Optimal tax and expenditure policy with aggregate uncertainty

Felix Bierbrauer[*]

University of Cologne, Germany

January 10, 2013

Abstract

We study optimal income taxation and public-goods provision under the assumption that the cross-section distributions of productive abilities or public-goods preferences are not known a priori. A conventional Mirrleesian treatment is shown to provoke manipulations of the policy mechanism by individuals with similar interests. The analysis therefore incorporates a requirement of coalition-proofness. The main result is that increased public-goods provision is associated with a more distortionary and a more redistributive tax system. With a conventional Mirrleesian treatment, the level of public-goods provision is not related to how distortionary or redistributive the tax system is.

Keywords: Optimal Taxation, Public goods, Mechanism Design.

JEL: C72; D72; H21.

[*] I am indebted to Martin Hellwig. This paper builds on our joint work on “Public-Good Provision in a Large Economy.” I am grateful for the comments of an anonymous referee. I also benefited from conversations with Sophie Bade, Martin Barbie, Marco Bassetto, Ted Bergström, Soren Blomquist, Pierre Boyer, Georges Casamatta, Christoph Engel, Mike Golosov, Hans Peter Grüner, Stephan Lauermann, Marco Sahm, Joel Slemrod, Christian Traxler, Aleh Tsyvinski, Georg von Heusinger and John Weymark.

1 Introduction

This paper looks at a classical problem in normative public economics: What are the characteristics of an optimal policy, consisting of a redistributive income tax schedule and expenditures on public goods.

We use a very simple model that differs, however, from the previous literature in one respect. It includes a problem of information aggregation, i.e., it is not known a priori how productive the economy's workforce is or what the demand for public goods looks like. This uncertainty about the data of the economy matters in the following sense: If we used a conventional approach to the policy problem – that is, an optimal income tax in conjunction with a version of the Samuelson rule that takes the marginal cost of public funds into account – then there would be scope for manipulations of the policy mechanism by groups of like-minded individuals.

Our analysis has two parts. In the first part, we develop a mechanism design approach that includes a set of incentive constraints which render such manipulations unattractive. In the second part, we characterize an optimal policy that satisfies these constraints. Our main results are as follows: First, uncertainty about the cross-section distribution of productive abilities does not generate an incentive problem. In this case, the optimal Mirrleesian policy is manipulation-proof. Second, uncertainty about the demand for public goods requires major deviations from the optimal Mirrleesian policy: If we think of, say, the United States and Sweden, as two countries which differ only in the demand for public goods, with a comparatively high demand for public goods in Sweden and a rather low demand for public goods in the United States, then a conventional Mirrleesian approach would reach the conclusion that the Swedes should have more public goods, but that the two countries should otherwise have identical tax and transfer systems. The optimal manipulation-proof policy, by contrast, implies, that the Swedes do not only have more public goods, but also a more redistributive and distortionary tax and transfer system. Moreover, relative to an optimal Mirrleesian policy, taxes, income transfers, and expenditures are distorted upwards in Sweden and distorted downwards in the US.

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.[1] 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 individuals, the fraction of high-skilled individuals with

[1] The paper thus contributes to 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), Kocherlakota (2005), or Bassetto and Phelan (2008). Predecessors are Hammond (1979) and Guesnerie (1995).

a high valuation of public goods, and the fraction of low-skilled individuals with a high valuation of 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 changes in the cross-section distribution of preferences or productive abilities. The aim of the paper is to characterize the *optimal* mechanism, i.e., the optimal response to changes in the cross-section distribution of preferences or productive abilities.

Our mechanism design approach invokes a requirement of robustness with respect to the specification of the individuals' probabilistic beliefs.[2] As has been shown in Bierbrauer (2009a), robust mechanism design in a large economy yields a framework that is equivalent to a Mirrleesian model of income taxation and public-goods provision.[3] The 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 the property that individuals have an incentive to lie about their public-goods preferences. Either the high-skilled individuals have an incentive to exaggerate their public-goods preferences or the low-skilled individuals 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.[4]

This problem arises because 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

[2] 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.

[3] To the best of my knowledge, an analysis of a Mirrleesian economy with aggregate uncertainty based on the solution concept of an interim Nash equilibrium has not yet been undertaken.

[4] This may seem counterintuitive because, in the political discourse, its seems that those who represent the low-skilled push for a larger public sector. However, often this is done for redistributive reasons and concerns the public provision of private goods, such as health care or education as opposed to the public provision of public goods, such as infrastructure or national defense.

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 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.[5]

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 or 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.[6]

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 the following main results. First, there is a fundamental difference between the implications of uncertainty about the cross-section distribution of preferences for public goods on the one hand, and uncertainty about the cross-section distribution of productive abilities on the other: while the assumption that the ability distribution is a priori unknown has no bearing on the set of admissible policies, the assumption of an unknown distribution of public-goods preferences 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

[5] 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.

[6] Several companion papers have also looked at this problem, albeit in more specific setups and with different solution concepts. Bierbrauer and Sahm (2010) focus on the aggregation of public-goods preferences via voting procedures. In addition, it is assumed that the cross-section distribution of public-goods preferences among the high-skilled and the low-skilled individuals is identical. Bierbrauer (2009b) also studies optimal non-linear income taxation and public-goods provision, but does not contain a rigorous foundation in terms of mechanism design theory, and, moreover, severely restricts the set of possible states of the economy. Bierbrauer (2011) looks at the communication of public-goods preferences under the assumption that a Ramsey tax system is used to finance public-goods provision.

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, and compare it to the optimal Mirrleesian policy. The main difference between these policies is how an increased demand for public goods affects the structure of the income tax system. If the number of people who value public goods highly goes up, then, under the Mirrleesian policy, the public-goods provision level goes up, but the tax policy remains unaffected. With coalition-proofness, by contrast, an increase in the demand for public goods goes together with an increase of direct income transfers from rich to poor individuals, and with an increase of marginal income tax rates. The intuition for these results is as follows: Suppose the Mirrleesian policy faces the problem that low-skilled people do not admit a high public-goods preference because they suffer too much from having to pay for the public good. Now, to fix this problem, an increase in the public-goods provision level has to be made more appealing to the low-skilled. This is achieved by promising them additional redistribution whenever the public-goods provision level goes up. To generate more utility for the low-skilled, direct income transfers have to be increased. This in turn makes it necessary to have a more distortionary income tax system.

Related literature. It has long been recognized in public economic theory that a society that wants to provide public goods and redistribute income 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. However, there is no unified treatment of these problems.

The theory of optimal taxation in the tradition of Mirrlees (1971) focuses on the problem of taxing 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.[7]

The theory of public-goods provision in the tradition of Clarke (1971) and Groves (1973), by contrast, focuses on problems of information aggregation. In this literature,

[7] See, for example, Boadway and Keen (1993), Gahvari (2006), or Kreiner and Verdelin (2010).

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 does this imply for public-goods provision? Should public spending expand if the demand for public goods goes up? If so, what are the implications for the shape of the tax system?

The remainder of the paper is organized as follows: Section 2 describes the economic environment. As a benchmark, Section 3 reviews the Mirrleesian approach to income taxation and public goods provision and relates it to a model of robust mechanism design. In Section 4 it is shown that the Mirrleesian outcome provokes collective deviations. In Section 5, we introduce the solution concept of a robust and coalition-proof interim Nash equilibrium. The optimal robust and coalition-proof mechanism for income taxation and public-goods provision is characterized in Section 6. The last section contains concluding remarks. All proofs are in the Appendix. The formal analysis repeatedly refers to results that can be found in Bierbrauer and Boyer (2010), a note which contains, for a simple model of non-linear income taxation, a complete analytical characterization of all Pareto-efficient tax schedules.

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

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

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 preference parameter θ^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. Moreover, it satisfies the Inada conditions so that $\lim_{c \to 0} u'(c) = \infty$ and $\lim_{c \to \infty} u'(c) = 0$. The disutility from productive effort depends on a skill parameter ω^i, which, again, takes either a high or a low value; for all i, $\omega^i \in \Omega = \{\omega_L, \omega_H\}$, where $0 < \omega_L < \omega_H$. Individuals are privately informed about their public-goods preference and about their skill level. To simplify the exposition, we assume that $\theta_L = \omega_L$ and that $\theta_H = \omega_H$.

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 high productivity, p_H is the fraction of high-skilled individuals with a high valuation of the public good, and p_L is the fraction of low-skilled individuals with a high valuation of the public good. The set of states is in the following denoted by $S = [0,1]^3$.

Aggregate uncertainty arises if the state of the economy is a priori unknown. In the following, we will occasionally limit the analysis to a subset of S in order to disentangle the implications of aggregate uncertainty that arises because the skill distribution is unknown from the implications of aggregate uncertainty that arises because the demand for public goods is unknown.[8] For instance, to isolate the implications of an unknown distribution of productivity parameters, it is convenient to assume that p_L and p_H are known quantities, so that different states are only distinguished by the population share of high-skilled individuals. Similarly, if we want to focus on the implications of uncertainty about the distribution of public-goods preferences among the low-skilled, we will treat p_H and f_H as given parameters and p_L as unknown, etc. Finally, we restrict ourselves to states with the property that $\frac{\omega_L}{\omega_H} \geq f_H$. This assumption simplifies the exposition. It implies that, for the optimization problems studied below, non-negativity constraints on consumption levels can be safely ignored.[9]

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 $c : S \times \Omega \times \Theta \mapsto \mathbb{R}_+$, and the output requirement is determined by $y : S \times \Omega \times \Theta \mapsto \mathbb{R}_+$.

The resource requirement of public-goods provision is captured by an increasing and convex cost function r which satisfies $\lim_{q \to 0} r'(q) = 0$ and $\lim_{q \to \infty} r'(q) = \infty$. A social

[8] Specifically, we will proceed in this way in Section 6 which deals with the characterization of optimal social choice functions. For the results in earlier sections, the set of states is taken to be $[0,1]^3$.

[9] For the given environment, Bierbrauer and Boyer (2010) provide a complete analytical characterization of the set of Pareto-efficient income tax schedules. It follows from their analysis – particularly, from arguments in the proof of Proposition 1 – that non-negativity constraints on consumption levels can be safely ignored, for every Pareto-efficient income tax schedule, if $\frac{\omega_L}{\omega_H} \geq f_H$.

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

$$\begin{aligned}
&f_H\Big(p_H(y(s,\omega_H,\theta_H)-c(s,\omega_H,\theta_H))+(1-p_H)(y(s,\omega_H,\theta_L)-c(s,\omega_H,\theta_L))\Big)\\
&+(1-f_H)\Big(p_L(y(s,\omega_L,\theta_H)-c(s,\omega_L,\theta_H))+(1-p_L)(y(s,\omega_L,\theta_L)-c(s,\omega_L,\theta_L))\Big)\\
&\geq r(q(s))\ .
\end{aligned} \tag{1}$$

Types and beliefs. The analysis below focuses 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.* Intuitively, an agent views the cross-section distributions of payoff types, as well as the beliefs of other agents about the cross-section distribution of payoff types (and also their higher order beliefs) as a random quantity. The agent's belief type is identified with a probability distribution of this random quantity.

Formally, a type space consists of a measurable space $(T,\mathcal{T})$, a measurable and surjective map π from T into $\Omega\times\Theta$, and a measurable map β from T into the space $\Delta(\Delta(T))$ of probability distributions over measures on T, denoted by β. The interpretation is as follows: There is a set of abstract types T. Now suppose that agent i has some abstract type $t^i\in T$. The function π determines the preference and the productivity parameter that enter agent i's payoff function. We therefore refer to $\pi(t^i)\in\Omega\times\Theta$ also as agent i's payoff type. The belief type $\beta(t^i)$ indicates the agent's probabilistic beliefs about the cross-section distribution of abstract types. Let $\Delta(T)$ be the set of possible cross-section distributions of types. Then, for any $X\subset\Delta(T)$, $\beta(X\mid t^i)$ is the probability that type t^i of agent i assigns to the event that the cross-section distribution of types belongs to the subset X of $\Delta(T)$.[10] We refer to the map $\beta: T\rightarrow\Delta(\Delta(T))$ 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 $\delta\in\Delta(T)$ 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, type t^i of agent i cannot rule out any event that has positive probability from the perspective of some other type $\bar{t}^i\neq t^i$.[11] A particular example of a moderately uninformative belief system is

[10] Observe that the set of possible belief types is the set of probability distribution over the set of possible cross-section distributions of types, $\Delta(\Delta(T))$.

[11] The absolute continuity assumption does not presuppose a common prior. For a discussion of moderately uninformative belief systems under a common prior assumption, see Bierbrauer and Hellwig (2010).

a "complete information type space", where all individuals "know" (assign probability 1 to) the true state and assign probability 0 to all other states.

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.[12] The function $Q : \Delta(A) \to \mathbb{R}_+$ gives the public-goods provision level as a function of the cross-sectional distribution of actions, and the functions $C : \Delta(A) \times A \to \mathbb{R}_+$ and $Y : \Delta(A) \times A \to \mathbb{R}_+$ 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. (Throughout, the capital letters Q, C, and Y refer to the outcome functions of a mechanism, and the small letters q, c and y refer to the different components of a social choice function.)

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 interim 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 an interim 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 $\pi : T \to \Omega \times \Theta$, there exists a mechanism that implements it on the type space $[(T, \mathcal{T}), \pi, \beta]$, for every moderately uninformative belief system β.[13]

[12] We do not (yet) restrict attention to direct mechanism and to truthtelling equilibria because, for the coalition-proof interim Nash equilibria that will be studied below, the revelation principle does not generally hold.

[13] 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.

3 A Mirrleesian approach

As a benchmark, we characterize the social choice functions that are robustly implementable as an interim 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. We will then demonstrate in subsequent sections that the optimal Mirrleesian policy provokes manipulations of the policy outcome by like-minded individuals.

Proposition 1 *The following statements are equivalent.*

(a) A social choice function (q, c, y) is robustly implementable as an interim Nash equilibrium.

(b) A social choice function (q, c, y) satisfies the following individual incentive compatibility constraints: For every $s \in S$, every $(\omega, \theta) \in \Omega \times \Theta$, and every $(\hat{\omega}, \hat{\theta}) \in \Omega \times \Theta$,

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

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 $(\omega, \theta) \in \Omega \times \Theta$,

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

for all $(\hat{\omega}, \hat{\theta}) \in \Omega \times \Theta$. 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 ω and every pair θ and $\hat{\theta}$,

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

so that two individuals who differ only in their valuation of public goods 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 $c(s, \omega, \theta) = c(s, \omega, \hat{\theta})$ and $y(s, \omega, \theta) =$

$y(s, \omega, \hat{\theta})$, for every s, ω, and every pair $(\theta, \hat{\theta})$.[14] In the following, we may hence drop the dependence of consumption levels and output requirements on public-goods preferences and write simply $c(s, \omega)$ and $y(s, \omega)$, respectively. With this notation, we can write the individual incentive compatibility constraints as follows: for every s, every ω, and every $\hat{\omega}$,

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

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

$$f_H(y(s,\omega_H) - c(s,\omega_H)) + (1 - f_H)(y(s,\omega_L) - c(s,\omega_L)) \geq r(q(s)) \ . \tag{6}$$

Social choice functions and income tax schedules. 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 $\tau : y \mapsto \tau(y)$ 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.[15] 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.

Total tax payments and marginal income tax rates in state s are then implicitly defined in the following way: The difference between an individual's contribution to the economy's output and his private goods consumption is the individual's tax payment. Hence, in state s, the tax payment of a type k-individual is given by $y(s, \omega_k) - c(s, \omega_k)$. We will also be interested in the tax payments net of the revenue requirement $r(q(s))$, which are defined by

$$n_k(s) := y(s,\omega_k) - c(s,\omega_k) - r(q(s)) \ .$$

The net tax payments are a measure of how *redistributive* the income tax system is.[16] Marginal income tax rates, by contrast, are a measure of how *distortionary* the tax system is. They are defined as the difference between an individual's marginal rate of

[14] 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 $y(s, \omega, \theta) - c(s, \omega, \theta) = y(s, \omega, \hat{\theta}) - c(s, \omega, \hat{\theta})$. This equality in conjunction with the fact that indifference curves in a $y - c$ diagram are strictly increasing and strictly convex, yields $c(s, \omega, \theta) = c(s, \omega, \hat{\theta})$ and $y(s, \omega, \theta) = y(s, \omega, \hat{\theta})$.

[15] Examples are Stiglitz (1982), Boadway and Keen (1993), Gahvari (2006), or Hellwig (2007).

[16] To see this, note that the resource constraints can be written as follows: For every $s = (f_H, p_H, p_L)$ it has to be the case that $f_H n_H(s) + (1 - f_H) n_L(s) \geq 0$. With an optimal policy this constraint is binding for every s, so that $n_H(s)$ can be interpreted as the direct income transfer that each high-skilled individual has to finance, and $n_L(s)$ is the direct income transfer that each low-skilled individual receives.

transformation between output y and consumption c, which equals 1 for each individual, and the individual's marginal rate of substitution, $\frac{1}{\omega u'(c)}$.[17] Hence, in state s, the marginal tax rate for a type k-individual is given by

$$\tau'_k(s) := 1 - \frac{1}{\omega_k u'(c(s,\omega_k))} \; .$$

The optimal Mirrleesian policy. An optimal utilitarian social choice function solves the following maximization problem: choose $q : S \to \mathbb{R}_+$, $c : S \times \Omega \to \mathbb{R}_+$ and $y : S \times \Omega \to \mathbb{R}_+$ in order to maximize expected utilitarian welfare $E[W(s)]$, where $W(s)$ is utilitarian welfare in state s, subject to the the constraints in (6) and (5).[18] 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, $q(s)$, $c(s,\omega_L)$, $y(s,\omega_L)$, $c(s,\omega_H)$ and $y(s,\omega_H)$ are chosen in order to maximize

$$W(s) \;\; = \;\; \bar{\theta}(s)q(s) + f_H\Big(u(c(s,\omega_H)) - \tfrac{y(s,\omega_H)}{\omega_H}\Big) + (1-f_H)\Big(u(c(s,\omega_L)) - \tfrac{y(s,\omega_L)}{\omega_L}\Big) \; ,$$

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 preference for public-goods provision in state s. As is well-known,[19] the solution to this problem is such that the incentive constraint for the high-skilled individuals is binding,

$$u(c(s,\omega_H)) - \frac{y(s,\omega_H)}{\omega_H} = u(c(s,\omega_L)) - \frac{y(s,\omega_L)}{\omega_H} \; , \tag{7}$$

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

$$u(c(s,\omega_L)) - \frac{y(s,\omega_L)}{\omega_L} > u(c(s,\omega_H)) - \frac{y(s,\omega_H)}{\omega_L} \; .$$

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 $y(s,\omega_L)$ and increasing $y(s,\omega_H)$, so that aggregate output remains unchanged. The resource constraint is also binding,

$$f_H(y(s,\omega_H) - c(s,\omega_H)) + (1-f_H)(y(s,\omega_L) - c(s,\omega_L)) = r(q(s)) \; . \tag{8}$$

[17] This is based on the first-order condition of the utility maximization problem that individuals face when confronted with an income tax schedule τ: choose c and y in order to maximize $u(c) - \frac{y}{\omega}$ subject to the constraint $c = y - \tau(y)$. The first order condition is $\tau'(y) = 1 - \frac{1}{\omega u'(c)}$.

[18] Expectations are taken with respect to the policy maker's subjective beliefs.

[19] A formal proof can be found in Weymark (1986) or Hellwig (2007).

Otherwise $y(s,\omega_L)$ and $y(s,\omega_H)$ 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 $q(s)$, $c(s,\omega_L)$, $y(s,\omega_L)$, $c(s,\omega_H)$ and $y(s,\omega_H)$. The results from this exercise are summarized in the following proposition, which we state without proof.

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

i) The optimal consumption levels satisfy

$$u'(c^*(s,\omega_H)) = \frac{1}{\omega_H} \quad \textit{and} \quad u'(c^*(s,\omega_L)) = \frac{1}{\omega_L}\frac{1-f_H\frac{\omega_H-\omega_L}{\omega_H}}{1-f_H\frac{\omega_H-\omega_L}{\omega_L}} \ .$$

ii) Let $\lambda(s) := \frac{f_H}{\omega_H} + \frac{1-f_H}{\omega_H}$.[20] *The optimal public-goods provision level satisfies the Samuelson rule, $\bar{\theta}(s) = \lambda(s)r'(q^*(s))$.*

iii) The optimal output requirements satisfy

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

where $e^*(s) := f_Hc^*(s,\omega_H) + (1-f_H)c^*(s,\omega_L) + r(q^*(s))$ 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. Moreover, the simple characterization of the optimal public-goods provision level is possible because an individual's willingness to work harder in exchange for increased private goods consumption does not depend on the supply of public goods. As shown in Boadway and Keen (1993), this rules out any reason to deviate from public-goods provision according to the Samuelson rule. With non-separable preferences such deviations could be attractive as a means of relaxing the burden of binding incentive constraints.

Figure 1 illustrates such a comparative statics exercise.[21] The population share of the high-skilled is fixed at $f_H = \frac{1}{2}$, and the fraction of the high-skilled with a high-valuation of the public good is also fixed at $p_H = \frac{1}{2}$. The curves in the figures show, respectively,

[20] It can be shown that marginal cost of public funds in state s is given by $\lambda(s)$.

[21] The figures are based specific functional forms, namely $u(c) = \sqrt{c}$ and $r(q) = \frac{1}{2}q^2$.

what happens with the low-skilled individuals' marginal income tax rate $\tau_L'(s)$, their net tax payment $n_L(s)$, and the public-goods provision level q as the share of the low-skilled individuals with a high valuation of the public good is continuously increased from $p_L = 0$ to $p_L = 1$. The figure illustrates that the public-goods provision level goes up as the number of low-skilled individuals who value the public good highly increases. In addition, the income tax system is unaffected by such a preference shock. Neither the marginal tax rate of the low-skilled individuals, nor their transfer income changes as their demand for public goods goes up.[22]

Insert Figure 1 here

Figure 1 documents the results of a particular comparative statics exercise, namely a change in the cross-section distribution of preferences for the public good among the low-skilled. A similar exercise can be undertaken with respect to changes of the cross-section distribution of the productivity parameter, i.e., a change of f_H, while p_L and p_H remain constant. Figure 2 documents the results of such an exercise. It shows how the optimal policy changes as the population share of the high-skilled individuals is increased from $f_H = \frac{1}{3}$ to $f_H = \frac{2}{3}$. It is assumed throughout that $p_L = \frac{3}{4}$ and that $p_H = \frac{1}{2}$.

Insert Figure 2 here

The curves show how, under the optimal Mirrleesian policy, characterized in Proposition 2, the public-goods provision level q, the low-skilled individuals' marginal income tax rate $\tau_L'(s)$, and their net tax payment $n_L(s)$ respond to such a productivity shock. As the average worker becomes more productive, the public goods provision level goes up and the tax system becomes more distortionary as reflected by the increase of the low-skilled individuals' marginal income tax rates. Proposition 2 implies that the high-skilled individuals' marginal income tax rate stays constant at a level of zero. Also, the net transfer that a low-skilled individual receives goes up. In this sense, the tax system becomes also more redistributive. It can also be shown that each high-skilled individual's net tax payment goes down. Hence, the reduced number of low-skilled individuals dominates the effect that each low-skilled individual receives a higher transfer.

Implementation. So far, we have assumed that an abstract mechanism $M = [(A, \mathcal{A}), Q, C, Y]$ is used to implement a social choice function (q, c, y). In a conventional Mirrleesian model of income taxation and pubic-goods provision, the state s of the economy is known, and the implementation of the optimal allocation via some income tax schedule T_s is straightforward. This observation extends to the present setting in which s is a priori unknown.

[22] Proposition 2 and the definition of marginal income taxes imply that the marginal income tax rate for the high-skilled is zero, for every s. This property is often referred to as *no distortion at the top.* Hence, also the high-skilled individuals' marginal tax rate remains unaffected.

However, individuals first have to send messages that enable the mechanism designer to learn what the state of the economy is, and what income tax schedule from the set of income tax schedules $\{T_s\}_{s \in S}$ has to be used.

4 Robust social choice functions are manipulable

In the following, we discuss two examples in order to question that the Mirrleesian incentive compatibility constraints in (5) suffice to make sure that a social choice function can be implemented. Specifically, we will argue that with such social choice functions individuals may have an incentive to coordinate their behavior in such a way that the optimal policy is manipulated.

Example 1: Public-goods preferences. With an optimal Mirrleesian policy as characterized in Proposition 2, an increase of the revenue requirement r by some $\varepsilon > 0$ implies that all individuals in the economy have to increase their output by ε, while their consumption levels remain unaffected. For a low-skilled individual, this implies a utility loss of $\frac{\varepsilon}{\omega_L}$, whereas the utility loss for the high-skilled individual equals only $\frac{\varepsilon}{\omega_H}$. 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 $V^* : S \times \Omega \times \Theta \to \mathbb{R}$, with

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

where (q^*, c^*, y^*) is the social choice function characterized in Proposition 2.

Now consider a low-skilled individual with a high valuation of public goods; i.e., $\omega = \omega_L$ and $\theta = \theta_H$. Recall that any state s can be written as $s = (f_H, p_H, p_L)$. One easily derives that,

$$\begin{aligned} \frac{\partial V^*(f_H,p_H,p_L,\omega_L,\theta_H)}{\partial p_L} &= \left(\theta_H \omega_L - r'(q^*(f_H, p_H, p_L))\right) \frac{1}{\omega_L} \frac{\partial q^*(f_H,p_H,p_L)}{\partial p_L} \\ &= \left(\theta_H \omega_L - \frac{\bar{\theta}(f_H,p_H,p_L)}{\lambda(f_H,p_H,p_L)}\right) \frac{1}{\omega_L} \frac{\partial q^*(f_H,p_H,p_L)}{\partial p_L} . \end{aligned} \tag{9}$$

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 also have a high valuation of public goods. This implies that $\bar{\theta}(f_H, p_H, p_L)$ is close to θ_H so that $\frac{\partial V^*(f_H,p_H,p_L,\omega_L,\theta_H)}{\partial p_L}$ is close to

$$\theta_H \left(\omega_L - \frac{1}{\lambda(f_H, p_H, p_L)}\right) \frac{1}{\omega_L} \frac{\partial q^*(f_H, p_H, p_L)}{\partial p_L} .$$

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

$$\frac{\partial V^*(f_H,p_H,p_L,\omega_L,\theta_H)}{\partial p_L} < 0 \ .$$

This situation is illustrated in Figure 3. 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 public-goods preference, p_L. The indirect utility function of these individuals 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 valuation of public goods, the mechanism designer would perceive p_L to be lower than it actually is and reduce the 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 public-goods preferences. Consequently, lying is a best response.

Insert Figure 3 here

Example 2: Productive Abilities. We can also question whether it is generally possible to acquire information on the fraction of high-skilled individuals, f_H. To demonstrate this, we 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 $z(s,\omega_L) := (c(s,\omega_L), y(s,\omega_L))$ and $z(s,\omega_H) := (c(s,\omega_H), y(s,\omega_H))$, and there is redistribution from the high-skilled to the low-skilled, $y(s,\omega_H) - c(s,\omega_H) > 0$ and $y(s,\omega_L) - c(s,\omega_L) < 0$. 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 4. 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.

Insert Figure 4 here

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 optimal Mirrleesian policy, characterized 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 valuation of public goods in Example 1 – 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.

5 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. We will then introduce the requirement of robustness, and provide a characterization of robust and coalition-proof social choice functions.

5.1 Coalition-proof interim Nash equilibrium

Consider an individual with payoff type (ω, θ) and a mechanism $M = [(A, \mathcal{A}), Q, C, Y]$. If this individual chooses an action a and if the cross-section distribution of actions equals α, then the resulting payoff for this individual is given by

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

To be able to state the individual's *expected* payoff in a concise way, we introduce some notation: A (mixed) strategy in the game induced by M is a function $\sigma : T \to \Delta(A)$ 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 an individual's type being equal to t, that $a(t)$ takes values in subset A' of A is in the following denoted by $\sigma(A' \mid t)$.

Now, if almost all individuals follow a strategy σ and the cross-section distribution of types is given by δ, this induces a cross-section distribution of actions that is given by $\alpha(\delta, \sigma)$. We assume that a law of large numbers for large economies holds, so that we can interpret $\sigma(A' \mid t)$ 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'.[23] Consequently, for a given δ, we can treat $\alpha(\delta, \sigma)$ as a non-random quantity.

The expected payoff of a type t individual from behaving according to a mixed strategy $\chi \in \Delta(A)$, if almost all other individuals behave according to σ, can now be written as

$$\tilde{U}_M(\sigma, \chi, t) := \int_{\Delta(T)} \int_A \tilde{u}_M(\alpha(\delta, \sigma), a, \omega(t), \theta(t)) \; d\chi(a) \; d\beta(\delta \mid t) \; .$$

Definition 1 *Given a mechanism M and a type space $[(T, \mathcal{T}), \pi, \beta]$, a strategy $\sigma^* : T \to \Delta(A)$ is said to be a coalition-proof interim Nash equilibrium if it is an interim Nash equilibrium, and there is no set of types $T' \subseteq T$ who can deviate to a strategy $\sigma'_{T'} : T' \to \Delta(A)$ 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 σ^* to types not in T', is an interim 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 $t \in T'$,

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

(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 $\omega(t') \neq \omega(t'')$, or $\theta(t') \neq \theta(t'')$, or $\beta(t') \neq \beta(t'')$ – with a strategy $\sigma''_{T''} : T'' \to \Delta(A)$, so that $(\sigma^*_{T \setminus T'}, \sigma'_{T' \setminus T''}, \sigma''_{T''})$ is an interim Nash equilibrium, and, for all $t \in T''$,

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

An equilibrium σ^* 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

[23] For a discussion of the law of large numbers in large economies, see Sun (2006), Al-Najjar (2004) or Judd (1985).

consistency requirements that such a deviation from an equilibrium strategy σ^* has to satisfy: the behavior that is prescribed by the deviation must induce a new interim 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.

The definition 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 ease of exposition, we do not model a possibly infinite chain of successive formations of subcoalitions. The reason that such a more elaborate treatment would yield the same conclusion is as follows: Coalition-formation takes place at the level of types. That is, a coalition is a subset T' of the set of types T, with the understanding that all individuals with types in that subset form a coalition. Put differently, a coalition is a collection of individuals with particular preferences and beliefs. With this approach, a subcoalition of some initial coalition is a subset T'' of some set of types $T' \subset T$, and a subsubcoalition would be a subset $T''' \subset T''$, etc. Now, if we allowed for a possibly infinite chain of successive formations of subcoalitions, this would make it more difficult to form a successful coalition and therefore make the set of implementable social choice functions weakly larger. The question then is whether it would make the set also strictly larger. For the given environment the answer is "no". The reason is that the proofs of Propositions 3 (necessary conditions for robustness and coalition-proofness) and 5 (sufficient conditions) below are based on the construction of coalitions and subcoalitions of minimal size, namely coalitions of individuals who all have the same payoff and belief types and which therefore possess no further subcoalitions. A requirement that these coalition and subcoalitions must themselves be coalition-proof would therefore have no bite.

We can think of the collective deviation as resulting from an own mechanism design problem that the deviating agents face. Condition (a) can be interpreted as an incentive compatibility constraint, so that behaving according to 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 Che and Kim (2006). It 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. The approach taken here is different in that we define a coalition-proof equilibrium for a game in normal form. This simplifies the exposition, without affecting the set of implementable social choice functions.[24]

We do not consider the possibility that members of a coalition may communicate with each other so as to acquire more precise information on the state of the economy, and to adjust their communication with the mechanism designer accordingly. For a given type space, allowing for such communication would tend to make coalition formation easier and therefore to make the set of implementable social choice functions smaller. However, below we will focus on social choice functions that are both robust and coalition-proof. This implies, in particular, that coalition-proofness has to hold on type spaces where all individuals "know" the state of the economy so that there is no need of such communication. It can be shown that the necessary condition of coalition-proofness on all such "complete information type spaces" is the key element in the characterization of coalition-proofness on all type spaces. One can therefore show that the set of robust and coalition-proof social choice functions would not shrink if we allowed for communication among deviators.[25]

5.2 Robust and coalition-proof implementation

Given a type space $[(T, \mathcal{T}), \pi, \beta]$, a social choice function (q, c, y) is said to be implementable as a coalition-proof interim Nash equilibrium, if there is a mechanism M and a strategy σ^* such that (i) σ^* is a coalition-proof interim Nash equilibrium on this type space, and (ii) the equilibrium allocation coincides with the prescription of the social choice function for every δ; i.e., we have that, for every δ,

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

and, for each $\delta \in \Delta(T)$ and $t \in T$,

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

$\sigma^*(t)$-almost surely. Note that since individuals play mixed strategies, $a(t)$ and also $C(\alpha(\delta, \sigma^*), a(t))$ and $Y(\alpha(\delta, \sigma^*), a(t))$ are random quantities with probability distribution

[24] Bierbrauer and Hellwig (2010) also study a problem of mechanism design in a large economy model with private information and aggregate uncertainty. They model an extensive form game of coalition-formation under conditions of incomplete information in the following way: (i) an overall mechanism designer first proposes a mechanism, (ii) then agents may choose to communicate with the designer of a collusive side mechanism, (iii) the designer of this collusive mechanism then uses the information that is contained in the reports that he receives to give individuals recommendations on how to behave in the overall mechanism, (iv) individuals choose whether or not to follow the recommendation and send a report to the overall mechanism. Coalition-proofness then requires that no subset of types can benefit from using such a collusive side mechanism. The necessary and sufficient conditions for coalition-proofness and robustness of a social choice function that follow from this very detailed approach are equivalent to those that follow from coalition-proofness and robustness as defined here.

[25] Again, see Bierbrauer and Hellwig (2010) for an explicit modeling of a communication stage in an extensive form game of coalition formation.

$\sigma^*(t)$. Equation (13) requires that $C(\alpha(\delta, \sigma^*), a(t))$ and $Y(\alpha(\delta, \sigma^*), a(t))$ are with probability 1 equal to the values of c and y that are stipulated by the social choice function for a type t individual in state δ.

We say that a social choice function is robustly implementable and coalition-proof, if, given $(T, \mathcal{T})$ and π, there is a mechanism M and a strategy σ^* such that requirements (i) and (ii) are fulfilled, for every belief system β.

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 interim Nash equilibrium, the revelation principle does not hold. It shows a social choice function that can be implemented as a coalition-proof interim Nash equilibrium if a non-direct mechanism is used, but which cannot be implemented if a direct mechanism is used.[26]

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 truthtelling 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 β 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

[26] For the solution concept of an interim Nash equilibrium, the revelation principle applies, i.e. any social choice function that can be reached by *some* interim Nash equilibrium of *some* mechanism, can also be reached as a truthtelling equilibrium of a direct mechanism. It is known that the more ambitious task of *full* implementation – i.e., finding a mechanism so that *every* interim Nash equilibrium of that mechanism reaches the social choice function – may require the use of non-direct mechanism, see, e.g., Jackson (2001) for an overview. While coalition-proofness is not the same as full implementation, there is a similarity in that both postulate that there must not exist further equilibria with certain properties.

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 δ' 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 lower part of Figure 5, where δ'' indicates a distribution of actions so that less than half of the population chooses a_2.

Insert Figure 5 here

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 5 arise 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 shown in the right part of Figure 5. 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.

Necessary conditions for robustness and coalition-proofness

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

$$V(f_H, p_H, p_L, \omega, \theta) = \theta q(f_H, p_H, p_L) + u(c(f_H, p_H, p_L, \omega)) - \frac{y(f_H, p_H, p_L, \omega)}{\omega} \ .$$

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), $V(f_H, p_H, p_L, \omega_H, \theta_L)$ is a non-increasing function of p_H, and $V(f_H, p_H, p_L, \omega_H, \theta_H)$ is a non-decreasing function of p_H, and (ii) for any given pair (f_H, p_H), $V(f_H, p_H, p_L, \omega_L, \theta_L)$ is a non-increasing function of p_L, and $V(f_H, p_H, p_L, \omega_L, \theta_H)$ is a non-decreasing function of p_L.*

To obtain an intuitive understanding of these conditions it is instructive to relate them to the incentive constraints which are familiar from the literature on the elicitation of public-goods preferences, which is based on models with a finite number of agents. Suppose that

f_H and p_H are known quantities so that a state s of the economy can be identified with the share of low-skilled individuals with a high valuation of public goods, p_L. Suppose that p_L can only take two values denoted by $\underline{p}_L$ and $\bar{p}_L$ with $\bar{p}_L > \underline{p}_L$. Denote by

$$v_L(p_L) = u(c(p_L, \omega_L)) - \frac{y(p_L, \omega_L)}{\omega_L}$$

the utility that low-skilled individuals derive from the consumption of private goods and work effort in state p_L. The requirement that $V(f_H, p_H, p_L, \omega_L, \theta_L)$ is a non-increasing function of p_L implies that

$$\theta_L q(\underline{p}_L) + v_L(\underline{p}_L) \geq \theta_L q(\bar{p}_L) + v_L(\bar{p}_L) \ . \tag{14}$$

Likewise, the requirement that $V(s, \omega_L, \theta_H)$ is a non-decreasing function of p_L implies that

$$\theta_H q(\bar{p}_L) + v_L(\bar{p}_L) \geq \theta_H q(\underline{p}_L) + v_L(\underline{p}_L) \ . \tag{15}$$

These constraints admit the following interpretation: There is a representative low-skilled agent. It is unknown whether this representative agent has a high or a low valuation of public goods. If his valuation is high, the outcome for state $\bar{p}_L$ is warranted and the incentive compatibility constraint in (15) requires that the representative agent likes this outcome better then the outcome for state $\underline{p}_L$, which would be warranted if his valuation was low. Likewise the incentive constraint in (14) requires that a representative low-skilled agent with a low valuation of public goods likes the outcome for state $\underline{p}_L$ better than the outcome for state $\bar{p}_L$. These constraints resemble those from the mechanism design literature on public goods provision that assumes that individuals value a public good and a private good called "money". Here, however, the agent is not trading off the utility from public-goods provision against monetary payments, but against the implications that changes in the income tax system have for his well-being. Proposition 3 requires that these constraints hold for any such pair $(\underline{p}_L, \bar{p}_L)$ with $\bar{p}_L > \underline{p}_L$. This is equivalent to the monotonicity constraints in part (ii) of Proposition 3. Part (i) of the proposition states the analogous constraints for a representative high-skilled agent.

The logic of the proof of Proposition 3 is as follows.[27] If, say, the constraint that $V(s, \omega_L, \theta_H)$ 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 $V(f_H, p_L, p_H, \omega_L, \theta_H) > V(f_H, p'_L, p_H, \omega_L, \theta_H)$. If we now consider a type space, so that all individuals assign mass 1 to a distribution of types δ with $s(\delta) = (f_H, p'_L, p_H)$, individuals with a low skill level and a high valuation of public goods have an incentive to lie. If they communicate a low as opposed to a high valuation to the mechanism designer – more specifically, if they, falsely, announce

[27] A difficulty for the proof of Proposition 3 is that we cannot rely on the revelation principle. However, given some equilibrium σ^*, a false communication of, say, the public-goods preference 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 $\theta(\hat{t}) \neq \theta(t)$.

a low valuation with probability $1 - \frac{p_L}{p'_L}$, and, truthfully, announce a high valuation with probability $\frac{p_L}{p'_L}$ – they will receive the outcome intended for the case that $s = (f_H, p_H, p_L)$, and are thereby made better off, i.e., requirement (b) in Definition 1 is fulfilled. Since the lie involves only a false communication of public-goods preferences and, by individual incentive compatibility, an individual's (c, y)-bundle does not depend on his public-goods preferences, 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 valuation of public goods do not benefit from understating their preferences is satisfied provided that $\frac{\partial V^*(f_H, p_H, p_L, \omega_L, \theta_H)}{\partial p_L} \geq 0$, where V^* is the indirect utility function that was defined in Section 4. By equation (9), this condition is violated whenever

$$\theta_H \omega_L < \frac{\bar{\theta}(f_H, p_H, p_L)}{\lambda(f_H, p_H, p_L)} . \tag{16}$$

The constraint that high-skilled individuals with a low valuation of public goods do not benefit from exaggerating their preferences is satisfied provided that $\frac{\partial V^*(f_H, p_H, p_L, \omega_H, \theta_L)}{\partial p_H} \leq 0$. It can be shown that this condition is violated whenever

$$\theta_L \omega_H > \frac{\bar{\theta}(f_H, p_H, p_L)}{\lambda(f_H, p_H, p_L)} . \tag{17}$$

Since $\theta_L \omega_H = \theta_H \omega_L$, equations (16) and (17) imply that the Mirrleesian social choice function violates coalition-proofness, whenever $\theta_H \omega_L \neq \frac{\bar{\theta}(f_H, p_H, p_L)}{\lambda(f_H, p_H, p_L)}$. 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 6. 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 $\theta_L \omega_H = \frac{\bar{\theta}(f_H, p_H, p_L)}{\lambda(f_H, p_H, p_L)}$, or, equivalently, along which $p_L = 1 - \frac{f_H}{1 - f_H} p_H$. For every state s that is not on this line, coalition-proofness fails. This shows that the failure of coalition-proofness is not limited to a tiny subset of the set of states $S = [0, 1]^3$. It occurs everywhere in the unit cube, with the exception of those sates that belong to the intersection of the unit cube and the plane defined by $\{(f_H, p_H, p_L) \mid p_L = 1 - \frac{f_H}{1 - f_H} p_H\}$.[28] These observations are summarized in the following Proposition.

Proposition 4 *The optimal Mirrleesian policy, characterized in Proposition 2, violates one of the necessary conditions for robustness and coalition-proofness in Proposition 3, for every $s \in S \setminus S_+$ where $S_+ := \{(f_H, p_H, p_L) \mid p_L = 1 - \frac{f_H}{1 - f_H} p_H\}$.*

[28] Without the assumptions $\theta_L = \omega_L$ and $\theta_H = \omega_H$ there could be subsets of S in which the Mirrleesian policy satisfies the necessary conditions in Proposition 3.

Insert Figure 6 here

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 $\Lambda(\epsilon)$ 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 $\epsilon > 0$,

$$u(c(s,\omega)) - \frac{y(s,\omega)}{\omega} \geq u(c(s,\hat{\omega})) - \frac{y(s,\hat{\omega})}{\omega} + \epsilon\ , \tag{18}$$

for every s, ω, and $\hat{\omega}$. These constraints require that, for every s, an individual with skill level ω prefers the "own" consumption-output bundle $(c(s,\omega), y(s,\omega))$ strictly over any alternative bundle $(c(s,\hat{\omega}), y(s,\hat{\omega}))$, where the parameter ϵ is the minimal intensity of this strict preference; that is, the constraints in (18) require that there is a little bit of slack in the incentive compatibility constraints in (5).

Proposition 5 *Every social choice function in $\Lambda(\epsilon)$ is robustly implementable as a coalition-proof interim Nash equilibrium.*

The proof is based on a direct mechanism that reaches the given social choice function in a truthtelling equilibrium.[29] We verify that this equilibrium is coalition-proof, whatever the belief system β 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 $\Lambda(\epsilon)$, 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 public-goods preferences. Any such deviation induces a new equilibrium because (c, y)-bundles do not depend on public-goods preferences. Suppose first that the types that participate all have the same payoff type, i.e., that they all have the same skill level and the same public-goods preference. For instance, suppose that they all have the skill level ω_H and the public-goods preference θ_L. If these individuals lie about their public-goods preferences, this implies that the mechanism designer ends up with the perception that p_H is higher than it actually is. By the constraints in Proposition 3, $V(f_H, p_H, p_L, \omega_H, \theta_L)$ 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.

[29] Hence, our proof implies, in particular, that if the sufficient conditions in Proposition 5 are satisfied, then we may, without loss of generality, focus on direct mechanisms and truthtelling equilibria.

Now suppose that the types who collectively lie about their public-goods preferences 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 (ω_H, θ_L) and individuals with payoff type (ω_H, θ_H) lie about their public-goods preferences. From an ex post perspective, either the (ω_H, θ_L)-individuals think that p_H, as perceived by the mechanism designer, is too high, or the (ω_H, θ_H)-individuals think that p_H is too low. Consequently, ex interim, individuals with payoff type (ω_H, θ_L) understand that, taking the lie of individuals with payoff type (ω_H, θ_H) as given, they are weakly better off if they communicate their characteristics truthfully. Likewise, the (ω_H, θ_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.

The assumption that the belief system is moderately uninformative is needed to rule out coalitions with heterogeneous preferences where homogeneous subcoalitions see no reason to withdraw their contribution simply because they assign probability 0 to the set of states in which the withdrawal would make them better off.

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

Propositions 3 and 5 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 $\Lambda(0)$. This procedure will be applied in the following Section.

The social choice functions in $\Lambda(0)$ are not generally coalition-proof, as we explain below, using Example 2 one more time. However, as follows from Proposition 5, strict individual incentive compatibility, or, equivalently, a tiny amount of slack in the individual incentive compatibility constraints in (5) suffices to ensure robust implementability as a coalition-proof interim Nash equilibrium. We therefore expect that every social choice function that belongs to $\Lambda(0)$ can be approximated by one that belongs to $\Lambda(\epsilon)$ for some small ϵ. This is also illustrated below in the context of Example 2.[30]

[30] A general proof could proceed as follows: For the problem to characterize an optimal robust and coalition-proof social choice function, formally studied in the next section, one replaces the incentive compatibility constraints in (5) by those in (18). If, for ϵ close to zero, the optimal social choice function

Example 2 revisited. To illustrate why the social choice functions in $\Lambda(0)$ 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 4, where the social choice function that is illustrated in Figure 4 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 4 is modified as follows: in both graphs, the bundle for high-skilled individuals is moved to a slightly higher indifference curve.[31] 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 **??**. 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.

5.3 On the separability of individual and collective incentive problems

The reasoning in section 5.2 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

depends in a continuous way on the parameter ϵ, then it follows that any social choice function that is optimal in the set $\Lambda(0)$ can be approximated by one that is robust and coalition-proof. Providing such a proof of continuity is, however, beyond the scope of this paper.

[31] To preserve feasibility, we may simultaneously have to move the low-skilled individuals to a slightly lower indifference curve.

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.

6 Optimal robust and coalition-proof social choice functions

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. In principle, there are three different sources of aggregate uncertainty: (i) uncertainty about the cross-section distribution of productive abilities, i.e., about f_H (ii) uncertainty about the distribution of public-goods preferences among low-skilled individuals, characterized by p_L, and (iii) uncertainty about the distribution of public-goods preferences among high-skilled individuals, or the value of p_H. To simplify the exposition, we will look at each source in isolation, i.e. we ask first what the optimal policy looks like if only f_H is unknown, then we ask what the the optimal policy looks like if only p_L is unknown, etc. [32]

We assume that the dimension in which aggregate uncertainty prevails is commonly known. This implies that the mechanism designer can prevent individuals from communicating their characteristics in a way that is inconsistent with what is commonly known about the set of states. He simply has to make the outcomes that are induced by such deviations sufficiently unattractive.

The case in which only f_H is unknown is straightforward. We have seen in the previous section that, even though there may be uncertainty about the population share of high-skilled individuals, eliciting this information does not require an essential adjustment of the policy mechanism. The optimal coalition-proof policy is therefore essentially equivalent to the optimal Mirleesian policy. Below we focus on the case in which p_L is unknown. The case in which p_H is unknown is very similar and is dealt with in the Appendix.

If p_L is unknown, the optimal Mirrleesian policy fails to be coalition-proof because, if p_L is sufficiently high, low-skilled individuals with a high valuation of public goods would benefit from understating their preferences. To derive an optimal coalition-proof policy, we therefore have to add the constraint that this behavior must not be attractive.

Given that f_H and p_H are assumed to be known, we can identify a state s of the economy with a particular value of p_L. We assume that the mechanism designer has subjective beliefs about the possible realizations of p_L. For simplicity, we assume that she

[32] A comprehensive treatment where f_H, p_H and p_L are all unknown would be feasible and give rise to conclusions that are qualitatively similar. However, this would come at the cost of a more involved analysis.

assumes that p_L is uniformly distributed over the unit interval.[33] The optimal policy can now be characterized as a solution to the following optimization problem: For every $p_L \in [0,1]$, Choose $c(p_L, \omega_L)$, $y(p_L, \omega_L)$, $c(p_L, \omega_H)$, $y(p_L, \omega_H)$ and $q(p_L)$ in order to maximize $E\left[W(p_L)\right]$ subject to the incentive constraints in (5), the resource constraints in (6), and the constraint $\frac{\partial V(p_L, \omega_L, \theta_H)}{\partial p_L} \geq 0$. We refer to this problem in the following as problem $\mathcal{P}_L(p_H, f_H)$.

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. 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, ρ),

$$\begin{array}{rcll} V_H(v_L, \rho) & := & \max & u(c_H) - \frac{y_H}{\omega_H} \\ & & \text{s.t.} & u(c_H) - \frac{y_H}{\omega_H} \geq u(c_L) - \frac{y_L}{\omega_H} \ , u(c_L) - \frac{y_L}{\omega_L} \geq u(c_H) - \frac{y_H}{\omega_L} \ , \\ & & & f_H(y_H - c_H) + (1 - f_H)(y_L - c_L) = \rho \ , u(c_L) - \frac{y_L}{\omega_L} = v_L \ . \end{array}$$

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 ρ.[34]

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 consider 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

$$\int_0^1 \{\bar{\theta}(p_L)q(p_L) + f_H V_H(v_L(p_L), r(q(p_L))) + (1 - f_H)v_L(p_L)\}dp_L$$

subject to the constraint, that for all p_L,

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

The constraint ensures that the utility of low-skilled individuals with a high valuation of public goods is a non-decreasing function of their population share. Otherwise coalition-proofness would be violated.[35] With a Mirrleesian approach, this constraint would not

[33] 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 a uniform prior simplifies the exposition. However, the logic of the analysis would remain the same with alternative assumptions about the mechanism designer's beliefs.

[34] For a complete characterization of the function V_H, see Proposition 1 in Bierbrauer and Boyer (2010).

[35] To achieve coalition-proofness we must also guarantee that low-skilled individuals with a low valuation are not tempted to exaggerate their preferences. This requires that $\theta_L q'(p_L) + v'_L(p_L) \leq 0$. Fortunately, it can be shown that this constraint is never binding. If the constraint $\theta_H q'(p_L) + v'_L(p_L) \geq 0$ binds, then it follows from $\theta_L < \theta_H$ and $q'(p_L) > 0$, that $\theta_L q'(p_L) + v'_L(p_L) < 0$. If the constraint does not bind, then the optimal coalition-proof policy coincides with the optimal Mirrleesian policy. Again, this implies that $\theta_L q'(p_L) + v'_L(p_L) < 0$.

be taken into account and the objective function could be maximized pointwise. Consequently, the optimal Mirrleesian policy is such that, for every p_L, $q(p_L)$ and $v_L(p_L)$ are characterized as the solution to the following first order conditions:

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

where V_{Hj} is the derivative of the function V_H with respect to its jth-argument. The first equation says that the social benefit of increased public-goods provision equals zero, and the second equation requires that the welfare gain from a marginal increase of the utility promise to the low-skilled individuals must be zero.

In the following, we say that there is an upward distortion of the public-goods provision level whenever $\bar{\theta} + f_H V_{H2}\; r'(q) < 0$, so that a reduction of the public-goods provision level would increase welfare. Analogously, we say that the public-goods provision level is distorted downwards when $\bar{\theta} + f_H V_{H2}\; r'(q) > 0$. We say that marginal income taxes and direct transfers to the low-skilled are distorted upwards if $f_H V_{H1} + 1 - f_H < 0$, so that reducing the utility promise to the low-skilled would increase welfare. Marginal income taxes and direct transfers to the high-skilled are said to be distorted upwards if $f_H V_{H1} + 1 - f_H > 0$. This choice of terminology is justified by the observation that there is a monotonic relation between the utility promise to the low-skilled and our measures of how distortionary and redistributive the tax system is: If v_L is higher than stipulated by the optimal Mirrleesian policy, then the underlying allocation is such that the low-skilled individual's marginal income tax rate as well as their income transfers are higher than under the Mirrleesian policy.[36]

Proposition 6 *The solution to Problem $\mathcal{P}_L(p_H, f_H)$ has the following properties. There are cutoff values γ^1 and γ^2 with $0 \leq \gamma^1 < \gamma^2 < 1$ such that:*

(a) For $p_L \leq \gamma^1$ the solution to $\mathcal{P}_L(p_H, f_H)$ coincides with the Mirrleesian policy.

(b) For $p_L \in (\gamma^1, \gamma^2)$, marginal income taxes, direct transfers to the low-skilled and public-goods provision levels are distorted downwards.

(c) For $p_L > \gamma^2$, marginal income taxes, direct transfers to the low-skilled and public-goods provision levels are distorted upwards.

Part (a) of the Proposition says that there may or may not be a parameter region, where the optimal coalition-proof policy and the optimal Mirrleesian policy coincide. Parts (b) and (c) say that there always is a region where the two policies do not coincide, i.e. a region in which the optimal coalition-proof policy is distorted. (Figure 7 provides a computed example in which the optimal policy coincides with the Mirrleesian policy for $p_L \leq 0.25$

[36] See Proposition 2 in Bierbrauer and Boyer (2010).

and is distorted otherwise.). These distortions have a particular pattern: First, public-goods provision and income taxation are always distorted in the same direction. If there is overprovision of public-goods, then there is also a tax system that is too distortionary and too redistributive. If there is underprovision of public goods, then the tax system is not redistributive enough and its distortions are too small. Second, upward distortions occur if the demand for public goods is high (p_L large), and downward distortions occur if the demand for public goods is low.

The proof of Proposition 6 relies mainly on two optimality conditions, which are formally derived in the Appendix. The first one is as follows:

$$\frac{1}{\theta_H}(\bar{\theta} + f_H V_{H2}\, r'(q)) = f_H V_{H1} + 1 - f_H \ . \tag{19}$$

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 $f_H V_{H1} + 1 - f_H$.[37] Equation (19) implies a *complementarity* between redistribution and public-goods provision: if we have an excessively distortionary and redistributive tax system, 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, $\bar{\theta} + f_H V_{H2}\, r'(q) < 0$, and vice versa.

An intuitive understanding of this finding may be obtained as follows: We need to deviate from the Mirrleesian policy, if the constraint $\theta_H q'(p_L) + v_L'(p_L) \geq 0$ binds so that the the utility of low-skilled individuals with a high valuation of public goods remains constant as p_L goes up, i.e. $\theta_H q(p_L) + v_L(p_L) = \text{const}$. Now suppose that, for some p_L, $q(p_L)$ is too high and $v_L(p_L)$ is too low from a welfare perspective. Then we could raise welfare by lowering $q(p_L)$ and increasing $v_L(p_L)$ without violating the constraint that $\theta_H q(p_L)$ and $v_L(p_L)$ add up to a constant. Hence, a situation where the public-goods provision level and the tax system are distorted in opposing directions cannot be optimal.

Proposition 6 stipulates that public-goods provision and redistribution should be distorted upwards if p_L is high, and that they should be distorted downwards otherwise. This follows from the second optimality condition, which stipulates that the "average distortion" must be zero, i.e., a solution has to be such that

$$\int_0^1 \{\bar{\theta} + f_H V_{H2}\, r'(q)\} dp_L = \int_0^1 \{f_H V_{H1} + 1 - f_H\} dp_L = 0 \ . \tag{20}$$

This condition says that any upward distortion of public-goods provision and redistribution that occurs over some subinterval of $[0, 1]$ has to be balanced by a downward distortion over some other subinterval. Given that this "budget condition" holds, it is

[37] The Mirrleesian policy also satisfies this since

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

However, as we have argued before (see Figure 4), it violates the constraint $\theta_H q'(p_L) + v_L'(p_L) \geq 0$ over some subinterval of $[0, 1]$.

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.

Figure 7 below illustrates the optimal tax and expenditure policy, again under the assumptions that $u(c) = \sqrt{c}$ and that $r(q) = \frac{1}{2}q^2$. Moreover, we assume that $f_H = p_H = \frac{1}{2}$. The figure illustrates how the public-goods provision, the net tax payment of the low-skilled, their marginal income tax rates, and the distortion relative to the Mirrleesian policy, as measured by $f_H V_{H1} + 1 - f_H$, vary with p_L. In this figure, thin grey lines represent the Mirrleesian policy and fat black lines represent the optimal coalition-proof policy.

Insert Figure 7 here

The figure illustrates that the optimal coalition-proof policy coincides with the optimal Mirrleesian policy for p_L close to zero. Then, there is a range in which income transfers to the low-skilled as well as their marginal income tax rates are lower than under the Mirrleesian policy. Moreover, the utility promise to the low-skilled, v_L, is lower: we have that $f_H V_{H1} + 1 - f_H > 0$ or, given that $f_H = \frac{1}{2}$, $\mid V_{H1} \mid< 1$. For high values of p_L, by contrast, direct income transfers to the low-skilled, their marginal income tax rate and their utility promise are distorted upwards.

The most striking difference between the Mirrleesian policy and the optimal coalition-proof policy concerns the relation between *changes in the public-goods provision level* and *changes in the tax policy.* If we consider an increase of p_L, i.e., a preference shock that leads to an increased supply of public goods, then, under the Mirrleesian policy, there is no effect at all on the tax policy. With the optimal coalition-proof policy, by contrast, the tax policy is affected: If the initial situation involves large expenditures on public goods, then a further increase of these expenditures goes together with a more redistributive and more distortionary income tax system. If the initial expenditure level is small, then increased spending on public goods has either no effect on the tax policy, or is associated with a less redistributive and less distortionary income tax system.

Empirical implications. Broadly speaking, Proposition 6 gives rise to the following conclusion: If the demand for public goods is low, then public-goods provision as well as marginal income taxes and direct income transfers are distorted downwards. If the demand for public goods is high, then all these policies are distorted upwards. Moreover, the numerical example suggests that an increase in expenditures is most likely associated with an increase of income transfers and increase of marginal income tax rates.

The main empirical implication of the analysis is the following: Suppose that we compare two countries whose policies are interpreted as resulting from the mechanism design exercise above. Then the prediction would be that the country that spends more on public goods will also have higher marginal income taxes and higher income transfers. Also, from Mirrleesian perspective, taxes, transfers and expenditures are too high in the country that spends a lot, and too low in the country that spends only little. For instance, if we interpret the US and Sweden as two countries who have essentially the same economic fundamentals (the same productivity of the average worker), but among the Swedes more people have a high valuation of public goods, then we would not only expect public expenditures to be higher in Sweden, but also the marginal income tax rates below the top, and the level of direct income transfers to the poor. We would not arrive at this prediction on the basis of the simple Mirrleesian model that led to Proposition 2.

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. The incorporation of this additional constraint therefore brings the outcomes of normative public economics closer to the forces that determine policy outcomes in the real world.

7 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, the optimal policy gives rise to a positive relationship between the expenditures on public goods, direct income transfers and marginal tax rates. The model predicts that if two countries differ in their public expenditure levels, then the country that spends more on public goods will also have more direct income transfers to the poor and a more distortionary tax system.

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. (2009a). A note on optimal income taxation, public-goods provision and robust mechanism design. *Journal of Public Economics*, 93:667–670.

Bierbrauer, F. (2009b). Optimal income taxation and public-good provision with endogenous interest groups. *Journal of Public Economic Theory*, 11:311–342.

Bierbrauer, F. (2011). Distortionary taxation and the free-rider problem. *International Tax and Public Finance*, forthcoming.

Bierbrauer, F. and Boyer, P. (2010). The Pareto-frontier in a simple Mirrleesian model of income taxation. Preprint 2010/16, 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.

Bierbrauer, F. and Sahm, M. (2010). Optimal democratic mechanisms for income taxation and public good provision. Journal of Public Economics.

Boadway, R. and Keen, M. (1993). Public goods, self-selection and optimal income taxation. *International Economic Review*, 34:463–478.

Che, Y. and Kim, J. (2006). Robustly collusion-proof implementation. *Econometrica*, 74:1063–1107.

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.

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 Proofs

A.1 Proof of Proposition 1

We fix $T, \mathcal{T}$, and π. By the standard version of the revelation principle, a social choice function (q, c, y) is implementable as an interim 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 an interim Nash equilibrium; i.e., for all t,

$$t \in \operatorname{argmax}_{t' \in T} \int_{\Delta(T)} U(Q(\delta), C(\delta, t'), Y(\delta, t'), \omega(t), \theta(t)) \, d\beta(\delta \mid t) \,, \tag{21}$$

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

$$Q(\delta) = q(s(\delta)) \tag{22}$$

and, for every t,

$$C(\delta, t) = c(s(\delta), \omega(t), \theta(t)) \text{ and } Y(\delta, t) = y(s(\delta), \omega(t), \theta(t)) \,. \tag{23}$$

We first show that $(b) \Rightarrow (a)$. Consider an incentive compatible social choice function (q, c, y). For an arbitrary belief system β construct a direct mechanism $M = [(T, \mathcal{T}), Q, C, Y]$ such that (22) and (23) hold. We seek to verify that, for every t,

$$\begin{aligned} t &\in \operatorname{argmax}_{t' \in T} \textstyle\int_{\Delta(T)} U(Q(\delta), C(\delta, t'), Y(\delta, t'), \omega(t), \theta(t)) \, d\beta(\delta \mid t) \\ &= \operatorname{argmax}_{t' \in T} \textstyle\int_{S} U(q(s), c(s, \omega(t'), \theta(t')), y(s, \omega(t'), \theta(t')), \omega(t), \theta(t)) \, d\hat{\beta}(s \mid t) \,, \end{aligned}$$

where, for any $S' \subset S$, $\hat{\beta}(S' \mid t) := \beta\left(\{\delta \in \Delta(T) \mid s(\delta) \in S'\} \mid t\right)$. Equivalently, for every t, $(\omega(t), \theta(t))$ solves

$$\max_{(\omega', \theta') \in \Omega \times \Theta} \int_{S} U(q(s), c(s, \omega', \theta'), y(s, \omega', \theta'), \omega(t), \theta(t)) \, d\hat{\beta}(s \mid t) \,.$$

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 β is such that for some s', $\beta(\{\delta \mid s(\delta) = s'\} \mid t) = 1$, for all t. Suppose that a direct mechanism (T, Q', C', Y') truthfully implements (q, c, y). Using conditions (22) and (23) to substitute for Q', C', and Y', the equilibrium condition in (21) becomes: for all t and all t',

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

or, equivalently, for all (ω, θ) and (ω', θ'),

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

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

Consider a class of types spaces which all have the same set of types $(T, \mathcal{T})$ and the same payoff type function π, but which may differ in the belief system β. Suppose there is a mechanism $M = [(A, \mathcal{A}), Q, C, Y]$ with an equilibrium σ^* that implements the social choice function (q, c, y) as a coalition-proof interim Nash equilibrium on all these type spaces. In particular, this requires that the mechanism reaches the social choice function; i.e., for every δ, conditions (12) and (13) are fulfilled.

We show that this implies that $V(s, \omega_L, \theta_H)$ 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

$$V(f_H, p_L, p_H, \omega_L, \theta_H) > V(f_H, p'_L, p_H, \omega_L, \theta_H) \ . \tag{24}$$

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 σ^* is a coalition-proof interim 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 $T' = \{t \mid (\omega(t), \theta(t)) = (\omega_L, \theta_H)\}$, 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 $\hat{t} \in \hat{T} = \{t \mid (\omega(t), \theta(t)) = (\omega_L, \theta_L)\}$, 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 truthtelling by all others. For types in T', define the "lie" $\ell'_{T'} : T' \to \Delta(T)$ 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 $h_{T\setminus T'} : T \setminus T' \to \Delta(T)$ 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 $\delta \in \Delta(T)$, an announced cross-sectional distribution of types $\bar{\delta}(\delta)$ with

$$\bar{\delta}(\tilde{T} \mid \delta) = \int_{t'\in T'} \ell'_{T'}(\tilde{T} \mid t')d\delta(t') + \int_{t\in T\setminus T'} h_{T\setminus T'}(\tilde{T} \mid t)d\delta(t) \ , \tag{25}$$

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

$$\sigma'_{T'}(A' \mid t') = \int_{\hat{t}\in T} \sigma^*(A' \mid \hat{t}) \ d\ell'_{T'}(\hat{t} \mid t') \ . \tag{26}$$

This construction ensures that, for every δ, 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 σ^* and the distribution of types equals $\bar{\delta}(\delta)$. Formally, for every δ,

$$\alpha(\bar{\delta}(\delta), \sigma^*) = \alpha(\delta, (\sigma^*_{T\setminus T'}, \sigma'_{T'})) \ . \tag{27}$$

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

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

Step 2: Consider a specific type space

Consider a type space with a belief system so that, for some δ 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

$$s(\bar{\delta}(\delta)) = (f_H, p_H, p_L) \ , \tag{28}$$

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 $(\sigma^*_{T\backslash T'}, \sigma'_{T'})$, the expected payoff of a type $t' \in T'$ equals

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

where

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

By Proposition 1, robust implementability of a social choice function as an interim 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 ω, but not on θ. 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, $\Phi(t') = V(s(\bar{\delta}(\delta)), \omega(t'), \theta(t'))$ so that so that, for every t' in T',

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

This observation in conjunction with equations (24) and (28) implies that types in T' are made strictly better off by this deviation.

Step 4: Show that, on this type space, $(\sigma^*_{T\backslash T'}, \sigma'_{T'})$ is an interim 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 σ^* 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 ω, but not on θ, 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 $(\sigma^*_{T\backslash T'}, \sigma'_{T'})$ 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 (28).

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

The deviating individuals have the same preferences, $(\omega(t'), \theta(t')) = (\omega_L, \theta_H)$, 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 5

Given a measurable space $(T, \mathcal{T})$ and a function $\pi = (w, \theta) : T \mapsto \Omega \times \Theta$, 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 $\delta \in \Delta(T)$ and all $t \in T$,

$$Q(\delta) = q(s(\delta)), C(\delta, t) = c(s(\delta), \omega(t)), \text{ and } Y(\delta, t) = y(s(\delta), \omega(t)) \; . \tag{29}$$

This construction ensures that the mechanism achieves the social choice function in a truthtelling equilibrium. More formally, the strategy $h : T \to \Delta(T)$ with $h(\{t\} \mid t) = 1$, for all t, is interim Nash equilibrium of the game induced by this mechanism, for every belief system β. 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 β.

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 $\ell'_{T'} : T' \to \Delta(T)$. We say that such a lie involves lies about skills if there is $t' \in T'$ so that

$$l(\hat{\omega} \mid t') := \ell'_{T'}(\{\hat{t} \mid \omega(\hat{t}) \neq \omega(t')\} \mid t') > 0 \; . \tag{30}$$

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 $\bar{\delta}(\delta) \in \Delta(T)$ (see the definition in equation (25)) 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 (29) holds, the expected payoff of an individual with a type $t' \in T'$ whose behavior satisfies (30) can be written as

$$\begin{aligned} &\int_{\Delta(T)} l(\hat{\omega} \mid t') \left\{ \theta(t) q(s(\bar{\delta}(\delta))) + u(c(s(\bar{\delta}(\delta)), \hat{\omega})) - \tfrac{y(s(\bar{\delta}(\delta)), \hat{\omega})}{\omega(t)} \right\} \\ &\quad + (1 - l(\hat{\omega} \mid t')) \left\{ \theta(t) q(s(\bar{\delta}(\delta))) + u(c(s(\bar{\delta}(\delta)), \omega(t)) - \tfrac{y(s(\bar{\delta}(\delta)), \omega(t))}{\omega(t)}) \right\} d\beta(\delta \mid t) \; , \end{aligned} \tag{31}$$

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

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

$$\int_{\Delta(T)} \left\{ \theta(t) q(s(\bar{\delta}(\delta))) + u(c(s(\bar{\delta}(\delta)), \omega(t)) - \tfrac{y(s(\bar{\delta}(\delta)), \omega(t))}{\omega(t)} \right\} d\beta(\delta \mid t) \; . \tag{32}$$

By the constraints in (18) we have that

$$\begin{aligned} &\theta(t) q(s(\bar{\delta}(\delta))) + u(c(s(\bar{\delta}(\delta)), \omega(t))) - \tfrac{y(s(\bar{\delta}(\delta)), \omega(t))}{\omega(t)} \\ &> \theta(t) q(s(\bar{\delta}(\delta))) + u(c(s(\bar{\delta}(\delta)), \hat{\omega}) - \tfrac{y(s(\bar{\delta}(\delta)), \hat{\omega})}{\omega(t)} \; , \end{aligned}$$

which implies that the expression in (32) is strictly larger than the expression in (31). This shows that, for a type $t' \in T'$, behaving in such a way that (30) holds is not a best response. Hence, $(h_{T \setminus T'}, \ell'_{T'})$ is not an interim 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 $\pi(t') = (\omega_H, \theta_L)$ 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

$$\ell'_{T'}(\{\hat{t} \mid \omega(\hat{t}) \neq \omega(t')\} \mid t') = 0 \ , \tag{33}$$

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

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

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

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

Since the given social choice function satisfies the monotonicity constraint

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

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 (33) holds. Consequently, we have for all δ 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 $\pi(t'') = (\omega_H, \theta_L)$ and that these individuals lie about their preference parameter with positive probability,

$$\ell'_{T'}(\{\hat{t} \mid \theta(\hat{t}) = \theta_H\} \mid t'') > 0 \ . \tag{35}$$

Given that the monotonicity constraint

$$\frac{\partial V(f_H, p_H, p_L, \omega_H, \theta_L)}{\partial p_H} \leq 0 \ ,$$

holds, these types will benefit from the deviation $\ell'_{T'}$ only if there is a subset D of $\Delta(T)$ with $\beta(D \mid t') > 0$ for all $t' \in T'$ with $\pi(t') = (\omega_H, \theta_L)$, 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 $\pi(t') = (\omega_H, \theta_L)$ 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 preference for public goods who announce a low preference with positive probability: there is $\hat{T}'' \subset T'$ so that $\hat{t}'' \in \hat{T}''$ $\pi(\hat{t}'') = (\omega_H, \theta_H)$, and

$$\ell'_{T'}(\{\hat{t} \mid \theta(\hat{t}) = \theta_L\} \mid \hat{t}'') > 0 \ . \tag{36}$$

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

$$\ell''_{\hat{T}''}(\{\hat{t} \mid \theta(\hat{t}) = \theta_L\} \mid \hat{t}'') < \ell'_{T'}(\{\hat{t} \mid \theta(\hat{t}) = \theta_L\} \mid \hat{t}'') \ . \tag{37}$$

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 δ 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 $\delta \in \Delta(T)$,

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

Given that the monotonicity constraint

$$\frac{\partial V(s, \omega_H, \theta_H)}{\partial p_H} \geq 0 \ , \tag{38}$$

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 $\beta(\tilde{D} \mid t') > 0$ so that (38) holds as a strict inequality. Finally, observe that $(h_{T \setminus T'}, \ell'_{T' \setminus \hat{T}''}, \ell''_{\hat{T}''})$ is an interim Nash equilibrium strategy because individuals communicate their skill levels truthfully and, individual outcomes do not depend on announced preference parameters (see equation (29)). Hence, if there is a subset $\tilde{D}$ of D with $\beta(\tilde{D} \mid t') > 0$ so that so that (38) 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 $\frac{\partial V(s, \omega_H, \theta_L)}{\partial p_H} \leq 0$ 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 Proposition 6

We use optimal control theory in order to characterize the solution to the optimization problem in the body of text. 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,

$$q' = g_1(u_1) \ , \quad \text{with} \quad g_1(u_1) = u_1 \ , \tag{39}$$

and

$$v_L' = g_2(u_2) \ , \quad \text{with} \quad g_2(u_2) = u_2 \ . \tag{40}$$

The monotonicity constraint $\theta_H q' + v_L' \geq 0$ can now be formulated as a constraint on the control variables,

$$h(u_1, u_2) \geq 0 \ , \quad \text{where} \quad h(u_1, u_2) = \theta_H u_1 + u_2 \ . \tag{41}$$

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 μ_1 is the costate variable associated with (39) and μ_2 is the costate variable associated with (40); 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 (41). The optimality conditions are as follows:[38] (i) The costate variables satisfy

$$\mu_1' = -\frac{\partial \mathcal{H}}{\partial q} \quad \text{and} \quad \mu_2' = -\frac{\partial \mathcal{H}}{\partial v_L} \ , \tag{42}$$

or, equivalently,

$$\mu_1' = -(\bar{\theta} + f_H V_{H2} \, r'(q)) \tag{43}$$

and

$$\mu_2' = -(f_H V_{H1} + 1 - f_H) \ . \tag{44}$$

(ii) The fact that we have free start values for the state variables q and v_L implies that[39]

$$\mu_1(0) = \mu_1(1) = 0 \quad \text{and} \quad \mu_2(0) = \mu_2(1) = 0 \ . \tag{45}$$

[38] 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 $r(q)$; see Proposition 1 in Bierbrauer and Boyer (2010).

[39] The end values are not free. It can be shown that the optimality conditions pin down the paths of the state and control variables in a way that the yields a particular end value of the state variables.

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

$$\frac{\partial \mathcal{L}}{\partial u_1} = 0 \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial u_2} = 0 \ , \tag{46}$$

and

$$\nu \geq 0 \ , \quad \text{and} \quad \nu h(u_1, u_2) = 0 \ . \tag{47}$$

Equations (46) can equivalently be written as

$$\mu_1 + \nu \theta_H = 0 \ , \tag{48}$$

and

$$\mu_2 + \nu = 0 \ . \tag{49}$$

Step 1: Derivation of the optimality conditions in the body of text, i.e., of equations (19) and (20)

Equations (45), (48), and (49) imply that

$$\nu(0) = \nu(1) = 0 \ . \tag{50}$$

Equations (48) and (49) also imply that

$$\mu_1' = -\frac{1}{\theta_H}\nu' \ , \tag{51}$$

and

$$\mu_2' = -\nu' \ . \tag{52}$$

Using (51) and (52) in conjunction with (43) and (44) yields

$$\frac{1}{\theta_H}(\bar{\theta} + f_H V_{H2}\, r'(q)) = f_H V_{H1} + 1 - f_H \ , \tag{53}$$

which is equal to equation (19) in the body of the text. To also derive (20), note that equations (50) and (52) imply that $\nu(1) - \nu(0) = -\int_0^1 \mu_2'(p_L)dp_L = 0$. Combining this with (44) yields

$$\int_0^1 \{f_H V_{H1}(v_L(p_L), r(q(p_L))) + 1 - f_H\} dp_L = 0 \ .$$

Step 2: Implications of the optimality conditions

The proof of Proposition 6 is based on three Lemmas.

Lemma 1 *Let $p_L = 1 - \frac{f_H}{1-f_H} p_H$, then condition (53) implies that $r'(q(p_L)) = \theta_H \omega_L$.*

Proof We first note that the function V_H introduced in Step 1 has the following property,[40]

$$V_{H2} = \frac{V_{H1}}{\omega_L} - \frac{1}{\omega_H} \; . \tag{54}$$

Now suppose that the Lemma is false. Then we have $r'(q) \neq \theta_H \omega_L$. Using the optimality condition (53), we may solve for $r'(q)$ and state this equivalently as

$$\bar{\theta} - \theta_H (f_H V_{H1} + 1 - f_H) \neq -\theta_H \omega_L f_H V_{H2} \; .$$

Using (54) to substitute for V_{H2}, we can rewrite this condition once more as $\frac{\bar{\theta}}{\lambda} \neq \theta_H \omega_L$.

However, $p_L = 1 - \frac{f_H}{1-f_H} p_H$ implies that $\bar{\theta} = \theta_H(1-f_H) + \theta_L f_H$ and also that $\frac{\bar{\theta}}{\lambda} = \theta_H \omega_L$. Hence, the assumption that the Lemma is false has led to a contradiction.

□

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

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

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

Suppose first that constraint (41) is binding, then this can be equivalently written as

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

where $Q := V_{H11} \theta_H^2 - 2 V_{H12} \theta_H r' + V_{H22} (r')^2$ is a quadratic form which is non-positive because the function V_H is concave, as follows from Proposition 1 and Lemma 12 in Bierbrauer and Boyer (2010). Using that $V_{H2} < 0$ (again, see Bierbrauer and Boyer (2010), Lemma 12), and that $r'' > 0$ establishes the result.

If the constraint (41) is not binding, then the optimal policy is equal to the Mirrleesian policy characterized in Proposition 2. It follows from this Proposition that the Mirrleesian policy is also such that $q'(p_L) > 0$. □

[40] See Lemma 12 of Bierbrauer and Boyer (2010).

Lemma 3 *For* $p_L \in (1 - \frac{f_H}{1-f_H}p_H, 1)$, $\nu''(p_L) < 0$*; for* $p_L = 1 - \frac{f_H}{1-f_H}p_H$, $\nu''(p_L) = 0$*; for* $p_L \in (0, 1 - \frac{f_H}{1-f_H}p_H)$, $\nu''(p_L) \geq 0$*, with a strict inequality if constraint (41) is binding.*

Proof Optimality conditions (52) and (44) imply that $\nu' = f_H V_{H1} + 1 - f_H$. Hence,

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

If constraint (41) 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 $V_{H12} = \frac{1}{\omega_L} V_{H11}$, we can rewrite this as

$$\nu'' = -f_H V_{H11} \frac{1}{\omega_L} q'(\theta_H \omega_L - r'(q)) \, . \tag{55}$$

It follows from Lemmas 1 and 2 that (i) $q'(\theta_H \omega_L - r'(q)) < 0$, if $p_L > 1 - \frac{f_H}{1-f_H}p_H$, (ii) that $q'(\theta_H \omega_L - r'(q)) = 0$, if $p_L = 1 - \frac{f_H}{1-f_H}p_H$ and (iii) that $q'(\theta_H \omega_L - r'(q)) > 0$, if $p_L < 1 - \frac{f_H}{1-f_H}p_H$. Moreover, it is shown in Proposition 1 of Bierbrauer and Boyer (2010) that $V_{H11} \leq 0$. Consequently, (55) implies (i) that $\nu'' < 0$ if constraint (41) binds and $p_L > 1 - \frac{f_H}{1-f_H}p_H$, (ii) that $\nu'' > 0$ if constraint (41) binds and $p_L < 1 - \frac{f_H}{1-f_H}p_H$ and (iii) that $\nu'' = 0$ if $p_L = 1 - \frac{f_H}{1-f_H}p_H$.

It has to be the case that (41) binds if $p_L > 1 - \frac{f_H}{1-f_H}p_H$. Suppose otherwise: then the solution to the policy problem coincides with the Mirrleesian policy. This policy is not coalition-proof whenever $p_L > 1 - \frac{f_H}{1-f_H}p_H$, a contradiction.

If constraint (41) does not bind for a subset of $(0, 1 - \frac{f_H}{1-f_H}p_H)$, then $\nu' = \nu'' = 0$ over this range.

□

Step 3: Completing the proof of Proposition 6

Since $\nu' = f_H V_{H1} + 1 - f_H = \frac{1}{\theta_H}(\bar{\theta} + f_H V_{H2}\, r'(q))$, the proof is complete, if we show that there are cutoff values γ_1 and γ_2 with $\gamma_1 < \gamma_2 < 1$ so that $p_L \in (\gamma_1, \gamma_2)$ implies $\nu' > 0$ and $p_L > \gamma_2$ implies $\nu' < 0$.

We know that $\nu(p_L) > 0$ for $p_L > 1 - \frac{f_H}{1-f_H}p_H$. Suppose first that $\nu(p_L) = 0$ for all $p_L \leq 1 - \frac{f_H}{1-f_H}p_H$. Then Lemma 3 implies that $\nu'' < 0$ for $p_L > 1 - \frac{f_H}{1-f_H}p_H$ and $\nu = \nu' = \nu'' = 0$, otherwise. Given that $\int \nu' dp_L = 0$, it has to be the case that there exist γ_1 and γ_2 with $1 - \frac{f_H}{1-f_H}p_H = \gamma_1$ so that $p_L \in (\gamma_1, \gamma_2)$ implies $\nu' > 0$ and $p_L > \gamma_2$ implies $\nu' < 0$.

Now suppose that $\nu(p_L) > 0$ for some $p_L < 1 - \frac{f_H}{1-f_H}p_H$. Consider the smallest p_L with this property and denote it by $\underline{p}_L$. Since $\nu(0) = 0$, it must be true that $\nu'(\underline{p}_L) > 0$.

By Lemma 3 we also have $\nu'' > 0$, for all $p_L \in (\underline{p}_L, 1 - \frac{f_H}{1-f_H} p_H)$. Hence $\nu' > 0$, for all $p_L \in (\underline{p}_L, 1 - \frac{f_H}{1-f_H} p_H)$. At $p_L = 1 - \frac{f_H}{1-f_H} p_H$, ν' starts to fall, and because of $\int \nu' dp_L = 0$ it must eventually become negative.

A.5 The optimal robust and coalition-proof policy if p_H is unknown

We assume that p_L is known, but that there is a shock to the preferences of the high-skilled, so that p_H has to be elicited from reports of high-skilled individuals about their public-goods preferences. The policy problem, henceforth referred to as Problem $\mathcal{P}_H(p_L, f_H)$, can be stated as follows: choose the functions $q : p_H \mapsto q(p_H)$ and $v_L : p_H \mapsto v_L(p_H)$ in order to maximize

$$\int_0^1 \{\bar{\theta}(p_H) q(p_H) + f_H V_H(v_L(p_H), r(q(p_H))) + (1 - f_H) v_L(p_H)\} dp_H$$

subject to the constraint, that for all p_L,

$$\theta_L q'(p_H) + \frac{d}{dp_H} V_H(v_L(p_H), r(q(p_H))) \leq 0 \ .$$

The constraint ensures that the utility of high-skilled individuals with a low valuation of public goods does not increase if more high-skilled individuals communicate a high valuation of public goods. The following Proposition characterizes the solution to Problem $\mathcal{P}_H(p_L, f_H)$. It is the mirror image of Proposition 6, i.e., it follows from exactly the same reasoning, so that we can omit a formal proof. Again, the main observation is that the tax system and the public-goods provision are always distorted in the same direction, with upward distortions occurring if the demand for public goods is high, and downward distortions if the demand for public goods is low.

Proposition 7 *The solution to Problem $\mathcal{P}_H(p_L, f_H)$ has the following properties. There are cutoff values η^1 and η^2 with $0 < \eta^1 < \eta^2 \leq 1$ such that:*

(a) For $p_H \geq \eta^2$, the solution to $\mathcal{P}_H(p_L, f_H)$ coincides with the Mirrleesian policy.

(b) For $p_H < \eta^1$, marginal income taxes, direct transfers to the low-skilled and public-goods provision levels are distorted downwards.

(c) For $p_H \in (\eta^1, \eta^2)$, marginal income taxes, direct transfers to the low-skilled and public-goods provision levels are distorted upwards.

Figure 8 illustrates the optimal tax and expenditure policy under the assumptions that $u(c) = \sqrt{c}$ and that $r(q) = \frac{1}{2} q^2$. Moreover, we assume that $f_H = p_L = \frac{1}{2}$. The figure illustrates how the public-goods provision, the net tax payment of the low-skilled, their marginal income tax rates, and the distortion relative to the Mirrleesian policy vary with p_H. The measure of the distortion is now given by $f_H + \frac{1-f_H}{V_{H1}}$.[41] Hence, an upward distortion corresponds to a situation where $\frac{1}{V_{H1}} > -1$. The figure illustrates once more that there is a comparatively large range of parameters, where increases of public expenditures are associated with a tax system that becomes more redistributive and distortionary, and a comparatively small range of parameters where higher expenditures go together with reduced transfers and smaller distortions.

Insert Figure 8 here

B Figures

[41] It is convenient to measure the distortion in this way because of the optimality condition $\int_0^1 \left\{ f_H + \frac{1-f_H}{V_{H1}} \right\} dp_H = 0$, which is the the analog to equation (20) above.

Figure 1: *Public-good provision levels, marginal income taxes of the low-skilled and net tax payments of the low-skilled, respectively, as a function of the share of low-skilled individuals with a high valuation of public goods.*

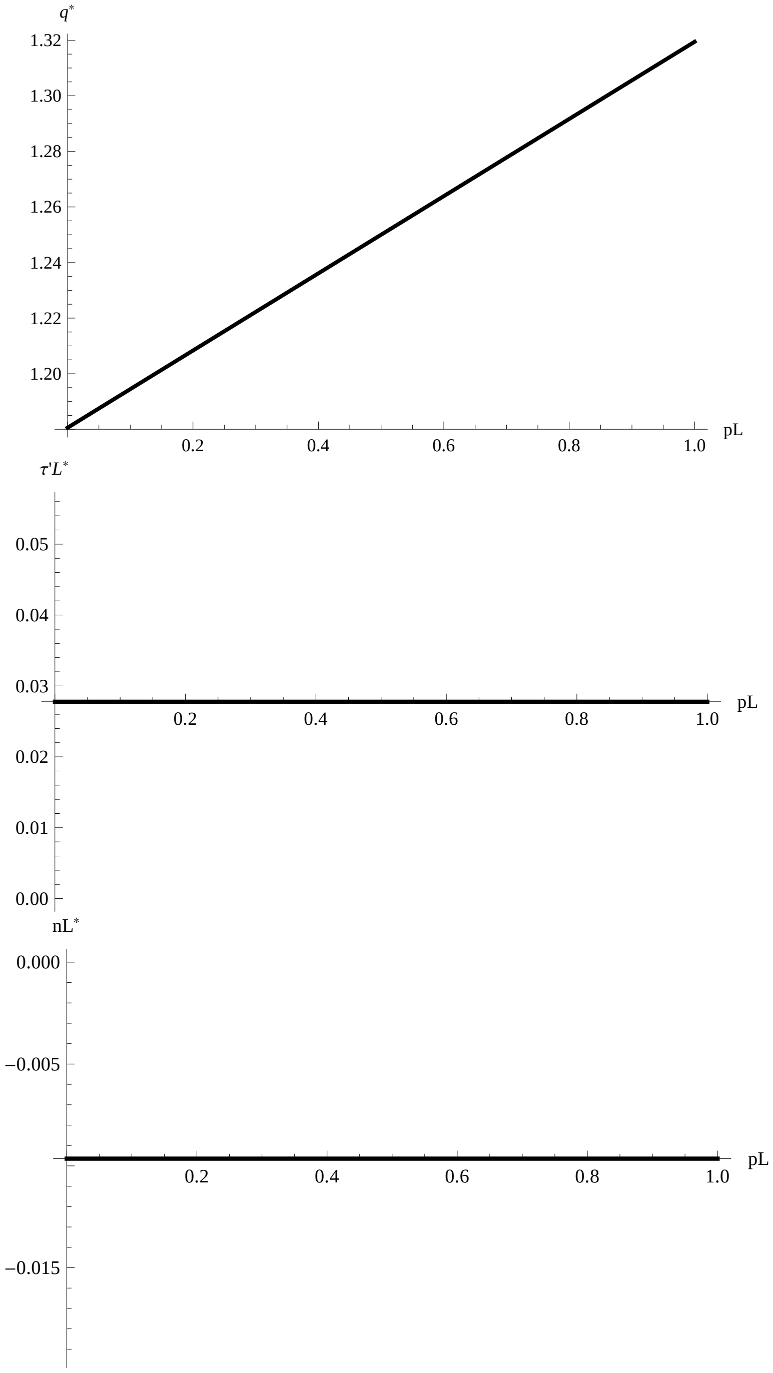


Figure 2: *Public-good provision levels, marginal income taxes of the low-skilled and net tax payments of the low-skilled, respectively, as a function of the population share of high-skilled individuals.*

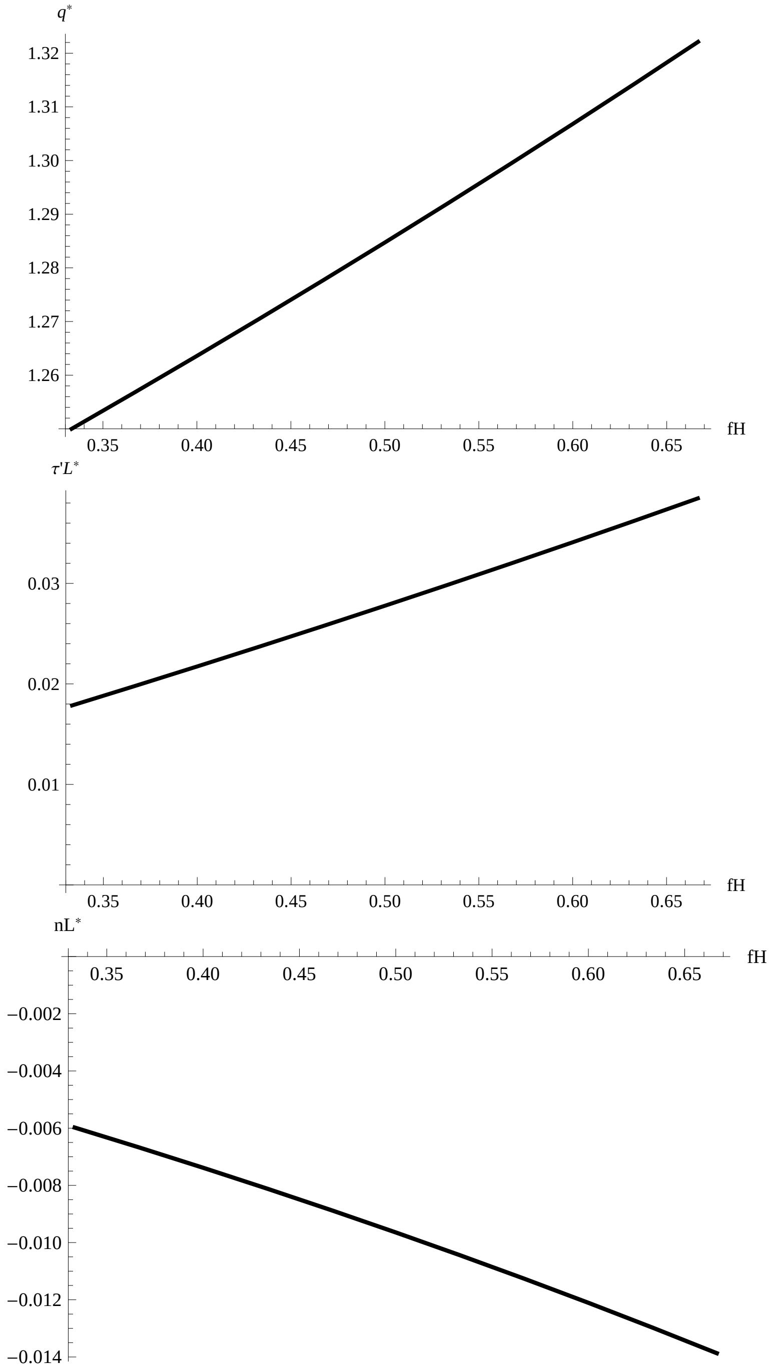


Figure 3: *Public-good provision levels, and utility of low-skilled individuals with a high valuation of public goods as a function of p_L, i.e., the share of low-skilled individuals with a high valuation of public goods.*

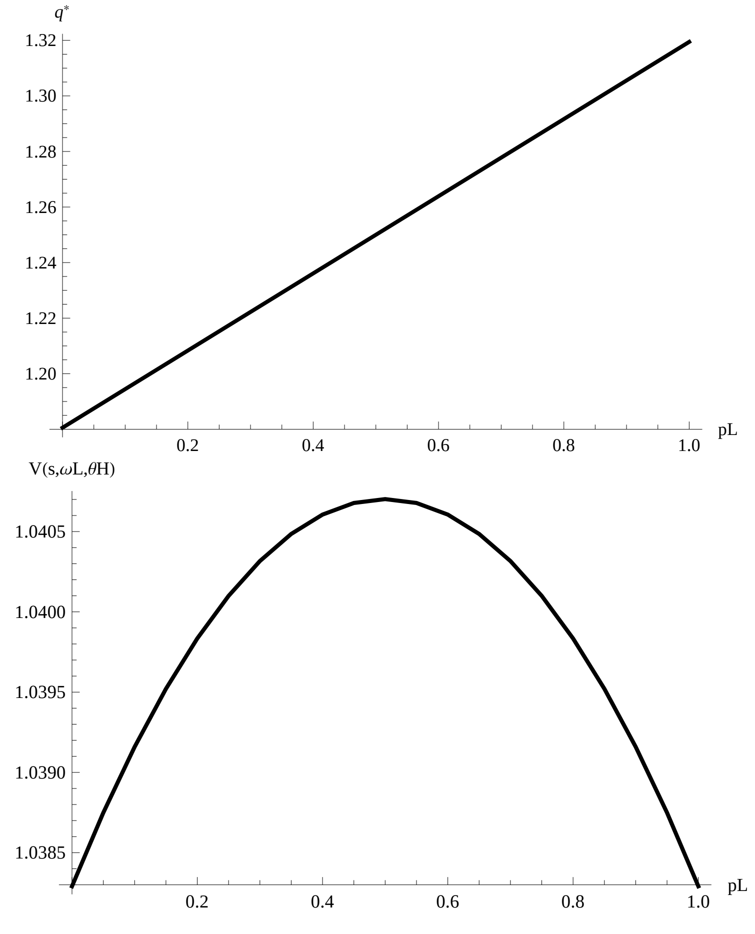
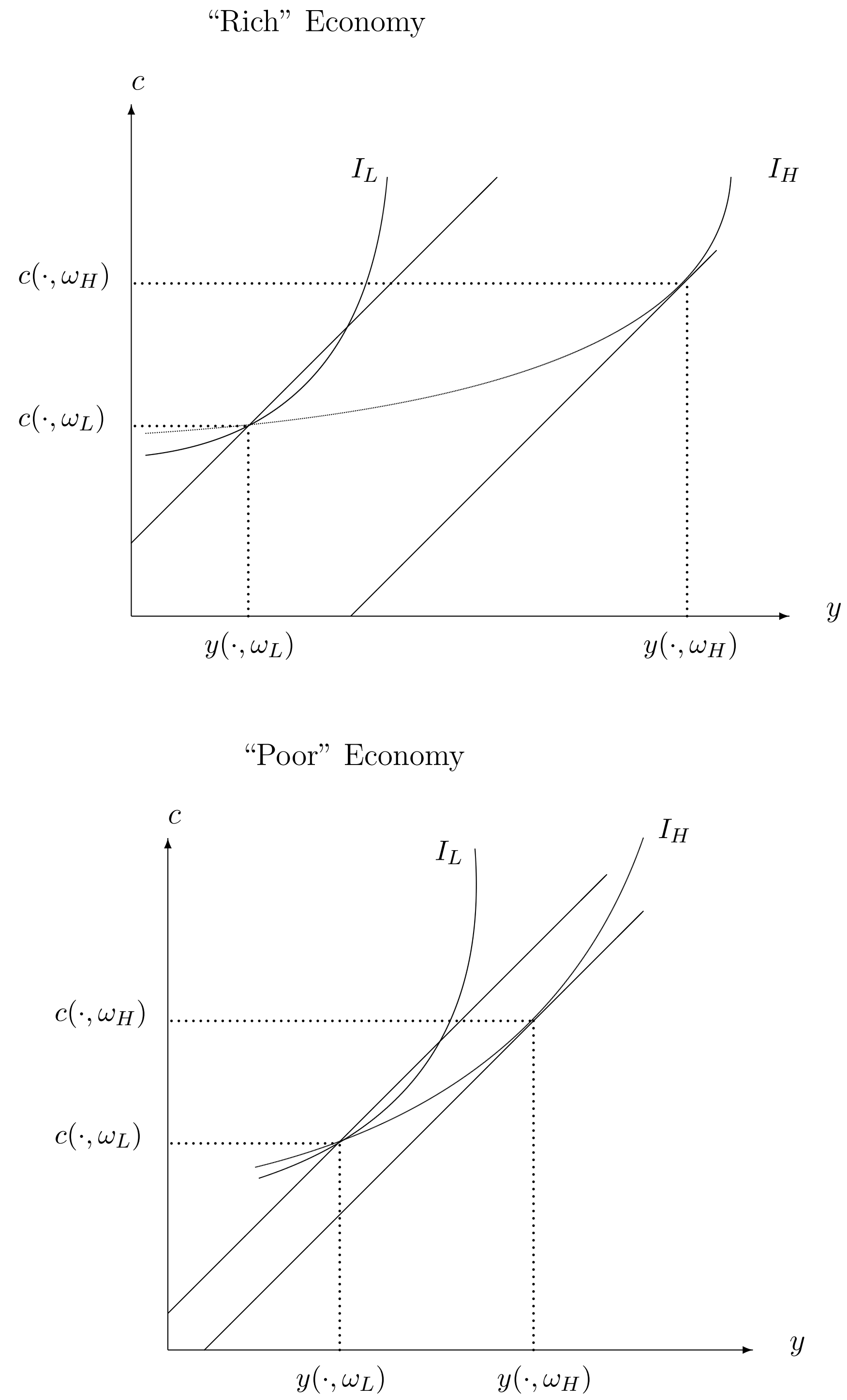


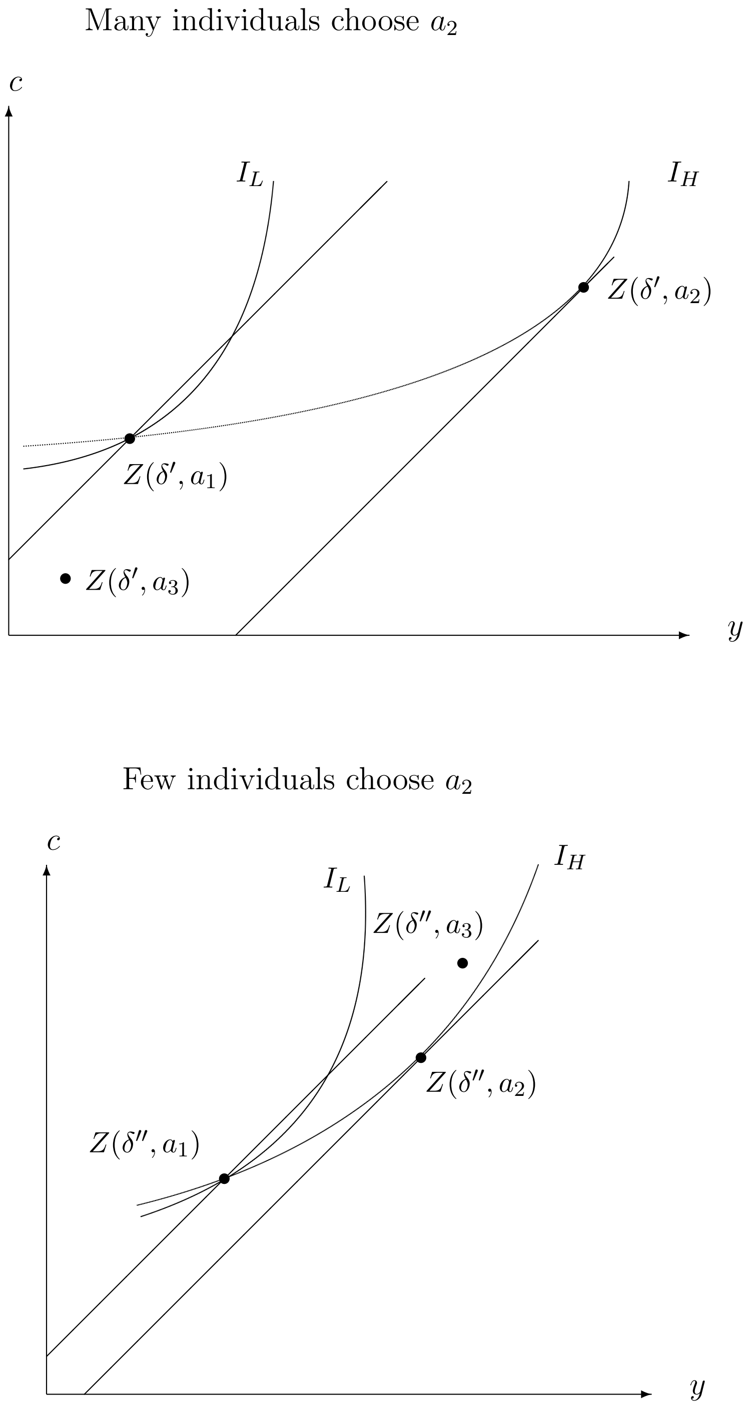
Figure 4: State-dependent redistribution



"Rich" Economy

"Poor" Economy

Figure 5: The non-direct mechanism


Many individuals choose a_2

Few individuals choose a_2

Figure 6: Violation of coalition-proofness

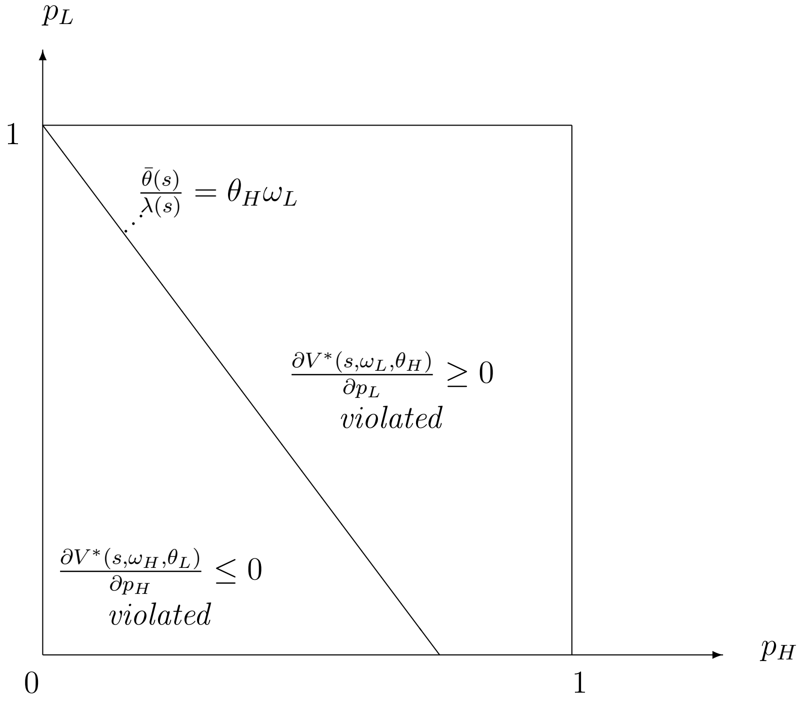

p_L
1
$\frac{\bar{\theta}(s)}{\lambda(s)} = \theta_H \omega_L$
$\frac{\partial V^*(s,\omega_L,\theta_H)}{\partial p_L} \geq 0$
violated
$\frac{\partial V^*(s,\omega_H,\theta_L)}{\partial p_H} \leq 0$
violated
p_H
0
1

Figure 7: *Optimal policy if p_L is unknown. Thin lines represent the Mirrleesian policy and fat lines represent the optimal coalition-proof policy.*

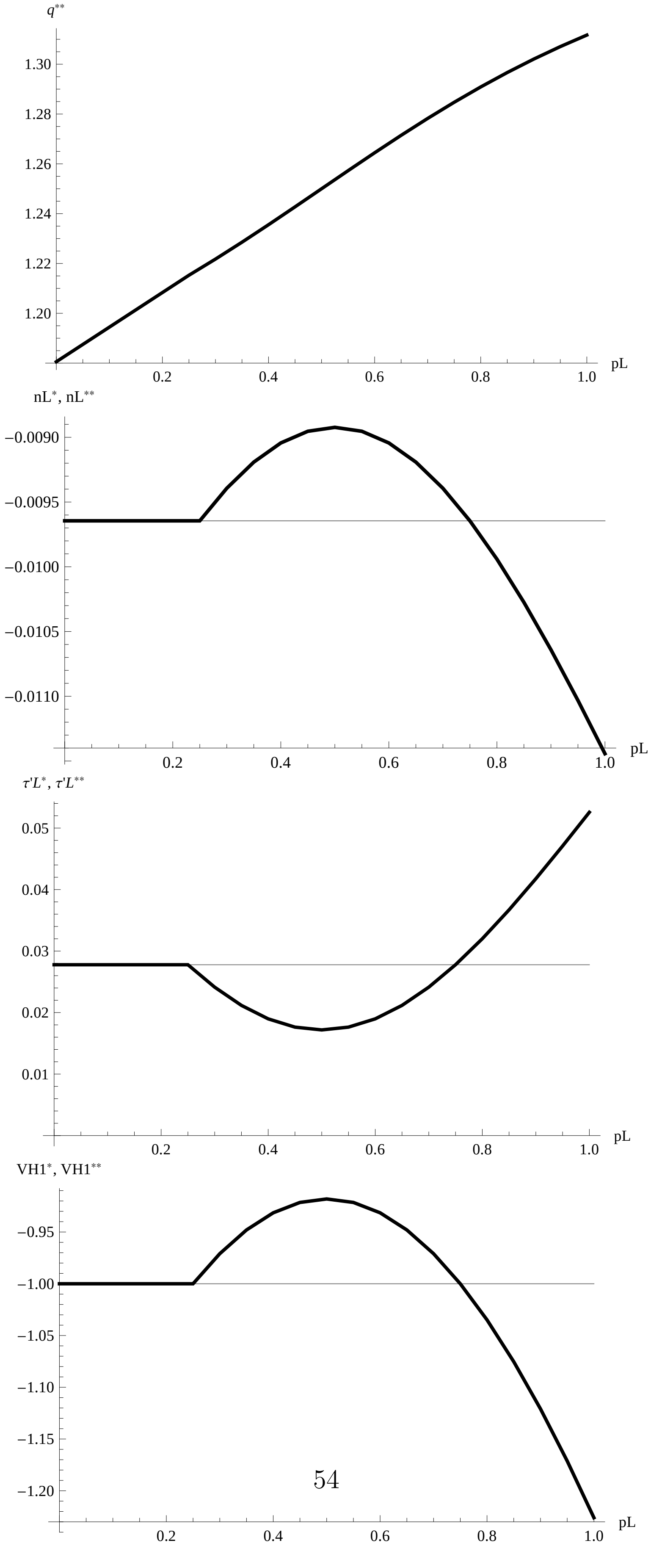
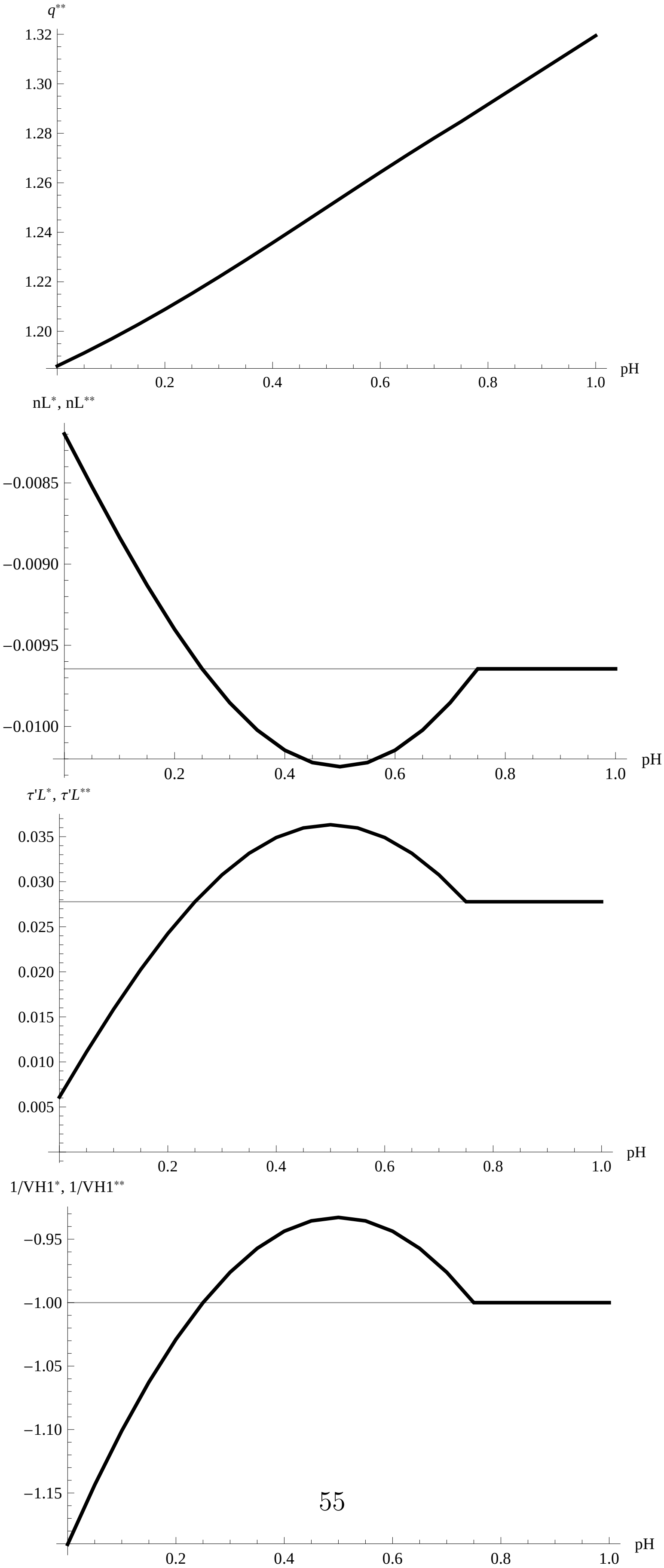


Figure 8: *Optimal policy if p_H is unknown. Thin lines represent the Mirrleesian policy and fat lines represent the optimal coalition-proof policy.*


q**
pH
nL*, nL**
pH
τ'L*, τ'L**
pH
1/VH1*, 1/VH1**
pH