $v_H^{**}(s) > v_H^*(s)$ , and  $q^{**}(s) < q^*(s)$ . Also, the implicit marginal tax rates are lower,  $\tau^{**}(s, w_L) < \tau^*(s, w_L)$  and  $\tau^{**}(s, w_H) \leq \tau^*(s, w_H)$ .*
- (b) *For  $p_H = \hat{p}_H$  and for  $p_H$  such that  $p_L = 1 - \frac{f_H}{1-f_H} p_H$  the allocation is undistorted.*
- (c) *For  $p_H > \hat{p}_H$  redistribution and public-goods provision are distorted upwards,  $v_L^{**}(s) > v_L^*(s)$ ,  $v_H^{**}(s) < v_H^*(s)$ , and  $q^{**}(s) > q^*(s)$ . Also, the implicit marginal tax rates are higher,  $\tau^{**}(s, w_L) > \tau^*(s, w_L)$  and  $\tau^{**}(s, w_H) = \tau^*(s, w_H)$ .*

## 5.3 Sufficiency

The social choice function  $(q^{**}, c^{**}, y^{**})$  which is characterized in Propositions 5 and 6 satisfies some of the necessary conditions for coalition-proofness which were identified in Proposition 3. In particular, if  $(q^{**}, c^{**}, y^{**})$  coincides with a the solution to a problem of the  $\mathcal{P}_L(p_H, f_H)$ -type, then, by the definition of this problem, the constraint that  $V(s, w_L, \theta_H)$  must be a non-decreasing function of  $p_L$  is satisfied. If it coincides with a the solution to a problem of the  $\mathcal{P}_H(p_L, f_H)$ -type, then, by the definition of this problem, the constraint that  $V(s, w_H, \theta_L)$  must be a non-increasing function of  $p_H$  is satisfied. The following Proposition establishes that all necessary conditions that were not explicitly taken into account by this approach, are also satisfied. Moreover, it is shown that, for any  $s$ , at most one these conditions holds as an equality. By Corollary 1 this

is an “almost” sufficient condition for coalition-proofness; i.e., if  $(q^{**}, c^{**}, y^{**})$  is not coalition-proof itself, then the Corollary implies that there is an alternative social choice function that is coalition-proof and arbitrarily close to  $(q^{**}, c^{**}, y^{**})$ .

**Proposition 7** *The social choice function  $(q^{**}, c^{**}, y^{**})$  satisfies all the constraints in Proposition 3. More specifically, for every  $s$ , one of those constraints holds as an equality, and all others hold as strict inequalities.*

It is straightforward to show that the constraint that  $V(s, w_H, \theta_H)$  must be a non-decreasing function of  $p_H$  is always satisfied: individuals with a high-skill level and a high taste parameter always have an above-average valuation of the public good. Hence, from their perspective the provision level is always too low, so that they want the policy maker to believe that  $p_H$  is high. Consequently,  $V^{**}(s, w_H, \theta_H)$  is a strictly increasing function of  $p_H$ . A symmetric argument implies that  $V^{**}(s, w_L, \theta_L)$  is a strictly decreasing function of  $p_L$ .

The more difficult part is to show that in the region where problems of the  $\mathcal{P}_L(p_H, f_H)$ -type are relevant for a characterization  $(q^{**}, c^{**}, y^{**})$ ,  $V^{**}(s, w_H, \theta_L)$  is a strictly decreasing function of  $p_H$ , and that, in the domain of  $\mathcal{P}_H(p_L, f_H)$ -problems,  $V^{**}(s, w_L, \theta_H)$  is a strictly increasing function of  $p_L$ . Recall that the Mirrleesian social choice function  $(q^*, c^*, y^*)$  also has this property (see Figure 4). In the Appendix, it is shown that it carries over to the present setting.

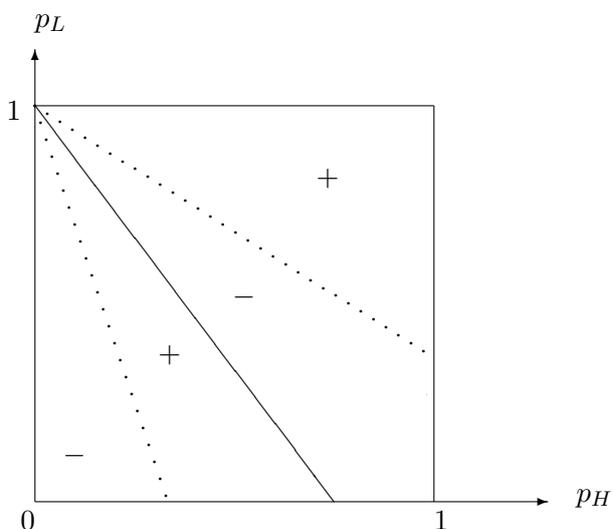
## 5.4 Summary and empirical implications

Figure 5 below summarizes the characterization of the optimal robust and coalition-proof social choice function  $(q^{**}, c^{**}, y^{**})$ . It compares it to the Mirrleesian social choice function  $(q^*, c^*, y^*)$ . Both are such that the public-goods provision level is strictly increasing in  $p_L$  and  $p_H$ .<sup>23</sup> With the Mirrleesian approach, this is neither accompanied by a departure from optimal utilitarian redistribution (we have throughout that  $f_H V_{H1} + 1 - f_H = 0$ ) nor by a change in marginal tax rates. By contrast, if we require coalition-proofness, preference shocks induce simultaneous changes of public-goods provision, redistribution, and marginal tax rates. Figure 6 illustrates this pattern. In this figure, a “+” sign indicates more public-goods provision, redistribution and taxation as compared to the Mirrleesian outcome. Likewise, a “−” sign indicates that there is less in any one of those dimensions.

The figure shows that upward distortions tend to be associated with high values of  $p_H$  and  $p_L$ , whereas downward distortions tend to be associated with low values of  $p_L$  and  $p_H$ . Equivalently, high public-goods provision levels are associated with high levels of redistribution and high marginal tax rates, whereas low public-goods provision levels are associated with low levels of redistribution and low marginal tax rates. Hence, at an empirical level, the optimal policy that is derived under the requirement of coalition-proofness implies a positive correlation between public-goods provision, redistribution and income tax rates. The conventional Mirrleesian approach in Section 3 would not give rise to such a correlation.

<sup>23</sup>For the provision rule  $q^{**}$ , this is formally shown in the Appendix, see Lemma 2.

Figure 6: The pattern of distortions



Of course, one has to be careful when trying to interpret the results from a normative exercise in a positive way. In the real world, public policy is not determined by optimal mechanism design. However, the requirement of coalition-proofness has a positive interpretation. It is a political economy constraint capturing the possibility that like-minded individuals may coordinate their actions – e.g., when they take voting decisions – to get their preferred policies. We may therefore hope that the incorporation of this additional constraint brings the outcomes of our normative analysis closer to the forces that determine policy outcomes in the real world.

## 6 Concluding Remarks

This paper has analyzed a large economy in which individuals are privately informed about their productive abilities and their preferences for public goods. Moreover, there is aggregate uncertainty with respect to the cross-sectional distribution of these characteristics. The analysis has identified two sets of incentive conditions for public policy. Individual incentive compatibility constraints take into account how individuals respond to an income tax system that determines their after-tax income as a function of their labor supply. Collective incentive compatibility constraints take care of the possibility that individuals may lobby for certain tax and expenditure policies, thus addressing the political reactions that may be triggered by the policy mechanism.

Collective incentive compatibility requires that if a group of individuals experiences a shift in their public-goods preferences such that their willingness to pay for a public good is increased, then it must be true that more of the public good is provided (otherwise these individuals understate their public-goods preferences) and that these individuals pay more taxes (otherwise they exaggerate their preferences). More generally speaking, the tax system confronts individuals with prices for public goods. These prices have to be set in an “appropriate” manner, namely in such a way that the “true” demand for public goods can be determined.

We have shown that introducing these considerations into a model of optimal taxation and public-goods provision induces an interdependence of tax and expenditure policies. At an empirical level, if the the optimal policy were implemented, this would give rise to a positive correlation between the expenditures on public goods, direct income transfers and marginal tax rates.

## References

- Al-Najjar, N. (2004). Aggregation and the law of large numbers in large economies. *Games and Economic Behavior*, 47:1–35.
- Bassetto, M. and Phelan, C. (2008). Tax riots. *Review of Economic Studies*, 75:649–669.
- Bergemann, D. and Morris, S. (2005). Robust mechanism design. *Econometrica*, 73:1771–1813.
- Bernheim, B., Peleg, B., and Whinston, M. (1986). Coalition-proof Nash equilibria I. Concepts. *Journal of Economic Theory*, 42:1–12.
- Bierbrauer, F. (2009). A note on optimal income taxation, public-goods provision and robust mechanism design. *Journal of Public Economics*, 93:667–670.
- Bierbrauer, F. and Boyer, P. (2010). The Pareto-frontier in a simple Mirrleesian model of income taxation. Mimeo, Max Planck Institute for Research on Collective Goods, Bonn.
- Bierbrauer, F. and Hellwig, M. (2010). Public-good provision in a large economy. Preprint 2010/02, Max Planck Institute for Research on Collective Goods, Bonn.
- Boadway, R. and Keen, M. (1993). Public goods, self-selection and optimal income taxation. *International Economic Review*, 34:463–478.
- Clarke, E. (1971). Multipart pricing of public goods. *Public Choice*, 11:17–33.
- Gahvari, F. (2006). On the marginal costs of public funds and the optimal provision of public goods. *Journal of Public Economics*, 90:1251–1262.
- Golosov, M., Kocherlakota, N., and Tsyvinski, A. (2003). Optimal indirect and capital taxation. *Review of Economic Studies*, 70:569–587.
- Groves, T. (1973). Incentives in teams. *Econometrica*, 41:617–663.
- Guesnerie, R. (1995). *A Contribution to the Pure Theory of Taxation*. Cambridge University Press.
- Hammond, P. (1979). Straightforward individual incentive compatibility in large economies. *Review of Economic Studies*, 46:263–282.
- Hellwig, M. (2007). A contribution to the theory of optimal utilitarian income taxation. *Journal of Public Economics*, 91:1449–1477.

- Jackson, M. (2001). A crash course in implementation theory. *Social Choice and Welfare*, 18:655–708.
- Judd, K. (1985). The law of large numbers with a continuum of i.i.d. random variables. *Journal of Economic Theory*, 35:19–25.
- Kamien, M. and Schwartz, N. (1991). *Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management*. Elsevier, New York.
- Kocherlakota, N. (2005). Zero expected wealth taxes: A Mirrlees approach to dynamic optimal taxation. *Econometrica*, 73:1587–1621.
- Kocherlakota, N. and Phelan, C. (2009). On the robustness of laissez-faire. *Journal of Economic Theory*, forthcoming.
- Kreiner, C. and Verdelin, N. (2010). Optimal provision of public goods: A synthesis. Mimeo, University of Copenhagen.
- Laffont, J. and Martimort, D. (1997). Collusion under asymmetric information. *Econometrica*, 65:875–911.
- Laffont, J. and Martimort, D. (2000). Mechanism design with collusion and correlation. *Econometrica*, 68:309–342.
- Ledyard, J. (1978). Incentive compatibility and incomplete information. *Journal of Economic Theory*, 18:171–189.
- Mirrlees, J. (1971). An exploration in the theory of optimum income taxation. *Review of Economic Studies*, 38:175–208.
- Stiglitz, J. (1982). Self-selection and Pareto-efficient taxation. *Journal of Public Economics*, 17:213–240.
- Sun, Y. (2006). The exact law of large numbers via Fubini extension and characterization of insurable risks. *Journal of Economic Theory*, 126:31–69.
- Weymark, J. (1986). A reduced-form optimal nonlinear income tax problem. *Journal of Public Economics*, 30:199–217.

## A Appendix

### A.1 Proof of Proposition 1

We fix  $T, \mathcal{T}$ , and  $\pi$ . By the standard version of the revelation principle, a social choice function  $(q, c, y)$  is implementable as a Bayes-Nash equilibrium by some mechanism  $M$  on a given type space, if and only if it is truthfully implementable, i.e., if and only if there exists a direct

mechanism  $M$  with an action set  $A = T$  and outcome functions  $Q$ , and  $C$  and  $Y$  such that (i) truthtelling is a Bayes-Nash equilibrium; i.e., for all  $t$ ,

$$t \in \operatorname{argmax}_{t' \in T} \int_{\mathcal{M}(T)} U(Q(\delta), C(\delta, t'), Y(\delta, t'), w(t), \theta(t)) d\beta(\delta | t), \quad (19)$$

and (ii) the equilibrium allocation is equal to the allocation stipulated by the social choice function; for every  $\delta$ ,

$$Q(\delta) = q(s(\delta)) \quad (20)$$

and, for every  $t$ ,

$$C(\delta, t) = c(s(\delta), w(t), \theta(t)) \text{ and } Y(\delta, t) = y(s(\delta), w(t), \theta(t)). \quad (21)$$

We first show that (b)  $\Rightarrow$  (a). Consider an incentive compatible social choice function  $(q, c, y)$ . For an arbitrary belief system  $\beta$  construct a direct mechanism  $M = [(T, T), Q, C, Y]$  such that (20) and (21) hold. We seek to verify that, for every  $t$ ,

$$\begin{aligned} t &\in \operatorname{argmax}_{t' \in T} \int_{\mathcal{M}(T)} U(Q(\delta), C(\delta, t'), Y(\delta, t'), w(t), \theta(t)) d\beta(\delta | t) \\ &= \operatorname{argmax}_{t' \in T} \int_S U(q(s), c(s, w(t'), \theta(t')), y(s, w(t'), \theta(t')), w(t), \theta(t)) d\hat{\beta}(s | t), \end{aligned}$$

where, for any  $S' \subset S$ ,  $\hat{\beta}(S' | t) := \beta(\{\delta \in \mathcal{M}(T) \mid s(\delta) \in S'\} | t)$ . Equivalently, for every  $t$ ,  $(w(t), \theta(t))$  solves

$$\max_{(w', \theta') \in W \times \Theta} \int_S U(q(s), c(s, w', \theta'), y(s, w', \theta'), w(t), \theta(t)) d\hat{\beta}(s | t).$$

This follows immediately from the fact that  $(q, c, y)$  is incentive compatible.

We now show that (a)  $\Rightarrow$  (b). Consider a type space where  $\beta$  is such that for some  $s'$ ,  $\beta(\{\delta \mid s(\delta) = s'\} | t) = 1$ , for all  $t$ . Suppose that a direct mechanism  $(T, Q', C', Y')$  truthfully implements  $(q, c, y)$ . Using conditions (20) and (21) to substitute for  $Q'$ ,  $C'$ , and  $Y'$ , the equilibrium condition in (19) becomes: for all  $t$  and all  $t'$ ,

$$\begin{aligned} &U(q(s'), c(s', w(t), \theta(t)), y(s', w(t), \theta(t)), w(t), \theta(t)) \\ &\geq U(q(s'), c(s', w(t'), \theta(t')), y(s', w(t'), \theta(t')), w(t), \theta(t)); \end{aligned}$$

or, equivalently, for all  $(w, \theta)$  and  $(w', \theta')$ ,

$$U(q(s), c(s, w, \theta), y(s, w, \theta), w, \theta) \geq U(q(s), c(s, w', \theta'), y(s, w', \theta'), w, \theta).$$

Since the choice of  $s'$  was arbitrary, the latter inequality holds for all  $s \in S$ . Hence,  $(q, c, y)$  is individually incentive-compatible.

## A.2 Proof of Proposition 3

Suppose there is a mechanism  $M = [(A, \mathcal{A}), Q, C, Y]$  with an equilibrium  $\sigma^*$  that robustly implements a social choice function  $(g, c, y)$  as a coalition-proof Bayes-Nash equilibrium. In particular, this requires that the mechanism reaches the social choice function; i.e., for every  $\delta$ , conditions (12) and (13) are fulfilled.

We show that this implies that  $V(s, w_L, \theta_H)$  must be a non-decreasing function of  $p_L$ . (All other claims in Proposition 3 follow from a symmetric argument.) Suppose otherwise, then there exist  $f_H, p_H, p_L$  and  $p'_L$  with  $p'_L > p_L$  so that

$$V((f_H, p_L, p_H), w_L, \theta_H) > V((f_H, p'_L, p_H), w_L, \theta_H). \quad (22)$$

In the following, we will construct a deviation and show that there is a type space so that it satisfies conditions (a), (b) and (c) in Definition 1. This contradicts the assumption that  $\sigma^*$  is a coalition-proof Bayes-Nash equilibrium on every type space.

### Step 1: Construction of a deviation

Intuitively, we seek to construct a deviation  $\sigma'_{T'}$  for individuals with a type in  $T' = \{t \mid (w(t), \theta(t)) = (w_L, \theta_H)\}$  which works as follows: a type  $t' \in T'$ , plays according to  $\sigma^*(t')$  with probability  $\frac{p_L}{p'_L}$  and plays according to  $\sigma^*(\hat{t})$ , where  $\hat{t} \in \hat{T} = \{t \mid (w(t), \theta(t)) = (w_L, \theta_L)\}$ , otherwise.

It proves convenient to define first two strategies that, with a direct mechanism, could be interpreted as a “lie” by a set of deviating types and as “honesty” or truth-telling by all others. For types in  $T'$ , define the “lie”  $\ell'_{T'} : T' \rightarrow \mathcal{M}(T)$  such that, for every  $t' \in T'$ ,  $\ell'_{T'}(\{t'\} \mid t') = \frac{p_L}{p'_L}$ , and  $\ell'_{T'}(\hat{T} \mid t') = 1 - \frac{p_L}{p'_L}$ . Let the function  $h_{T \setminus T'} : T \setminus T' \rightarrow \mathcal{M}(T)$  be such that for all  $t \in T \setminus T'$ ,  $h_{T \setminus T'}(\{t\} \mid t) = 1$ . Observe that the pair  $(\ell'_{T'}, h_{T \setminus T'})$  induces, for each  $\delta \in \mathcal{M}(T)$ , an announced cross-sectional distribution of types  $\bar{\delta}(\delta)$  with

$$\bar{\delta}(\tilde{T} \mid \delta) = \int_{t' \in T'} \ell'_{T'}(\tilde{T} \mid t') d\delta(t') + \int_{t \in T \setminus T'} h_{T \setminus T'}(\tilde{T} \mid t) d\delta(t), \quad (23)$$

for any subset  $\tilde{T}$  of  $T$ .

With reference to  $\ell'_{T'}$ , we now define a strategy  $\sigma'_{T'}$  for the game induced by mechanism  $M$  in the following way: for every  $t' \in T'$  and every subset  $A'$  of  $A$ , let

$$\sigma'_{T'}(A' \mid t') = \int_{\hat{t} \in \hat{T}} \sigma^*(A' \mid \hat{t}) d\ell'_{T'}(\hat{t} \mid t'). \quad (24)$$

This construction ensures that, for every  $\delta$ , the distribution of actions that results if individuals with types in  $T'$  behave according to  $\sigma'_{T'}$  and all others follow  $\sigma^*_{T \setminus T'}$  equals the distribution of actions that results if all individuals follow  $\sigma^*$  and the distribution of types equals  $\bar{\delta}(\delta)$ . Formally, for every  $\delta$ ,

$$\alpha(\bar{\delta}(\delta), \sigma^*) = \alpha(\delta, (\sigma^*_{T \setminus T'}, \sigma'_{T'})). \quad (25)$$

To see that this true, note that for any subset  $A'$  of  $A$ ,

$$\begin{aligned}
& \alpha(A' \mid \bar{\delta}(\delta), \sigma^*) \\
&= \int_{t' \in T} \sigma^*(A' \mid t') d\bar{\delta}(t \mid \delta) \\
&= \int_{t' \in T} \sigma^*(A' \mid t') d \left( \int_{t \in T'} \ell'_{T'}(t' \mid t) d\delta(t) + \int_{t \in T \setminus T'} h_{T \setminus T'}(t' \mid t) d\delta(t) \right) \\
&= \int_{t \in T'} \int_{t' \in T} \sigma^*(A' \mid t') d\ell'_{T'}(t' \mid t) d\delta(t) + \int_{t \in T \setminus T'} \int_{t' \in T} \sigma^*(A' \mid t') dh_{T \setminus T'}(t' \mid t) d\delta(t) \\
&= \int_{t \in T'} \sigma'_{T'}(A' \mid t) d\delta(t) + \int_{t \in T \setminus T'} \sigma^*_{T \setminus T'}(A' \mid t) d\delta(t) \\
&= \alpha(A' \mid \delta, (\sigma^*_{T \setminus T'}, \sigma'_{T'})) .
\end{aligned}$$

### Step 2: Consider a specific type space

Consider a type space with a belief system so that, for some  $\delta$  such that  $s(\delta) = (f_H, p_H, p'_L)$ ,  $\beta(\{\delta\} \mid t) = 1$ , for all  $t$ . The distribution of types  $\bar{\delta}(\delta)$  that is communicated to the mechanism if types in  $T'$  behave according to  $\sigma'_{T'}$  and types in  $T \setminus T'$  behave according to  $\sigma^*_{T \setminus T'}$ , therefore is such that

$$s(\bar{\delta}(\delta)) = (f_H, p_H, p_L) , \tag{26}$$

with probability 1.

### Step 3: Show that, on this type space, the deviation makes the deviators better off

By equations (12) and (13), given the strategy  $(\sigma^*_{T \setminus T'}, \sigma'_{T'})$ , the expected payoff of a type  $t' \in T'$  equals

$$\Pi(t') := \frac{p_L}{p'_L} V(s(\bar{\delta}(\delta)), w(t'), \theta(t')) + \left(1 - \frac{p_L}{p'_L}\right) \Phi(t') ,$$

where

$$\Phi(t') := E \left[ \theta(t') q(s(\bar{\delta}(\delta))) + u(c(s(\bar{\delta}(\delta)), w(\hat{t}), \theta(\hat{t}))) - \frac{y(s(\bar{\delta}(\delta)), w(\hat{t}), \theta(\hat{t}))}{w(t')} \mid \hat{t} \in \hat{T} \right] .$$

By Proposition 1 robust implementability of a social choice function as a Bayes-Nash equilibrium implies individual incentive compatibility of a social choice function. As we observed in Section 3, this in turn implies that, for any  $s$ ,  $c$  and  $y$  may depend on  $w$ , but not on  $\theta$ . Also, note that, by construction of the set  $\hat{T}$ , types in  $T'$  choose only actions that communicate their skill level truthfully to the mechanism. Hence,  $\Phi(t') = V(s(\bar{\delta}(\delta)), w(t'), (t'))$  so that so that, for every  $t'$  in  $T'$ ,

$$\Pi(t') = V(s(\bar{\delta}(\delta)), w(t'), (t')) .$$

This observation in conjunction with equations (22) and (26) implies that types in  $T'$  are made strictly better off by this deviation.

**Step 4: Show that, on this type space,  $(\sigma_{T \setminus T'}^*, \sigma_{T'}')$  is a Bayes-Nash equilibrium**

Consider an alternative type space with a belief system so that, for all  $t$ ,

$$\beta(\{\delta \mid s(\delta) = (f_H, p_L, p_H)\} \mid t) = 1 .$$

Since  $\sigma^*$  robustly implements the given social choice function, behaving according to  $\sigma^*(t)$  is a best response for every type  $t$ , given these beliefs. Since, for any  $s, c$  and  $y$  may depend on  $w$ , but not on  $\theta$ , behaving according to  $\sigma^*(\hat{t})$ , for some  $\hat{t} \in \hat{T}$  is also a best response for an individual with type  $t' \in T'$ .

But this implies that behaving according to  $(\sigma_{T \setminus T'}^*, \sigma_{T'}')$  is also a best response for each type  $t$  under the assumption made in Step 2, namely that the type space is such that

$$\beta(\{\delta \mid s(\delta) = (f_H, p'_L, p_H)\} \mid t) = \beta(\{\delta \mid s(\bar{\delta}(\delta)) = (f_H, p_L, p_H)\} \mid t) = 1 ,$$

which implies that the deviation satisfies (26).

**Step 5: Show that, on this type space, the deviation is subcoalition-proof**

The deviating individuals have the same preferences,  $(w(t'), \theta(t')) = (w_L, \theta_H)$ , for all  $t' \in T'$ , and the same beliefs, by the assumptions made in Step 2. Hence, there exists no strict subset of  $T'$  which could undermine the subcoalition-proofness of the deviation  $\sigma_{T'}'$ .

**A.3 Proof of Proposition 4**

Given a measure space of types  $(T, \mathcal{T})$  and a function  $\pi = (w, \theta) : T \mapsto W \times \Theta$ , and given a social choice function  $(q, c, y) \in \Omega(\epsilon)$ , we construct a direct mechanism  $M = [(T, \mathcal{T}), Q, C, Y]$  so that, for all  $\delta \in \mathcal{M}(T)$  and all  $t \in T$ ,

$$Q(\delta) = q(s(\delta)), C(\delta, t) = c(s(\delta), w(t)), \text{ and } Y(\delta, t) = y(s(\delta), w(t)) . \quad (27)$$

This construction ensures that the mechanism achieves the social choice function in a truthtelling equilibrium. More formally, the strategy  $h : T \rightarrow \mathcal{M}(T)$  with  $h(\{t\} \mid t) = 1$ , for all  $t$ , is Bayes-Nash equilibrium of the game induced by this mechanism, for every belief system  $\beta$ . This was shown in the proof of Proposition 1. In the following, we seek to show that this equilibrium is coalition-proof on every type space with a moderately uninformative belief system  $\beta$ .

**Step 1: No deviations that involve lies about skills**

Suppose there is a set of types  $T'$  who deviate from  $h$  and instead behave according to a lie  $\ell'_{T'} : T' \rightarrow \mathcal{M}(T)$ . We say that such a lie involves lies about skills if there is  $t' \in T'$  so that

$$l(\hat{w} \mid t') := \ell'_{T'}(\{\hat{t} \mid w(\hat{t}) \neq w(t')\} \mid t') > 0 . \quad (28)$$

We show in the following that any such deviation violates condition (a) in Definition 1 and therefore does not challenge the coalition-proofness of the truthtelling equilibrium.

Let  $\bar{\delta}(\delta) \in \mathcal{M}(T)$  (see the definition in equation (23)) be the cross-section distribution of types that is communicated to the mechanism if types in  $T'$  behave according to  $\ell'_{T'}$  and types in  $T \setminus T'$  behave according to  $h_{T \setminus T'}$ .

Given that (27) holds, the expected payoff of an individual with a type  $t' \in T'$  whose behavior satisfies (28) can be written as

$$\int_{\mathcal{M}(T)} l(\hat{w} | t') \left\{ \theta(t)q(s(\bar{\delta}(\delta))) + u(c(s(\bar{\delta}(\delta))), \hat{w}) - \frac{y(s(\bar{\delta}(\delta)), \hat{w})}{w(t)} \right\} \\ + (1 - l(\hat{w} | t')) \left\{ \theta(t)q(s(\bar{\delta}(\delta))) + u(c(s(\bar{\delta}(\delta))), w(t)) - \frac{y(s(\bar{\delta}(\delta)), w(t))}{w(t)} \right\} d\beta(\delta | t), \quad (29)$$

where  $\hat{w} \neq w(t)$ .

Now suppose that the individual in question would instead communicate his skill level truthfully with probability 1. The resulting payoff equals

$$\int_{\mathcal{M}(T)} \left\{ \theta(t)q(s(\bar{\delta}(\delta))) + u(c(s(\bar{\delta}(\delta))), w(t)) - \frac{y(s(\bar{\delta}(\delta)), w(t))}{w(t)} \right\} d\beta(\delta | t). \quad (30)$$

By the constraints in (16) we have that

$$\theta(t)q(s(\bar{\delta}(\delta))) + u(c(s(\bar{\delta}(\delta))), w(t)) - \frac{y(s(\bar{\delta}(\delta)), w(t))}{w(t)} \\ > \theta(t)q(s(\bar{\delta}(\delta))) + u(c(s(\bar{\delta}(\delta))), \hat{w}) - \frac{y(s(\bar{\delta}(\delta)), \hat{w})}{w(t)},$$

which implies that the expression in (30) is strictly larger than the expression in (29). This shows that, for a type  $t' \in T'$ , behaving in such a way that (28) holds is not a best response. Hence,  $(h_{T \setminus T'}, \ell'_{T'})$  is not a Bayes-Nash equilibrium strategy.

## Step 2: No deviation so that all participating individuals have the same preferences

Suppose the deviating set of types  $T'$  is such that  $t' \in T'$  and  $\hat{t}' \in T'$  imply that  $\pi(t') = \pi(\hat{t}')$ . For the sake of concreteness assume that  $\pi(t') = (w_H, \theta_L)$  for all  $t' \in T'$ . We know by Step 1 that there is no deviation that involves lies about skills and challenges the coalition-proofness of equilibrium  $h$ . Hence, suppose that all participating individuals truthfully communicate their skills

$$\ell'_{T'}(\{\hat{t} | w(\hat{t}) \neq w(t')\} | t') = 0, \quad (31)$$

and that some lie about their taste parameter with positive probability,

$$\ell'_{T'}(\{\hat{t} | \theta(\hat{t}) = \theta_H\} | t') > 0. \quad (32)$$

Consequently, for every  $\delta$ ,  $s(\delta) = (f_H(\delta), p_H(\delta), p_L(\delta))$  and  $s(\bar{\delta}(\delta)) = (f_H(\bar{\delta}(\delta)), p_H(\bar{\delta}(\delta)), p_L(\bar{\delta}(\delta)))$  are such that

$$f_H(\delta) = f_H(\bar{\delta}(\delta)), \quad p_H(\delta) < p_H(\bar{\delta}(\delta)) \quad \text{and} \quad p_L(\delta) = p_L(\bar{\delta}(\delta)).$$

Since the given social choice function satisfies the monotonicity constraint

$$\frac{\partial V(s, w_H, \theta_L)}{\partial p_H} \leq 0,$$

this deviation will fail to make the participating types better off, i.e., it violates condition (b) in Definition 1, and therefore does not challenge the coalition-proofness of the truthtelling equilibrium.

### Step 3: No deviation with heterogeneous preferences

Now suppose that the deviating set of types  $T'$  is such that there are  $t' \in T'$  and  $\hat{t}' \in T'$  so that  $\pi(t') \neq \pi(\hat{t}')$ . Again, we may assume that the deviation involves no lies about skills so that (31) holds. Consequently, we have for all  $\delta$  that  $f_H(\delta) = f_H(\bar{\delta}(\delta))$ .

Assume, for the sake of concreteness, that there is  $T'' \subset T'$  so that  $t'' \in T''$  implies that  $\pi(t'') = (w_H, \theta_L)$  and that these individuals lie about their taste parameter with positive probability,

$$\ell'_{T'}(\{\hat{t} \mid \theta(\hat{t}) = \theta_H\} \mid t'') > 0 . \quad (33)$$

Given that the monotonicity constraint

$$\frac{\partial V(s, w_H, \theta_L)}{\partial p_H} \leq 0 ,$$

holds, these types will benefit from the deviation  $\ell'_{T'}$  only if there is a subset  $D$  of  $\mathcal{M}(T)$  with  $\beta(D \mid t') > 0$  for all  $t' \in T'$  with  $\pi(t') = (w_H, \theta_L)$ , which has the following property:  $\delta \in D$  implies that

$$p_L(\delta) \neq p_L(\bar{\delta}(\delta)) , \quad \text{or} \quad p_H(\delta) > p_H(\bar{\delta}(\delta)) .$$

Since we have limited information to type spaces with moderately uninformative belief systems,  $\beta(D \mid t') > 0$  for all  $t' \in T'$  with  $\pi(t') = (w_H, \theta_L)$  implies in fact that  $\beta(D \mid t') > 0$  for all  $t' \in T'$ , i.e., all participants of the deviation assign positive probability mass to the set  $D$ .

Suppose that the set  $D$  is such that  $p_H(\delta) > p_H(\bar{\delta}(\delta))$ , for all  $\delta \in D$ . (The alternative cases so that  $\delta \in D$  implies  $p_L(\delta) < p_L(\bar{\delta}(\delta))$  or  $p_L(\delta) > p_L(\bar{\delta}(\delta))$  can be treated in exactly the same way.) This implies that the set  $T'$  includes high-skilled individuals with a high taste parameter who announce a low taste parameter with positive probability: there is  $\hat{T}'' \subset T'$  so that  $\hat{t}'' \in \hat{T}''$   $\pi(\hat{t}'') = (w_H, \theta_H)$ , and

$$\ell'_{T'}(\{\hat{t} \mid \theta(\hat{t}) = \theta_L\} \mid \hat{t}'') > 0 . \quad (34)$$

The assumptions that, for every  $s$ , at most one of the monotonicity constraints in Proposition 3 is binding and that the belief system is moderately uninformative have the following implication: there is a subset  $\tilde{D}$  of  $D$  so that  $\beta(\tilde{D} \mid t') > 0$  for all  $t' \in T'$ , and conditional on  $\delta \in \tilde{D}$ , types in  $T''$  or types in  $\hat{T}''$  are made strictly better off if they reduce the probability of a lie, taking the behavior of all other individuals as given. To see this, suppose first that types in  $\hat{T}''$  change their behavior and now follow a strategy  $\ell''_{\hat{T}''}$  with

$$\ell''_{\hat{T}''}(\{\hat{t} \mid \theta(\hat{t}) = \theta_L\} \mid \hat{t}'') < \ell'_{T'}(\{\hat{t} \mid \theta(\hat{t}) = \theta_L\} \mid \hat{t}'') . \quad (35)$$

Let  $\hat{\delta}(\delta)$  be the cross-section distribution of types that is communicated to the mechanism given that the true cross-section distribution of types is  $\delta$  and that individuals behave according to the strategy profile  $(h_{T \setminus T'}, \ell'_{T' \setminus \hat{T}''}, \ell''_{\hat{T}''})$ . We have that, for all  $\delta \in \mathcal{M}(T)$ ,

$$p_H(\hat{\delta}(\delta)) > p_H(\bar{\delta}(\delta)) \quad \text{and} \quad p_L(\hat{\delta}(\delta)) = p_L(\bar{\delta}(\delta)) .$$

Given that the monotonicity constraint

$$\frac{\partial V(s, w_H, \theta_H)}{\partial p_H} \geq 0, \quad (36)$$

holds, for all  $s$ , the outcome of this deviation makes all types in  $\hat{T}''$  weakly better off. It makes them also strictly better off, provided that there is a subset  $\tilde{D}$  with  $\beta(\tilde{D} | t') > 0$  so that (36) holds as a strict inequality. Finally, observe that  $(h_{T \setminus T'}, \ell'_{T' \setminus \hat{T}''}, \ell''_{\hat{T}''})$  is a Bayes-Nash equilibrium strategy because individuals communicate their skill levels truthfully, and, individual outcomes do not depend on announced taste parameters (see equation (27)). Hence, if there is a subset  $\tilde{D}$  of  $D$  with  $\beta(\tilde{D} | t') > 0$  so that so that (36) holds as a strict inequality, the deviation  $\ell'_{T'}$  fails to be subcoalition-proof.

Now assume that there is no such set  $\tilde{D}$ . Then, since for every  $s$  at most one monotonicity constraints in Proposition 3 is binding, it has to be the case that the monotonicity constraint  $\frac{\partial V(s, w_H, \theta_L)}{\partial p_H} \leq 0$  holds as a strict inequality with probability 1, conditional on the event  $\delta \in D$ . But this implies that now individuals with types in  $T''$  benefit from reducing the probability of a lie. Again, this implies that  $\ell'_{T'}$  fails to be subcoalition-proof.

#### A.4 Proof of Corollary 1

Consider a social choice function  $(q, c, y) \in \Omega(0)$ . Since  $\bigcup_{0 \leq \epsilon \leq \bar{\epsilon}} \Omega(\epsilon)$  is compact, we can construct a sequence of social choice functions  $\{(q^k, c^k, y^k)\}_{k=1}^{\infty}$  with  $(q^k, c^k, y^k) \in \Omega(\frac{\epsilon}{k})$ , for each  $k$ , which converges to  $(q, c, y)$ . By continuity of  $U$  this implies that also, for each  $s, w$ , and  $\theta$ ,  $U(q^k(s), c^k(s, w), y^k(s, w), w, \theta)$  converges to  $U(q(s), c(s, w), y(s, w), w, \theta)$ .

#### A.5 Proof of Proposition 5

##### Step 1: Reformulate problem $\mathcal{P}_L(p_H, f_H)$

The function  $V_H$ , defined by

$$\begin{aligned} V_H(v_L, \rho) &:= \max u(c_H) - \frac{y_H}{w_H} \\ \text{s.t.} & \quad u(c_H) - \frac{y_H}{w_H} \geq u(c_L) - \frac{y_L}{w_H}, u(c_L) - \frac{y_L}{w_L} \geq u(c_H) - \frac{y_H}{w_L}, \\ & \quad f_H(y_H - c_H) + (1 - f_H)(y_L - c_L) = \rho, u(c_L) - \frac{y_L}{w_L} = v_L, \end{aligned}$$

is the Pareto-frontier in a simplified version of the Mirrleesian optimal income tax problem that does not contain a decision on public-goods provision. The properties of this Pareto-frontier are extensively studied in Bierbrauer and Boyer (2010). The derivations that follow make repeated use of these properties.

At the optimal allocation, it has to be the case that, for all possible values of  $p_L$ , the utility of the high-skilled satisfies

$$v_H^{**}(s) = V_H(v_L^{**}(s), r(q^{**}(s))). \quad (37)$$

To see why this is true, note that the monotonicity constraint

$$\frac{\partial V(s, w_L, \theta_H)}{\partial p_L} \geq 0,$$

restricts how the low-skilled individuals' utility may vary with  $s$ , or, equivalently, with  $p_L$ . There is no such constraint for the high-skilled. Hence, once we have determined  $v_L(s)$  and  $r(q(s))$ , optimality considerations require that we make  $v_H(s)$  as large as possible. This implies that a solution to problem  $\mathcal{P}_L(p_H, f_H)$  has to satisfy (37), for every  $s$ . This makes it possible to reformulate problem  $\mathcal{P}_L(p_H, f_H)$  in such a way that we can treat the utility level of the low-skilled and the public-goods provision level as choice variables.

Since, for the analysis of problem  $\mathcal{P}_L(p_H, f_H)$ ,  $p_H$  and  $f_H$  are fixed parameters, we may interpret both the utility of the low-skilled  $v_L$  and the public-goods provision level as functions of  $p_L$ , and suppress the dependence on the whole vector  $s = (f_H, p_H, p_L)$ . With a slight abuse of notation, problem  $\mathcal{P}_L(p_H, f_H)$  may therefore be stated as follows: choose the functions  $v_L : p_L \mapsto v_L(p_L)$  and  $q : p_L \mapsto q(p_L)$  in order to maximize

$$\int_{\kappa(p_H)}^1 \{\bar{\theta}(p_L)q(p_L) + f_H V_H(v_L(p_L), r(q(p_L))) + (1 - f_H)v_L(p_L)\} dp_L$$

subject to the monotonicity constraint, that for all  $p_L \in [\kappa(p_H), 1]$ , with  $\kappa(p_H) := \max\{0, 1 - \frac{f_H}{1-f_H}p_H\}$ ,

$$\theta_H q'(p_L) + v'_L(p_L) \geq 0 .$$

## Step 2: Statement of optimality conditions

We use optimal control theory in order to characterize the solution to this optimization problem. Specifically, we treat  $q$  and  $v_L$  as states variables. The control variables  $u_1$  and  $u_2$  are equal to  $q'$  and  $v'_L$ ; that is, they satisfy the following equations of motion,

$$q' = g_1(u_1) , \quad \text{with} \quad g_1(u_1) = u_1 , \tag{38}$$

and

$$v'_L = g_2(u_2) , \quad \text{with} \quad g_2(u_2) = u_2 . \tag{39}$$

The monotonicity constraint can now be formulated as a constraint on the control variables,

$$h(u_1, u_2) \geq 0 , \quad \text{where} \quad h(u_1, u_2) = \theta_H u_1 + u_2 . \tag{40}$$

The optimality conditions for this problem can be conveniently stated by making use of the following Hamiltonian

$$\mathcal{H}(q, v_L, u_1, u_2) = \bar{\theta}(p_L)q + f_H V_H(v_L, r(q)) + (1 - f_H)v_L + \mu_1 g_1(u_1) + \mu_2 g_2(u_2) ,$$

where  $\mu_1$  is the costate variable associated with (38) and  $\mu_2$  is the costate variable associated with (39); and of the Lagrangean

$$\mathcal{L}(q, v_L, u_1, u_2) = \mathcal{H}(q, v_L, u_1, u_2) + \nu h(u_1, u_2) ,$$

where  $\nu \geq 0$ , is the multiplier associated with (40). The optimality conditions are as follows:<sup>24</sup>

(i) The costate variables satisfy

$$\mu'_1 = -\frac{\partial \mathcal{H}}{\partial q} \quad \text{and} \quad \mu'_2 = -\frac{\partial \mathcal{H}}{\partial v_L}, \quad (41)$$

or, equivalently,

$$\mu'_1 = -(\bar{\theta} + f_H V_{H2} r'(q))\phi_L \quad (42)$$

and

$$\mu'_2 = -(f_H V_{H1} + 1 - f_H)\phi_L, \quad (43)$$

where  $V_{Hj}$  denotes the partial derivative of the function  $V_H$  with respect to its  $j$ -th argument.

(ii) The fact that we have free start and end values for the control variables implies that

$$\mu_1(\kappa(p_L)) = \mu_1(1) = 0 \quad \text{and} \quad \mu_2(\kappa(p_L)) = \mu_2(1) = 0. \quad (44)$$

(iii) The control variables satisfy the following first order and complementary slackness conditions:

$$\frac{\partial \mathcal{L}}{\partial u_1} = 0 \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial u_2} = 0, \quad (45)$$

and

$$\nu \geq 0, \quad \text{and} \quad \nu h(u_1, u_2) = 0. \quad (46)$$

Equations (45) can equivalently be written as

$$\mu_1 + \nu \theta_H = 0, \quad (47)$$

and

$$\mu_2 + \nu = 0. \quad (48)$$

### Step 3: Some implications of the optimality conditions

We know that, for  $p_L \in (\kappa(p_H), 1)$  the constraint (40) is binding so that, over this range, we have

$$\nu(p_L) > 0. \quad (49)$$

Also, equations (44), (47), and (48) imply that

$$\nu(\kappa(p_L)) = \nu(1) = 0. \quad (50)$$

Finally, (47), and (48) also imply that

$$\mu'_1 = -\frac{1}{\theta_H} \nu', \quad (51)$$

---

<sup>24</sup>For a derivation of these optimality conditions, see Kamien and Schwartz (1991), pp. 195-197. These conditions are necessary and sufficient provided that the Lagrangean  $\mathcal{L}$  is concave in  $(q, v_L, u_1, u_2)$ . Since it is linear in  $u_1$ , and  $u_2$ , this follows from the fact that  $V_H$  is a concave function of  $v_L$  and  $r(q)$ . A proof of this assertion can be found in Bierbrauer and Boyer (2010).

and

$$\mu'_2 = -\nu' . \quad (52)$$

Using (51) and (52) in conjunction with (42) and (43) yields

$$\frac{1}{\bar{\theta}_H} (\bar{\theta} + f_H V_{H2} r'(q)) = f_H V_{H1} + 1 - f_H . \quad (53)$$

The following three Lemmas establish some properties that will prove useful subsequently.

**Lemma 1** *Condition (53) implies that, for all  $p_L \in (\kappa(p_H), 1)$ ,  $r'(q(p_L)) \geq \theta_H w_L$ .*

**Proof** We first note that the function  $V_H$  introduced in Step 1 has the following property,<sup>25</sup>

$$V_{H2} = \frac{V_{H1}}{w_L} - \frac{1}{w_H} . \quad (54)$$

Now suppose that the Lemma is false. Then we have

$$r'(q) < \theta_H w_L .$$

Using the optimality condition (53) we may solve for  $r'(q)$  and state this inequality equivalently as

$$\bar{\theta} - \theta_H (f_H V_{H1} + 1 - f_H) \leq -\theta_H w_L f_H V_{H2} .$$

Using (54) to substitute for  $V_{H2}$ , we can rewrite this condition once more as

$$\frac{\bar{\theta}}{\lambda} < \theta_H w_L ,$$

which contradicts the assumption that  $p_L \geq \kappa(p_L)$ . □

**Lemma 2** *For all  $p_L \in (\kappa(p_H), 1)$ ,  $q'(p_L) > 0$ .*

**Proof** If we totally differentiate equation (53) with respect to  $p_L$ , we obtain

$$\bar{\theta}' = f_H (V_{H11} v'_L + V_{H21} r' q' - (V_{H21} v'_L + V_{H22} r' q') r' - V_{H2} r'')$$

Using that constraint (40) is binding, this can be equivalently written as

$$\bar{\theta}' = -f_H (Q + V_{H2} r'') q'$$

where

$$Q := V_{H11} \theta_H^2 - 2V_{H12} \theta_H r' + V_{H22} (r')^2$$

---

<sup>25</sup>See Bierbrauer and Boyer (2010) for a proof.

is a quadratic form which is non-positive because the function  $V_H$  is jointly concave in  $v_L$  and  $\rho$ , see Bierbrauer and Boyer (2010) for a proof. Using that  $V_{H2} < 0$  (again, see Bierbrauer and Boyer (2010)), and that  $r'' > 0$  establishes the result.  $\square$

**Lemma 3** For all  $p_L \in (\kappa(p_H), 1)$ ,  $\nu''(p_L) \leq 0$ .

**Proof** Optimality conditions (52) and (43) imply that

$$\nu' = f_H V_{H1} + 1 - f_H .$$

Hence,

$$\nu'' = f_H V_{H11} v'_L + f_H V_{H12} r'(q) q'$$

Since constraint (40) is binding, this can be equivalently written as

$$\nu'' = q'(-f_H V_{H11} \theta_H + f_H V_{H12} r'(q)) .$$

Since (54) implies that  $V_{H12} = \frac{1}{w_L} V_{H11}$ , we can rewrite this as

$$\nu'' = -f_H V_{H11} w_L q'(\theta_H w_L - r'(q)) . \tag{55}$$

It follows from Lemmas 1 and 2 that

$$q'(\theta_H w_L - r'(q)) \leq 0 .$$

Moreover, it is shown in Bierbrauer and Boyer (2010) that  $V_{H11} \leq 0$ . Consequently, (55) implies that  $\nu'' \leq 0$ .  $\square$

#### Step 4: Implications for public-good provision and redistribution

**Lemma 4** There exists  $\hat{p}_L \in (\kappa(p_H), 1)$  so that

$$i) p_L < \hat{p}_L \text{ implies } v_L^{**}(s) < v_L^*(s), v_H^{**}(s) > v_H^*(s) \text{ and } q^{**}(s) < q^*(s).$$

$$ii) p_L = \hat{p}_L \text{ implies } v_L^{**}(s) = v_L^*(s), v_H^{**}(s) = v_H^*(s) \text{ and } q^{**}(s) = q^*(s).$$

$$iii) p_L > \hat{p}_L \text{ implies } v_L^{**}(s) > v_L^*(s), v_H^{**}(s) < v_H^*(s) \text{ and } q^{**}(s) > q^*(s).$$

**Proof Part A.** Equation (53) states that the marginal utilitarian welfare gain from increased public-goods provision is proportional to the marginal utilitarian welfare gain from increased redistribution, i.e., from an increase of  $v_L$ .  $v_L^{**}(s) < v_L^*(s)$  implies that the latter is positive. Equation (53) then requires that also  $q^{**}(s) < q^*(s)$ , and vice versa. Moreover, since the function  $V_H$  is strictly decreasing in  $v_L$  and  $\rho$ , we have that  $v_L^{**}(s) < v_L^*(s)$  implies that  $v_H^{**}(s) =$

$V_H(v_L^{**}(s), r(q^{**}(s))) > v_H^*(s) = V_H(v_L^*(s), r(q^*(s)))$ . This proves that

$$v_L^{**}(p_L) < v_L^*(p_L) \iff q^{**}(p_L) < q^*(p_L) \iff v_H^{**}(p_L) > v_H^*(p_L) .$$

Analogously, one shows that

$$v_L^{**}(p_L) = v_L^*(p_L) \iff q^{**}(p_L) = q^*(p_L) \iff v_H^{**}(p_L) = v_H^*(p_L) ,$$

and

$$v_L^{**}(p_L) > v_L^*(p_L) \iff q^{**}(p_L) > q^*(p_L) \iff v_H^{**}(p_L) < v_H^*(p_L) .$$

*Part B.* It thus remains to be shown that there is  $\hat{p}_L$  so that  $p_L < \hat{p}_L$  implies  $v_L^{**}(s) < v_L^*(s)$ ,  $p_L = \hat{p}_L$  implies  $v_L^{**}(s) = v_L^*(s)$ , and  $p_L > \hat{p}_L$  implies  $v_L^{**}(s) > v_L^*(s)$ . Optimality conditions (52) and (43) imply that

$$\nu' = f_H V_{H1} + 1 - f_H .$$

Hence, this is equivalent to showing that there is  $\hat{p}_L$  so that  $p_L < \hat{p}_L$  implies  $\nu' > 0$ ,  $p_L = \hat{p}_L$  implies  $\nu' = 0$ , and  $p_L > \hat{p}_L$  implies  $\nu' < 0$ . This follows from the following observations:  $\nu(p_L) > 0$ , for  $p_L \in (\kappa(p_H), 1)$  since the constraint (40) is binding;  $\nu(\kappa(p_L)) = \nu(1) = 0$  as an implication of the optimality conditions (44), (47) and (48); and finally the observation that  $\nu''(p_L) \leq 0$ , for  $p_L \in (\kappa(p_H), 1)$ , in Lemma 3.  $\square$

**Lemma 5** *Suppose the given parameter  $p_H$  is such that  $\eta(p_H) \geq 0$ , then  $v_L^{**}(\kappa(p_H)) = v_L^*(p_L)$ ,  $q_L^{**}(\kappa(p_H)) = q_L^*(p_L)$ , and  $v_H^{**}(\kappa(p_H)) = v_H^*(p_L)$ .*

**Proof** By the arguments in *Part A* of the proof Lemma 4 it suffices to show that  $q_L^{**}(\kappa(p_H)) = q_L^*(\kappa(p_H))$ . Suppose otherwise, i.e.,  $q_L^{**}(\kappa(p_H)) \neq q_L^*(\kappa(p_H))$ . Lemma 1 implies that, for all  $p_L$ ,  $q_L^{**}(p_L) \geq q_L^*(\kappa(p_L))$ . Hence,  $q_L^{**}(\kappa(p_H)) > q_L^*(\kappa(p_L))$ , i.e., there is overprovision of the public good, so that, for  $p_L = \kappa(p_H)$  we have,<sup>26</sup>

$$\bar{\theta} + f_H V_{H2} r'(q) < 0 .$$

From (42) this implies that for  $p_L = \kappa(p_H)$ ,  $\mu'_1 > 0$ . Optimality condition (51) then implies that  $\nu' < 0$ . However,  $\nu(p_L) > 0$ , for  $p_L \in (\kappa(p_H), 1)$  since the constraint (40) is binding; and  $\nu(\kappa(p_H)) = 0$  as an implication of the optimality conditions (44). As a consequence, (47) and (48) imply that, for  $p_L = \kappa(p_H)$ , we need to have  $\nu' > 0$ . Hence, the assumption that  $q_L^{**}(\kappa(p_H)) \neq q_L^*(\kappa(p_L))$  has led to a contradiction, and must be false.  $\square$

<sup>26</sup>Note that  $q_L^{**}(\kappa(p_H)) > q_L^*(\kappa(p_H))$  implies that  $v_L^{**}(\kappa(p_H)) > v_L^*(\kappa(p_H))$ . Jointly, these two observations imply that  $|V_{H2}|$  is larger than at an undistorted allocation.

## Step 6: implications for marginal tax rates

For an undistorted allocation, the marginal tax rates are those in Proposition 2. It is shown in Bierbrauer and Boyer (2010) that more redistribution in comparison to this benchmark, i.e.,  $v_L^{**}(p_L) > v_L^*(p_L)$  implies that  $\tau^{**}(p_L, w_L) > \tau^*(p_L, w_L)$ , i.e., the distortion at the bottom gets more severe, whereas  $\tau^{**}(p_L, w_H) = \tau^*(p_L, w_H)$ , so that there is no distortion at the top. It is also shown that  $v_L^{**}(p_L) < v_L^*(p_L)$  implies that  $\tau^{**}(p_L, w_L) < \tau^*(p_L, w_L)$  and that  $\tau^{**}(p_L, w_H) \leq \tau^*(p_L, w_H)$ .

The reason is as follows: if, starting from  $v_L^*(p_L)$ , redistribution in favor of the low-skilled is reduced, we eventually reach a region of the Pareto-frontier where no incentive constraint is binding and the implicit marginal tax rates of high- and low-skilled individuals are equal to 0;  $\tau^{**}(p_L, w_L) = 0$  and  $\tau^{**}(p_L, w_H) = 0$ . If we reduce  $v_L$  further, we get to a region where the low-skilled individuals' incentive constraint binds which implies no distortion at the bottom,  $\tau^{**}(p_L, w_L) = 0$ , and an upward distortion of labour supply for the high-skilled,  $\tau^{**}(p_L, w_H) < 0$ .

Combining these observations with Lemmas 4 and 5 proves the statements about implicit marginal tax rates in Proposition 5.

## A.6 Proof of Proposition 6

The proof of Proposition 6 is the exact mirror image of the proof of Proposition 6, and is therefore omitted.

## A.7 Proof of Proposition 7

We fix  $f_H$  at an arbitrary level so that we may suppress the dependence of  $s$  on  $f_H$ , and write simply  $s = (p_H, p_L)$ . Suppose that  $\frac{\bar{\theta}(p_H, p_L)}{\lambda} > \theta_H w_L$  so that the social choice function  $(q^{**}, c^{**}, y^{**})$  is determined as the solution to a collection of subproblems of the  $\mathcal{P}_L(p_H, f_H)$ -type. (The proof under the alternative assumption that  $\frac{\bar{\theta}(p_H, p_L)}{\lambda} < \theta_H w_L$  would follow from exactly the same arguments.) The proof makes use of the following Lemma:

**Lemma 6** *Suppose that  $\frac{\bar{\theta}(p_H, p_L)}{\lambda} > \theta_H w_L$ . Then*

$$\frac{\partial v_L^{**}(p_H, p_L)}{\partial p_H} = -\theta_H \frac{\partial q^{**}(p_H, p_L)}{\partial p_H}.$$

**Proof** Since  $\frac{\partial V(s, w_L, \theta_H)}{\partial p_L} \geq 0$  is binding for  $\frac{\bar{\theta}(p_H, p_L)}{\lambda} > \theta_H w_L$ , we have that, for a given  $p_H$ ,

$$\frac{\partial v_L^{**}(p_H, p_L)}{\partial p_L} = -\theta_H \frac{\partial q_L^{**}(p_H, p_L)}{\partial p_L}.$$

This implies that there exists a number  $\alpha(\kappa(p_H))$  so that, for every  $p_L \in [\kappa(p_H), 1]$ ,

$$\alpha(\kappa(p_H)) = \theta_H q^{**}(p_H, p_L) + v_L^{**}(p_H, p_L). \quad (56)$$

To prove the Lemma, we show that  $\alpha'(\alpha(\kappa(p_H)))\kappa'(p_H) = 0$ . For values of  $p_H$  so that  $0 \geq \eta(p_H)$ , this follows trivially from the observation that  $\kappa(p_H) = 0$  and hence also  $\kappa'(p_H) = 0$ . For values

of  $p_H$  so that  $0 < \eta(p_H)$ , it follows from Proposition 6 that  $\alpha(\kappa(p_H))$  is the utility level induced by an undistorted allocation. Hence,

$$\alpha(\kappa(p_H)) = \theta_H q^*(p_H, \kappa(p_H)) + u(c^*(p_H, \kappa(p_H), w_L)) - \frac{y^*(p_H, \kappa(p_H), w_L)}{w_L}.$$

Upon using Proposition 2, we find that

$$\alpha'(\kappa(p_H))\kappa'(p_H) = \left( \theta_H - \frac{1}{w_L} r'(q^*(p_H, \kappa(p_H))) \right) \frac{d}{dp_H} (q^*(p_H, \kappa(p_H))) \kappa'(p_H),$$

where

$$\theta_H - \frac{1}{w_L} r'(q^*(p_H, \kappa(p_H))) = \theta_H - \frac{1}{w_L} \frac{\bar{\theta}(p_H, \kappa(p_H))}{\lambda} = 0.$$

□

**Step 1: if  $\frac{\partial V(s, w_L, \theta_H)}{\partial p_L} \geq 0$  binds, then  $\frac{\partial V(s, w_H, \theta_L)}{\partial p_H} < 0$**

We seek to show that

$$\frac{\partial V(s, w_H, \theta_L)}{\partial p_H} = \theta_L \frac{\partial}{\partial p_H} q^{**}(p_H, p_L) + \frac{\partial}{\partial p_H} v_H^{**}(p_H, p_L) < 0.$$

From Step 1 in the proof of Proposition 6, we know that

$$v_H^{**}(p_H, p_L) = V_H(v_L^{**}(p_H, p_L), r(q^{**}(p_H, p_L))).$$

Hence, we seek to show that

$$\theta_L \frac{\partial}{\partial p_H} q^{**}(p_H, p_L) + V_{H1} \frac{\partial}{\partial p_H} v_L^{**}(p_H, p_L) + V_{H2} r' \frac{\partial}{\partial p_H} q^{**}(p_H, p_L) < 0,$$

or, equivalently, by Lemma 6, that

$$(\theta_L - \theta_H V_{H1} + V_{H2} r') \frac{\partial}{\partial p_H} q^{**}(p_H, p_L) < 0.$$

A straightforward adaptation of the arguments in the proof of Lemma 2 reveals that

$$\frac{\partial}{\partial p_H} q^{**}(p_H, p_L) > 0.$$

To complete the argument, it therefore remains to be shown that

$$\theta_L - \theta_H V_{H1} + V_{H2} r' < 0. \tag{57}$$

To see that this is true note that optimality condition (53) in the proof of Proposition 6 implies that

$$-\theta_H V_{H1} + V_{H2} r' = \frac{1 - f_H}{f_H} \theta_H - \frac{\bar{\theta}(p_H, p_L)}{f_H}.$$

Consequently, (57) holds if and only if

$$\theta_L f_H + \theta_H (1 - f_H) < \bar{\theta}(p_H, p_L). \tag{58}$$

To see that this is always fulfilled, recall that, by assumption,

$$\bar{\theta}(p_H, p_L) > \theta_H w_L \lambda, \quad (59)$$

in the interior of the region where the subproblems of the  $\mathcal{P}_L(p_H, f_H)$ -type determine  $(q^{**}, c^{**}, y^{**})$ . Further, upon using that  $\lambda = \frac{f_H}{w_H} + \frac{1-f_H}{w_L}$ ,  $w_H = \theta_H$  and  $w_L = \theta_L$ , it is straightforward to verify that

$$\theta_H w_L \lambda = \theta_L f_H + \theta_H (1 - f_H). \quad (60)$$

Hence, (60) and (59) imply that (58) is true.

**Step 2: if  $\frac{\partial V(s, w_L, \theta_H)}{\partial p_L} \geq 0$  binds, then  $\frac{\partial V(s, w_H, \theta_H)}{\partial p_H} > 0$**

We seek to show that

$$\frac{\partial V(s, w_H, \theta_H)}{\partial p_H} = \theta_L \frac{\partial}{\partial p_H} q^{**}(p_H, p_L) + \frac{\partial}{\partial p_H} v_H^{**}(p_H, p_L) > 0.$$

From Step 1 in the proof of Proposition 6 we know that

$$v_H^{**}(p_H, p_L) = V_H(v_L^{**}(p_H, p_L), r(q^{**}(p_H, p_L))).$$

Hence, we seek to show that

$$\theta_H \frac{\partial}{\partial p_H} q^{**}(p_H, p_L) + V_{H1} \frac{\partial}{\partial p_H} v_L^{**}(p_H, p_L) + V_{H2} r' \frac{\partial}{\partial p_H} q^{**}(p_H, p_L) > 0,$$

or, equivalently, by Lemma 6, that

$$(\theta_L - \theta_H V_{H1} + V_{H2} r') \frac{\partial}{\partial p_H} q^{**}(p_H, p_L) > 0.$$

Since  $\frac{\partial}{\partial p_H} q^{**}(p_H, p_L) > 0$ , it therefore remains to be shown that

$$\theta_H - \theta_H V_{H1} + V_{H2} r' > 0. \quad (61)$$

To see that this is true, note that optimality condition (53) in the proof of Proposition 6 implies that

$$-\theta_H V_{H1} + V_{H2} r' = \frac{1 - f_H}{f_H} \theta_H - \frac{\bar{\theta}(p_H, p_L)}{f_H}.$$

Consequently, (61) holds if and only if

$$\theta_H f_H + \theta_H (1 - f_H) > \bar{\theta}(p_H, p_L).$$

This inequality is obviously fulfilled for all  $(p_H, p_L) \neq (1, 1)$ .

**Step 3: if  $\frac{\partial V(s, w_L, \theta_H)}{\partial p_L} \geq 0$  binds, then  $\frac{\partial V(s, w_L, \theta_L)}{\partial p_L} < 0$**

We seek to show that

$$\frac{\partial V(s, w_L, \theta_L)}{\partial p_L} = \theta_L \frac{\partial}{\partial p_L} q^{**}(p_H, p_L) + \frac{\partial}{\partial p_L} v_L^{**}(p_H, p_L) < 0.$$

This follows from  $\frac{\partial}{\partial p_L} q^{**}(p_H, p_L) > 0$ , the fact that  $\theta_L < \theta_H$  and that  $\frac{\partial V(s, w_L, \theta_H)}{\partial p_L} \geq 0$  binds, so that

$$\theta_H \frac{\partial}{\partial p_L} q^{**}(p_H, p_L) + \frac{\partial}{\partial p_L} v_L^{**}(p_H, p_L) = 0.$$