Optimal Taxation with Non-sophisticated Agents *

Pei Cheng Yu †

February 28, 2015

Abstract

This paper studies a model of taxation where the government faces non-sophisticated agents with private information on productivity. Non-sophisticated agents (also known as partially naive agents) are not fully aware of impending changes in their preferences, in contrast to sophisticated agents who can fully forecast their future preference changes. In particular, this paper addresses the problem of inadequate savings resulting from time inconsistent individuals who are partially aware of their present bias. I demonstrate how the government can design an optimal truth-telling incentive scheme to achieve any redistribution outcome despite information asymmetry. In other words, the private information on productivity does not impede the government from implementing the first best allocation. This result holds under very general conditions with agents who have imperfect forecasting abilities, and can be implemented using income specific non-linear savings subsidies. As extensions, I examine other environments where this result would not hold and asymmetric information would distort the set of implementable allocations. This includes economies with diversely naive agents, when the degree of non-sophistication is uncertain, and settings with restrictions on the tax instruments.

Keywords: Optimal taxation, Time inconsistency, Screening, Noncommon priors

JEL Classification Numbers: D03, D62, D82, D84, D86, D91, H21

*I thank David Rahman, Aldo Rustichini, Manuel Amador, Martin Szydlowski, Beth Allen and Jan Werner for many helpful comments. The views expressed herein are those of the authors and not necessarily reflect those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

†University of Minnesota and FRB Minneapolis (e-mail: *yuxxx680@umn.edu*)

1 Introduction

The growing field of behavioral economics has produced mounting evidence suggesting the existence of systematic biases inherent in human behavior. This is a departure from the familiar rational expectations framework. This paper is motivated by the recent experimental and empirical evidence which documents time inconsistent behavior, which may support arguments for government intervention. I aim to incorporate this particular behavioral bias into a normative framework, and examine the optimal tax policy in a Mirrlees setting. I analyze a model of agents with limited cognitive abilities and find significant differences in the optimal tax policy from what is suggested by a model with rational agents and asymmetric information in productivity. In particular, the government can implement the welfare maximizing allocation without any distortions. In other words, the government can implement the same allocations as the allocations in an environment without private information. As a result, the social welfare in my baseline model is better than in an environment with fully rational agents and private information.

The optimal design of labor taxes is a key issue in public finance. Mirrlees [(1971)] introduced a model with asymmetric information on production efficiency in workers, and derived a set of tax policies that could implement the set of truth-telling allocations with the least distortion in effort provision. This paper introduces another layer of concern when considering the appropriate labor tax: agents have time-inconsistent preferences with limited awareness of an impending change in their time preference. In essence, agents are *partially naive*. In light of recent empirical and experimental evidence, I focus on agents with present bias. When agents are partially naive of their time-inconsistency, a government wishing to help the agent ameliorate (for example, to help the agent save enough for retirement) this behavioral bias may inadvertently affect the agents' incentives to work. An agent's cognitive bias adds an additional layer of concern over the efficiency and equity trade-off in a traditional Mirrlees economy. The main contribution of this paper is to describe the interaction between the adverse selection problem and the naiveté problem, and characterize the optimal policy when the government has redistributional motives.

Empirical evidence has shown that the behavioral biases exhibited is non-negligible. For example, DellaVigna and Malmendier [(2006)] studied gym membership data and showed that 80% of monthly gym members would have been better off had they chosen to pay per visit. Several empirical studies have demonstrated the pervasiveness of such biases in a wide array of settings.[1] Skeptics would argue that such behavioral bias may go away

[1]For example, Ausubel [(1999)] and Shui and Ausubel [(2005)] have found similar biases in the credit card market. For more examples, DellaVigna [(2009)] provides an overview of the empirical evidence for behavioral economics.

once we consider more important issues such as retirement savings or investment portfolios. However, evidence has shown the contrary. For example, Madrian and Shea [(2001)] studied participation in the 401(k) plan and found evidence of strong default effects on the participation decision of individuals. O'Donoghue and Rabin [(2001)] show that a model of time-inconsistent individuals with naiveté can help explain the strong influence of the default option on retirement savings.

Evidence of behavioral anomalies suggests that there could be room for implementing paternalistic mechanisms to correct for time inconsistent behavior. Sunstein and Thaler [(2003)] and O'Donoghue and Rabin [(2003), (2006)] have called for the government to implement policies that could help individuals make the *right* choice. However, would implementing such corrective policies influence an individual's economic behavior in other aspects of his/her life? The main interest of this paper is to examine how such policies interact with the incentives of an individual to work.

I find that the optimal set of allocations differs from the Mirrlees allocations. In particular, the government can implement the first best allocation without distorting the labor provision of any agents! More specifically, the main result shows that the government can achieve full redistribution without sacrificing output efficiency. This surprising result is due to the fact that with naive agents, the government can induce efficient labor provision by promising a large payoff in a certain good. However, after the preferences change, the agents no longer value the good they were promised, and would instead prefer the proportion of goods that corresponds to redistribution. In essence, the government can *fool* the agents and does not need to deliver on the promise. As a result, the government can proceed to implement the redistributive policy without paying any information rent. With a well chosen fooling mechanism, the incentive to work is not hampered by the private information of the agents when the agents are naive.

I also consider two types of partial naiveté: magnitude and frequency. I show that for both types of partial naiveté, the government can implement the first best result for any degree of partial naiveté. As a result, there is a discontinuity in welfare with respect to the cognitive abilities of the agents. If agents are sophisticated (agents who are self-aware of their preference change), then they require information rents to induce truth-telling. However, for any degree of naiveté, the government can always implement the desired labor provision and redistribution without transferring any information rents. This is true even when agents are partially naive but very close to being sophisticated. This result is similar to Heidhues and Koszegi [(2010)].

These results apply to a very general setting with agents who experience preference changes and are not accurate in their predictions of these changes. This paper will focus on

the problem of inadequate savings. I model this behavior by adopting the interpretation of time inconsistent behavior and naiveté in O'Donoghue and Rabin [(2001)]. The agents are hit with a temptation to save too little when they need to decide to save and are not fully aware of this when they are making their labor decisions.

I then consider an environment where sophisticated agents coexist with the partially naive agents. In this environment, the government would need to screen for the agents' production efficiency and also their cognitive abilities. The two screening mechanisms would interact with each other and the optimal policies and welfare would differ from our benchmark model without the sophisticated agents. I show that the sophisticated agents with high productive efficiency would receive an information rent, while the naive counterparts do not. As a result, the naive agents are asked to provide more labor and end up consuming less to achieve a more equitable outcome. This result shows that naive agents are akin to fully inelastic agents, so are targeted by governments to raise tax revenue.

The main result of this paper relies on the fact that the government is allowed to deceive the agents by exploiting their naiveté. I consider an extension where the government is not certain about the severity of the taste change the agents will experience. In other words, the government is uncertain about the probability of a taste change or about the degree of the present bias. This limits the willingness of the government to *bet* against the agents' beliefs.

In other extensions of the model, I explore the effects of limiting the government's set of policy instruments. Following Krusell, Kuruscu and Smith [(2010)], I consider linear tax-transfer schemes to provide the agent with the appropriate intertemporal incentives. A linear tax on either savings or current consumption is appealing since it only serves to adjust the market prices while the consumers' consumption choice set remains untouched. Recent work also suggests the adoption of 'minimal' paternalistic policies to minimize government intervention.[2] With a linear tax wedge affecting the intertemporal substitution, the tax will affect the incentives for the agents to work. In the model, the naive agents who are choosing their labor supply are unaware of the fact that the linear tax on savings or consumption serves to help them correct their future temptation of saving too little. Due to this unawareness, the agents will interpret the linear tax as an unnecessary distortion on their incentives to work, and may choose to work at an inefficient level. The government facing heterogeneous agents in production abilities will have to factor in the effect the linear tax has on labor provision when he designs a tax policy to separate the different types of agents. Finally, I consider an extension where the set of implementable allocations is restricted due to political competition.

[2]For example, Sunstein and Thaler [(2003)] have argued for paternalistic policies to correct for biases in behavior that do not coerce agents.

1.1 Related Literature

This paper is closely related to two strands of literature: the optimal taxation literature and the behavioral contracting literature. The paper aims to combine the two fields. Several works have already attempted to analyze the optimal government policies for maximizing the welfare of agents who suffer from temptation and self-control problems. Contrary to exploitative contracting, behavioral public economics aims to find the optimal *paternal* policy. For example, O'Donoghue and Rabin [(2006)] and Gruber and Koszegi [(2004)] have examined the utilization of government policies to curb addictive behavior. O'Donoghue and Rabin [(2003)] suggested using a mechanism design approach to find the most efficient policies when agents suffer from bounded rationality. This paper adopts such an approach.[3] Several papers have also used a similar mechanism design or Ramsey type approach to characterize the optimal paternalistic government policies. A brief introduction of the papers most related to my work is given below.

This paper is closely related to Krusell, Kuruscu and Smith [(2010)] in that their work also study the optimal taxation of consumers who suffer from temptation. They find that the government should subsidize future consumption in an effort to correct the agent's impatience and tendency to save too little. This is in contrast to the no capital taxation result of Chamley-Judd, and is different from the usual no capital taxation in the mean presented in Golosov, Kocherlakota and Tsyvinski [(2003)]. My work differs from Krusell, Kuruscu and Smith [(2010)] in two aspects. Firstly, I consider non-sophisticated agents, while theirs are sophisticated. Secondly, their environment is a complete information one, while I introduce asymmetric information in productive efficiency. I introduce non-sophisticated agents because experimental evidence have found that humans are not fully aware of their own future preferences, and are also not very adept in learning about them. A discussion on the naiveté assumption is given in Section 7.

Amador, Werning and Angeletos [(2006)] have also examined government policies that could help agents with temptation. They study agents who suffer from temptation and are subject to future taste shocks. The government would like to help the agent overcome his temptation problems, but also allow the agent some room to accommodate the stochastic taste shock. However, the tendency to save too little confounds with the unobserved taste shock which creates a trade-off between a commitment policy and a flexible one. They find that a minimum savings rule is optimal. Their work also considers a sophisticated agent, and the adverse selection problem is in the agent's taste shock or marginal utility. The main difference lies in that my work seeks to explore how government policies aimed at helping a

[3] Though this paper adopts a normative framework, it is not meant to support the implementation of paternalistic policies. A discussion of paternalism is provided in Section 7.

boundedly rational agent could distort his labor provision. Therefore, I have a production economy, while Amador, Werning and Angeletos [(2006)] focus on an endowment economy.

A few papers have studied the optimal taxation problem with asymmetric information and quasi-hyperbolic discounting agents. Bassi [(2010)] considers an environment where the hyperbolic discount factor is also non-observable, which creates a two-dimensional screening problem for the government. Guo and Krause [(2015)] study an environment where the government does not have full commitment. These two papers share a common goal with ours, but they all consider sophisticated agents or in settings where naiveté plays no role. This paper is the first to consider the impact of cognitive limitations on the optimal taxation problem.

Several papers have examined taxation models where individuals are differentiated along two or more dimensions. Cremer, Pestieau and Rochet [(2001)] examine a model where both the productivity level and endowments are not observed by the government. Cremer, Pestieau and Rochet [(2003)] extend the model to an overlapping generations setting and endogenize individual endowments as inherited wealth. Beaudry, Blackorby and Szalay [(2009)] examines an economy where agents could participate in both market and non-market production and have different productivity levels for both sectors. The government is unable to observe the productivity levels in both sectors. Most closely related to this paper in terms of the policy issue concerned is Diamond and Spinnewijn [(2011)]. Their paper discusses a model with heterogeneity in both productivity and time preference. This paper is also concerned with the policy on savings, but the heterogeneity lies in the agent's awareness of their underlying present bias. This type of multidimensional screening, where both the skill and cognitive ability of individuals are also not observable, has not yet been analyzed in public policy.[4]

The paper is organized as follows. Section 2 outlines the setup of the model and works out the results for our benchmark model and some preliminary cases (without asymmetric information or without temptation). Section 3 applies the results in section 2 to examine the problem of inadequate savings. Section 4 examines the optimal taxation of diversely naive agents, which includes sophisticated agents. Section 6 explores a model where governments are also not certain about the probability of the event of a taste change. Section 6 explores the two cases where the government is constrained in their set of policy instruments: linear taxes and political constraints. Section 7 discusses the assumption of naiveté, the selection of the welfare criteria and on paternalism. Section 8 concludes the paper. All proofs can be found in the appendix.

[4]It is to the best of the author's knowledge that this is the first paper to discuss multidimensional screening with naiveté in any economic setting.

2 The General Model

Following Spiegler [(2011)], I analyze the optimal allocation under two types of partial naiveté: magnitude naiveté and frequency naiveté. I will show that, regardless of the type of naiveté, a government with redistributive motives can implement the first best allocation in an environment with partially naive agents despite the presence of information asymmetry in a very general setup.

2.1 Setup of General Model

Consider an economy with $|N| \geq 2$ goods produced with labor or other goods and a continuum of agents denoted by the set $I = [0, 1]$. There are M types of agents denoted by the set $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$. The types are distributed according to $\Pr(\theta = \theta_m) = \pi_m > 0$, for all $\theta_m \in \Theta$ with $\sum_{m=1}^{M} \pi_m = 1$.

The production of good n depends on the labor input l_n and the vector amount of other inputs $\mathbf{x}_n \in \mathrm{R}_+^N$. Let $y_n = F_n(\mathbf{x}_n, l_n; \theta_m) \in \mathrm{R}_+$ denote a continuous and differentiable production process strictly increasing in l_n and $\mathbf{x}_n$ of a type m agent for good n. Let $l_n = G_n(y_n, \mathbf{x}_n; \theta_j)$ denote the inverse of $F_n(\mathbf{x}_n, l_n; \theta_j)$ with fixed input $\mathbf{x}_n$. Each type of agent differs in their labor production efficiency: $F_n(\mathbf{x}_n, l_n; \theta_j) > F_n(\mathbf{x}_n, l_n; \theta_k)$, for any labor input $l_n > 0$ and $\mathbf{x}_n$ and good n with $\theta_j > \theta_k$. Therefore, a higher value of θ is associated with higher production efficiency. If a good does not depend on labor input, then it does not depend on θ. The production of at least one of the goods requires labor.

As is standard in Mirrlees taxation, I assume that both the production efficiency of each agent and their labor input $l = (l_1, l_2, \ldots, l_N)$ are not observable by the government. The government can only observe output $y = (y_1, y_2, \ldots, y_N)$ and input x.

2.1.1 Consumer Utility

The agents have the following utility before production

$$U(c, l),$$

where $c = (c_1, c_2, \ldots, c_N)$. We will refer to U as the *ex-ante* utility. The agents' utility changes after production, but before consumption, to

$$V(c; l),$$

where l is sunk. We will refer to V as the *ex-post* utility. Let U and V be continuously differentiable and let it be strictly increasing and concave in consumption: $\frac{\partial U}{\partial c_n} > 0, \frac{\partial^2 U}{\partial c_n^2} < 0$ and $\frac{\partial V}{\partial c_n} > 0, \frac{\partial^2 V}{\partial c_n^2} < 0$. Let U be strictly decreasing and convex in labor: $\frac{\partial U}{\partial l_n} < 0, \frac{\partial^2 U}{\partial l_n^2} < 0$. Finally, for any good $n \in N$, let $\lim_{c_n \to 0} \frac{\partial U}{\partial c_n} = +\infty$ and $\lim_{c_n \to 0} \frac{\partial V}{\partial c_n} = +\infty$ to ensure an interior solution for consumption. Also, for any good $n \in N$, let $\lim_{l_n \to 0} \frac{\partial U}{\partial l_n} = 0$ and $\lim_{l_n \to +\infty} \frac{\partial U}{\partial l_n} = +\infty$ so the labor supply is always strictly positive and finite.

We will assume that the utility from consumption is different: $U \neq V$. More precisely, we assume the marginal rate of substitution for some consumption goods is different for U than for V.

Assumption 1 *There exists $j, k \in N$ such that $\frac{\frac{\partial U}{\partial c_k}}{\frac{\partial U}{\partial c_j}} \neq \frac{\frac{\partial V}{\partial c_k}}{\frac{\partial V}{\partial c_j}}$.*

Assumption 1 along with strictly increasing and concave utility implies a *single crossing condition* on the indifference curves for the ex-ante and ex-post utility of the two goods, j and k. It allows the government to implement policies that would seem attractive to the ex-ante agent while remaining undesirable for the ex-post agent. If the ex-post and ex-ante utility satisfy Assumption 1, then the preference exhibits *taste change*. I will impose an additional standard assumption on the agents' preferences: the marginal rate of substitution between consumption and output is smaller for more efficient agents.

Assumption 2 *For any good $n \in N$ that depends on labor for production, the ex-ante preferences satisfy the single crossing property: $\frac{\partial}{\partial \theta}\left(-\frac{\frac{\partial U}{\partial y_n}}{\frac{\partial U}{\partial c_n}}\right) < 0$*

Assumption 2 makes separation of productivity types optimal for the government. Both Assumption 1 and Assumption 2 will be crucial in the proof of the main result of this paper.

2.1.2 Types of Non-sophistication

The agents are *partially naive*. There are two common ways to model partial naiveté. Loewenstein, O'Donoghue and Rabin [(2003)] and Heidhues and Koszegi [(2010)] have interpreted partial naiveté as the underestimation of the *magnitude* of a taste change. Eliaz and Spiegler [(2006)] have interpreted partial naiveté as the underestimation of the *likelihood* of a taste change. Following Spiegler [(2011)], I will refer to the former as *magnitude naiveté* and the latter as *frequency naiveté*.

Definition 1 *For some $\alpha \in (0, 1]$, agents are partially naive in magnitude if, with probability one, they perceive their ex-post utility to be*

$$W(c, l) = \alpha U(c, l) + (1 - \alpha) V(c, l).$$

Definition 1 defines magnitude naiveté. If $\alpha < 1$, then agents are certain that their preferences will change. However, since α is bounded away from 0, agents underestimate the degree of their taste change.

Definition 2 *Agents are partially naive in frequency if, they believe their ex-post utility to be $V(c,l)$ with probability $1-\alpha$, where $\alpha \in (0,1]$. In other words, let $W(c,l)$ denote the expected ex-post utility of the agent:*

$$W(c,l) = \alpha U(c,l) + (1-\alpha)V(c,l).$$

Definition 2 defines frequency naiveté. In essence, if $\alpha < 1$, the agents attach a positive probability to the likelihood of a change in the preference. However, since α is bounded away from 0, they underestimate the probability of their preferences changing.

Under both definitions, if $\alpha = 1$, the agents are naive and never foresee the preference change. I assume that $\alpha > 0$, so that the agents are are never fully sophisticated.

2.1.3 Timing

The timing of the model is shown in Figure 1.

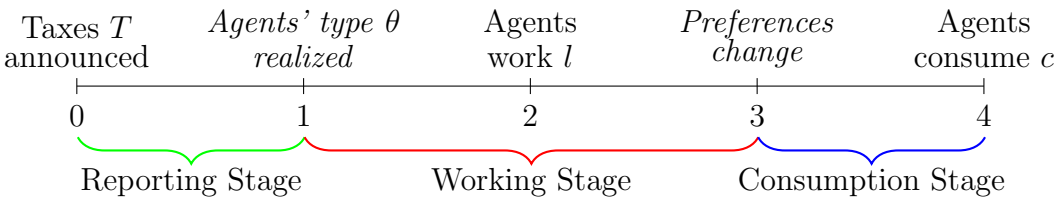

Figure 1: Timing of Events

At date 0, the government designs the tax system. By the law of large numbers, the government knows the measure of each type of agent even before the individual agents learn their productivity types. At date 1, the agent's type is realized. At date 2, the agents report their types and produce according to the tax schedule laid out for their announced types. They make decisions according to the ex-ante utility from date 0 to date 2. At date 3, before the agents make their consumption decisions, their preferences switch to the ex-post utility. At date 4, the agents make their consumption decisions based on the ex-post utility. The three stages are highlighted by the actions of the agent. During the 'choice stage,' the agent chooses the type announcement he reports to the government. During the 'work stage,'

he works according to the type he reported in the previous stage. During the 'consumption stage,' he chooses his consumption level. Notice that the government is restricted to introducing its tax plan before the agents' types are realized.

2.1.4 Tax Instruments

The government is allowed to present a menu of tax options for the agents. In the form of its resulting allocations, the government issues the following menu $\left\{(c^R(\theta_m), c^I(\theta_m), l(\theta_m)\right\}_{\theta_m \in \Theta}$. More concretely, after the tax schedule is announced, partially naive agents of all types 'mentally' choose a set of allocations $(c^I(\theta), l(\theta))$. However, after production, the partially naive agent will 'actually' choose allocation $c^R(\theta)$ to maximize the ex-post utility. The superscript I represents 'imaginary', since the allocation $c^I(\theta)$ is never actually chosen, but were the perceived choices before the preference switch. While the superscript R represents 'reality,' since allocations $c^R(\theta)$ are actually chosen after the preference change, but were not planned before production began.

For both types of partial naiveté, agents do not fully anticipate their preference change before the consumption stage, which makes the tax menu non-redundant. In essence, both 'imaginary' and 'real' allocations matter. However, the imaginary allocations are evaluated differently under the two types of partial naiveté. For magnitude naiveté, agents make their labor decision based on $U(c, l)$ while anticipating a taste change of $W(c, l)$. Therefore, they require $(c^I(\theta), l(\theta))$ to be more appealing than $(c^R(\theta), l(\theta))$ under $W(c, l)$, and the reporting strategy is evaluated using $U(c, l)$. While for frequency naiveté, a agents make their labor decision based on their expected ex-post utility $W(c, l)$. More specifically, they require $(c^I(\theta), l(\theta))$ to be more appealing than $(c^R(\theta), l(\theta))$ under $U(c, l)$, and the reporting strategy is evaluated at $W(c, l)$. Notice that for both types of partial naiveté, the real allocation is more appealing than the imaginary allocation under the ex-post utility.

I would assume that the government has full commitment; in essence, once the tax schedule is announced at date 0, the government is fully committed to carrying out the taxes as promised. Also, the taxation of labor income occurs in the periods when labor decisions are made.

I will follow the primal approach in characterizing the optimal allocations, and then find the tax instruments or political institutions that can implement the optimal allocations.

2.1.5 Welfare Criteria

The government has superior knowledge of the agents' change in preference, and would try to help the agents. Implicitly, I assume the partially naive agents do not draw any

inferences from the policies the government enacts. This is because they do not share the same prior as the government. More specifically, partial naiveté embeds a non-common prior assumption.

The government evaluates allocations at date 0 according to the following welfare criteria

$$\sum_{m=1}^{M} \pi_m \psi \left[U(c^R(\theta_m), l(\theta_m)) \right],$$

where $\left(c^R(\theta_m), l(\theta_m)\right)$ denotes the real allocation for the type m agent. I assume that $\psi \circ U$ is a strictly increasing and concave function, so that government has a redistributive motive. Notice that if $\psi \circ U = U$, then the welfare criteria is utilitarian, just the sum of the agents' ex-ante utility.

Much of the literature on dynamically inconsistent preferences have evaluated welfare with the ex-ante utility. I adopt the ex-ante utility relation as the main welfare criterion because it reflects the agents' long-term planning, while the ex-post utility reflects the agents' short-term temptations. In other words, the ex-post utility is not immune to regret and a benevolent government would consider the adverse implications if the agents give in to their urges. Under such perspectives, the choice of the welfare criteria is non-arbitrary, since the actions undertaken with regard to the ex-post preferences can be regarded as a systematic mistake the agents make, as in Bernheim and Rangel [(2004)]. I will show that the main idea of this paper, which is the government is able to implement first best allocations, is robust to changes in the welfare criteria.

2.2 The Planning Problem

With full commitment by the government, the revelation principle implies the government can focus on a direct mechanism that elicits truth telling. With a slight abuse of notation, I will define $G(y(\theta_{m'}), \mathbf{x}(\theta_{m'}); \theta_m) \in \mathrm{R}_+^N$ as the labor input vector for a type θ_m agent pretending to be a type $\theta_{m'}$ agent.

2.2.1 Planning Problem with Magnitude Naiveté

Under magnitude naiveté, the government's problem is

$$\max_{\{c^R(\theta_m), c^I(\theta_m), l(\theta_m)\}_{m=1}^M} \sum_{m=1}^{M} \pi_m \psi \left[U(c^R(\theta_m), l(\theta_m)) \right], \tag{1}$$

subject to

$$\sum_{m=1}^{M} \pi_m \left[F_n \left(\mathbf{x}_n(\theta_m), l_n(\theta_m); \theta_m \right) - c_n^R(\theta_m) \right] = 0, \forall n \in N. \tag{2}$$

$$U(c^I(\theta_m), l(\theta_m)) \geq U \left[c^I(\theta_{m'}), G(y(\theta_{m'}), \mathbf{x}(\theta_{m'}); \theta_m) \right], \forall \theta_m, \theta_{m'} \in \Theta, \theta_m \neq \theta_{m'}, \tag{3}$$

$$W(c^I(\theta_m), l(\theta_m)) \geq W(c^R(\theta_m), l(\theta_m)), \forall \theta_m \in \Theta, \tag{4}$$

$$V(c^R(\theta_m), l(\theta_m)) \geq V(c^I(\theta_m), l(\theta_m)), \forall \theta_m \in \Theta. \tag{5}$$

The government budget constraint is shown in (2). Constraint (3) is the incentive compatibility constraint. Notice that the agents' decision to report is determined by W evaluated at the imaginary allocation. This is because I require that the agents perceive they would choose the imaginary allocation for any effort level, even if they deviated from truth-telling. Constraints (4) and (5) are the fooling constraints.

In the magnitude naiveté interpretation, the agents are certain that their preference will change, and they anticipate that change by evaluating their reporting strategy using the imaginary allocation. This is because they believe that once their preferences switch to $W(c, l)$, they would prefer the imaginary allocation over the real allocation. However, the agents underestimate the magnitude of the taste change and would instead prefer to choose the real allocation over the imaginary allocation after the preference switch.

2.2.2 Planning Problem with Frequency Naiveté

For frequency naiveté, the government maximizes (1) subject to the government budget constraint (2) and the fooling constraint (5) with the following incentive compatibility constraint $\forall \theta_m, \theta_{m'} \in \Theta, \theta_m \neq \theta_{m'}$,

$$\begin{aligned} &\alpha U(c^I(\theta_m), l(\theta_m)) + (1-\alpha) U(c^R(\theta_m), l(\theta_m)) \\ &\geq \alpha U \left[c^I(\theta_{m'}), G(y(\theta_{m'}), \mathbf{x}(\theta_{m'}); \theta_m) \right] + (1-\alpha) U \left[c^R(\theta_{m'}), G(y(\theta_{m'}), \mathbf{x}(\theta_{m'}); \theta_m) \right], \end{aligned} \tag{6}$$

and the fooling constraint for the imaginary allocation

$$U(c^I(\theta_m), l(\theta_m)) \geq U(c^R(\theta_m), l(\theta_m)), \forall \theta_m \in \Theta. \tag{7}$$

The difference between frequency naiveté and magnitude naiveté lies in the beliefs of the future preference. In frequency naiveté, the agents place a strictly positive probability on their preferences remaining the same. In other words, they believe with some probability that they will choose the imaginary allocation evaluated at the ex-ante preference $U(c, l)$, which is represented in (7). However, in magnitude naiveté, the agents are certain that their

preferences would change, but they under estimate the extent of this shift.

2.2.3 More on the Constraints

If there exists a type m such that at least one of the fooling constraints is non-binding, then $c^I(\theta_m) \neq c^R(\theta_m)$, and the government is *fooling* the type m agent. In other words, the government is exploiting the agents' partial naiveté.

Notice the imaginary allocations are not required to satisfy the government budget constraint. This is because the government only cares about the real allocation, and views the imaginary allocations as an *empty* promise. In other words, the government is certain about the degree of the naiveté and present bias of the agents, so it places no weight on a future where it may need to actually honor the delivery of imaginary allocations. Another concern with the government budget constraint is how the agents do not realize that the aggregate imaginary allocations violate the budget constraint. This is because each agent is infinitesimally small, and even though an agent believes he/she would consume the imaginary allocation, he/she does not consider what other agents believe and how they would behave.

2.3 The Effects of (Partial) Naiveté

I will first analyze the case with fully naive agents ($\alpha = 1$), and show that the first best welfare can be attained. I will then analyze the optimal allocation for both types of partial naiveté.

2.3.1 Fully Naive Agents: $\alpha = 1$

By Assumption 1 and Assumption 2, we can show that the government can achieve the first best allocation. In other words, surprisingly, private information does not matter in an environment where all agents harbor some naiveté.

Proposition 1 *The optimal allocation for the environment where agents have private information on productivity and are fully naive about their preference changes is the same as the allocation in the environment without private information and naiveté.*

Proposition 1 states that the private information problem can be alleviated if the agents are naive. This is because the government can enact policies to fool the agents into believing a particular allocation would be realized in the future, which can provide the necessary incentives for the agents to report truthfully. After their preferences change, the duped agents would find the first best allocation superior to the imaginary allocation. In other words, with

the imaginary allocations, the government is able to provide the information rents necessary for truth-telling. However, these rents are imaginary. After the preference of the agents change, the government is able to implement the first best allocation without paying the information rents. Indeed, it necessarily follows that it is optimal for the government to deceive the agents when the are fully naive regardless of their productivity type.

Corollary 3 *If $\alpha = 1$, it is optimal for the government to fool all types of agents.*

The key to deceiving the agents is to load the rents on the good that they value during the reporting stage, but would not value as much relative to other goods after the preference change. By Assumption 1, suppose $\frac{\partial U}{\partial c_k}/\frac{\partial U}{\partial c_j} > \frac{\partial V}{\partial c_k}/\frac{\partial V}{\partial c_j}$, then the agents value good k more than good j at the reporting stage. The government can then promise more of good k than good j for the imaginary allocations to elicit truthful reports. However, after the preference change, the promise of more good k is less appealing, and the agents would no longer choose the imaginary allocations but the real allocations, with less of good k.

2.3.2 Partially Naive Agents: $\alpha < 1$

I will now show that for partially naive agents of both types of bias, magnitude naiveté or frequency naiveté, for any $\alpha \in (0, 1)$, the government is able to achieve the first best allocation.

Proposition 2 *The optimal allocation for the environment where agents have private information on productivity and are partially naive in magnitude or frequency about their preference changes is the same as the allocation in the environment without private information and partial naiveté.*

The proof and interpretation of Proposition 2 are the same as Proposition 1. It is interesting to note that there is a discontinuity in the optimal welfare with respect to the cognitive limitations of the agents. With magnitude naiveté, the government is able to achieve first best welfare for any $\alpha \in (0, 1]$. In other words, the government is able to devise a mechanism to fool the agents which can eventually implement the first best allocation as long as the agents have some naiveté about their preferences in the future. However, it is well known that with fully sophisticated agents ($\alpha = 0$), the government can only implement the Mirrlees allocations which requires information rent for the productive types. As a result, there is a discontinuity in welfare which was first described in Heidhues and Koszegi [(2010)].

A more surprising result for partial naiveté is that for naiveté in frequency, we also get results similar to naiveté in magnitude. Spiegler [(2011)] has shown the optimal contract to

be continuous with respect to cognitive limitations in a second-degree price discrimination setting for frequency naiveté. However, Proposition 2 shows that this continuity result does not hold in the optimal taxation setting even with naiveté in frequency.

Similar to 3, the government is able to achieve first best welfare with a fooling mechanism regardless of the agents' degree of naiveté and the type of naiveté, magnitude or frequency.

Corollary 4 *If $\alpha < 1$ and agents are partially naive in magnitude or frequency, then it is optimal for the government to fool all types of agents.*

3 The Savings Problem

I will now consider a special case of the general model with N consumption goods. The agents live for two periods. They produce and make consumption and savings decision in the first period, and consume the saved goods in the second period.

Following the usual Mirrlees setup, the production technology is linear, $F(l;\theta_j) = \theta_j l$. Therefore, in a competitive equilibrium, the wages are equated to the marginal productivity of labor. There is also a storage (savings) technology that transfers one unit of good in the first period to one unit of second period good. (Alternatively, they have access to a bond with interest rate 0.)

The agents have the following ex-ante utility

$$U(c,k,l) = u(c) - h(l) + w(k).$$

They face the following ex-post utility function, where l is taken as given

$$V(c,k,l) = u(c) - h(l) + \beta w(k).$$

The period utilities are continuously differentiable and satisfy the usual concavity assumptions $u', -u'' > 0$ and $w', -w'' > 0$, while the dis-utility from labor satisfies $h', h'' > 0$. Also, $\lim_{c\to 0} u'(c) = +\infty$ and $\lim_{k\to 0} w'(k) = +\infty$, so consumption in both periods will be strictly positive.

I will focus on the case with present bias, where $\beta < 1$. We can interpret β as measuring the degree or severity of the present bias. A smaller β represents a stronger bias for present consumption. I will refer to β as measuring the degree of temptation the agents sufferer from. Following O'Donoghue and Rabin [(2001)], a partially naive agent in magnitude perceives his degree of present bias to be $\hat{\beta} \in (\beta, 1]$ before reporting his type. Notice if $\hat{\beta} = 1$, then the agent is naive and unaware of his present bias. Similar to the general model, the perceived

present bias is always strictly greater than the actual present bias, $\hat{\beta} > \beta$, so the agents are not sophisticated.

In the present setup, the present bias is similar to a temptation shock that the agent does not foresee perfectly. With a simple transformation, the model is similar to a quasi-hyperbolic model with partial naiveté

$$\begin{aligned} U_1(c,k,l) &= -\tilde{h}(l) + \hat{\beta}\delta\left[u(c) + \delta w(k)\right], \\ U_2(c,k) &= u(c) + \hat{\beta}\delta w(k), \\ U_3(k) &= w(k), \end{aligned}$$

where $\delta = 1$ and $h(\cdot) = \frac{1}{\hat{\beta}}\tilde{h}(\cdot)$. Agents live for three periods. In the first period, agents choose their labor supply. In the second period, agents make consumption and savings decision. Finally, in the third period, agents consume their retirement savings. If $\hat{\beta} = \beta$, then the agents are sophisticated and the transformed model is similar to Laibson [(1997)], and if not, then it is similar to the model with cognitive limitations as presented in O'Donoghue and Rabin [(2001)][5].

A partially naive agent in frequency believes that his preferences change to β with probability $1-\alpha$ and it would stay the same with probability α. If $\alpha = 1$, then the agent is naive, and if $\alpha = 0$, then the agent is fully sophisticated. A partially naive agent corresponds to a belief where $\alpha \in (0,1)$.

Notice the savings problem is a simplified setup of the general problem with preference changes. The government's planning problem is also similar. Note that Assumption 1 and Assumption 2 are automatically satisfied.

Corollary 5 *It is optimal to fool all productivity types. The optimal allocation for the environment with private information and (partial) naiveté is the same as the allocation in the first best environment without private information and* $\beta = 1$.

To demonstrate how Corollary 5 works, consider an economy with two productivity types $\Theta = \{\theta_b, \theta_g\}$, where $\theta_g > \theta_b$, and let the government be utilitarian, so that $\psi \circ U = U$. Let $\left(c^R, k^R, l\right)$ denote the first best allocation. Since $\psi \circ U = U$, the first best allocation is equated across types, and the marginal cost of effort is equated to the marginal benefit of consumption. Let us examine the fully naive case, so there is no distinction between

[5] A three period quasi-hyperbolic discounting model where the agents work for the first three periods is presented in Appendix B. The main arguments still hold.

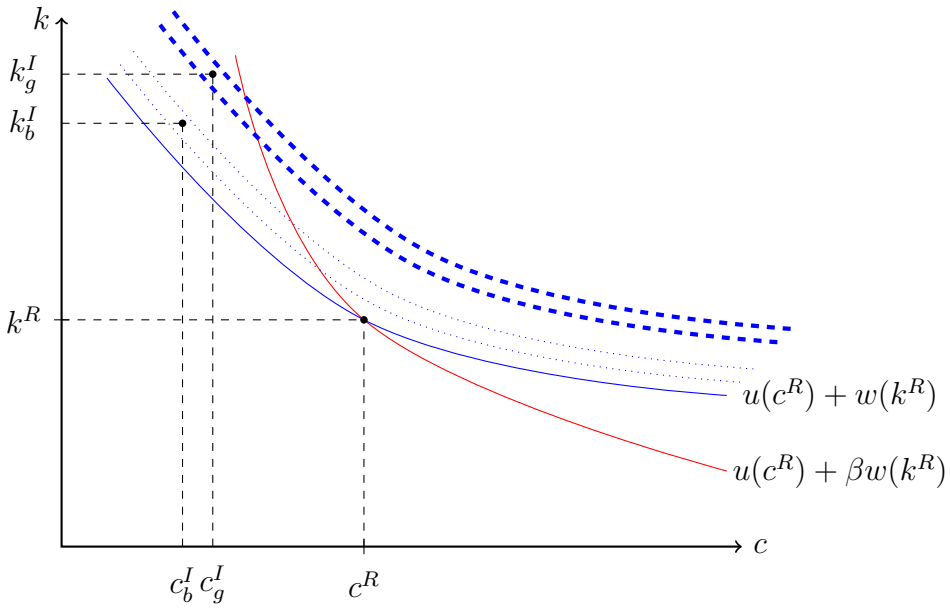

Figure 2: Indifference Curves

magnitude and frequency naiveté. The incentive compatibility constraints are

$$u(c_g^I) - h(l_g) + w(k_g^I) \geq u(c_b^I) - h\left(\frac{\theta_b l_b}{\theta_g}\right) + w(k_b^I), \tag{8}$$

$$u(c_b^I) - h(l_b) + w(k_b^I) \geq u(c_g^I) - h\left(\frac{\theta_g l_g}{\theta_b}\right) + w(k_g^I), \tag{9}$$

and the fooling constraints are

$$u(c_g^R) + \beta w(k_g^R) \geq u(c_g^I) + \beta w(k_g^I), \tag{10}$$

$$u(c_b^R) + \beta w(k_b^R) \geq u(c_b^I) + \beta w(k_b^I), \tag{11}$$

$$u(c_g^I) + w(k_g^I) \geq u(c_g^R) + w(k_g^R), \tag{12}$$

$$u(c_b^I) + w(k_b^I) \geq u(c_b^R) + w(k_b^R). \tag{13}$$

In figure 2, the solid blue curve represents the indifference curve of the ex-ante utility at allocation (c^R, k^R). The solid red curve represents the indifference curve of the ex-post utility at allocation (c^R, k^R). The imaginary allocations have to be in the area bounded by the solid line indifference curves. Any allocation within this area satisfies the inequalities (10), (11), (12) and (13). Furthermore, the incentive compatibility constraints, (8) and (9), provide an upper and lower bound to the difference in ex-ante utility of the two types of

agents. In essence,

$$h\left(\frac{\theta_g l_g}{\theta_b}\right) - h(l_b) \geq \left[u(c_g^I) + w(k_g^I)\right] - \left[u(c_b^I) + w(k_b^I)\right] \geq h(l_g) - h\left(\frac{\theta_b l_b}{\theta_g}\right).$$

Therefore, given the first best labor provision, the imaginary allocations have to be within the dashed indifference curves, where the good type's imaginary allocation is within the bold dashed area, and the bad type's is within the light dotted area.

3.1 Implementation: Income Taxation and Retirement Savings

To implement the first best allocation in this environment, the government can rely on savings subsidies that are non-linear and productivity specific. Consider the fully naive case with two productivity types and an utilitarian government. Denote the first best allocation as $(c^*, k^*, l_m^*)_{m\in\Theta}$, which are the real allocations the government wishes to implement. They satisfy the following marginal conditions: intertemporal substitution $u'(c_m^*) = w'(k_m^*)$, full insurance across types $c_g^* = c_b^*$ and $k_g^* = k_b^*$, and intratemporal substitution $u'(c_m^R) = \frac{1}{\theta_i} v'(l_m^*)$. It also satisfies the government budget constraint: $\pi_g \theta_g l_g^* + \pi_b \theta_b l_b^* = c^* + k^*$. Therefore, the real savings subsidies is the same for all types and it is chosen to smooth consumption across periods optimally:

$$1 + \tau^* = \beta.$$

The transfers have to satisfy the government budget constraint: for the high productivity agents,

$$T_g^* = \pi_b(\theta_g l_g^* - \theta_b l_b^*) + (1-\beta)k^*,$$

and for the low productivity agents,

$$T_b^* = -\pi_g(\theta_g l_g^* - \theta_b l_b^*) + (1-\beta)k^*.$$

The government can select any imaginary allocation that satisfies the fooling and incentive compatibility constraints, say $\left(c_m^I, k_m^I\right)_{m\in\Theta}$. It can proceed to pin down the imaginary savings subsidy

$$1 + \tau_m^I = \frac{w'(k_m^I)}{u'(c_m^I)}.$$

Using the savings subsidy, it can easily find the imaginary transfers

$$T_m^I = \theta_m l_m^* - \left(c_m^I + (1+\tau_m^I)k_m^I\right).$$

As a result, agent m faces the following policy menu: $\left\{\left(\tau_m^I, T_m^I\right); (\tau^*, T_m^*)\right\}$.

From the derivation above, the implementation typically involves non-linear savings subsidies, $\tau_m^I \neq \tau^*$. Such non-linearities is already prevalent in the present tax system. For example, a feature of IRA and 401(k) accounts is that annual contribution are capped: rate of return below cap is higher than the rate of return above the cap, which creates a non-linear intertemporal budget constraint. The model suggests that more elaborate or complicated retirement savings tax systems may improve both consumption smoothing and redistribution. For example, if the government selects the imaginary allocations such that $c_m^I < c^*$ and $k_m^I > k^*$, then it is possible that the rate of return below a certain cap is lower than the rate of return above the cap. More importantly, the model suggests that savings subsidy that differs for each productivity level can help screen the agents.

Furthermore, the implementation can also utilize insights from Thaler and Benartzi [(2004)] by exploiting the tendency for agents to exhibit status quo bias. The same behavioral bias that results in inadequate savings can cause procrastination, which leads to inertia or status quo bias. In the context of my model, the default taxes are set at (τ^*, T_m^*), with the option of changing to $\left(\tau_m^I, T_m^I\right)$ post-production for agent m. Though exploiting the possible status quo bias is not needed in my model, it makes sense to set the default to (τ^*, T_m^*) for implementation.

Finally, notice that this implementation adheres to the rules of libertarian paternalism[6] because the freedom of choice is not compromised in this setup. The agents are allowed to choose to consume at the imaginary allocations, but would not. A detailed discussion of paternalism is provided in Section 7.

4 Model with Diversely Naive Agents

The previous model had agents differing in their production efficiency while sharing the same cognitive features. In this section, I consider a setting where governments are facing dynamically inconsistent agents who differ in their cognitive abilities.

It is easy to show that our results from the previous section still applies to a setting with diversely naive agents if all agents are non-sophisticated. In other words, as long as all agents are bounded away from sophistication, the government is still capable of achieving the first best allocations. To see this, consider a population of magnitude naive agents with perceived present bias $\hat{\beta}$ distributed within the boundaries $\left[\underline{\beta}, \overline{\beta}\right]$, where $\underline{\beta}$ is strictly greater than the true level of present bias β. The government can target the least naive agents, agents with

[6]Sunstein and Thaler [(2008)] have argued for 'libertarian paternalist' policies that would 'nudge' individuals to choosing the appropriate course of action without compromising the freedom of choice.

preceived present bias $\underline{\beta}$, and provide them with the appropriate incentives to reveal their productivity types. A separating mechanism for $\underline{\beta}$ will also work for any $\hat{\beta} > \underline{\beta}$, regardless of the joint distribution of productivity and cognitive limitation. This is because providing incentives for the least naive agents for truth-telling is the most difficult, so any incentives that could separate the productivity of the least naive agents will also be truth-telling for more naive agents. This also holds for agents with naiveté in frequency. As a result, for the rest of this section, we will be focusing on a population that also contains sophisticated agents.

Since both productivity and degree of naiveté of the agents are unobserved by the government, the optimal policy has to solve a multidimensional screening problem. The government would like to know which agents are productive so they could be encouraged to produce more. However, the form of the incentive scheme would depend on whether the agents are sophisticated or not. As was already shown, non-sophisticated agents can be manipulated into producing the appropriate level of output without any costs. However, sophisticated agents will be immune to manipulation and deception and will require actual information rents for full revelation of their types.

I will assume (without loss of generality) that agents are either sophisticated or fully naive, $\hat{\beta} \in \{\beta, 1\}$. Let $\Theta = \{\theta_b, \theta_g\}$ with $\theta_g > \theta_b$. Let $\Pi\left(\theta_i, \hat{\beta}\right)$ denote the joint distribution of productivity and naiveté. For simplicity, let π_m^j denote the measure of a type $m \in \{b, g\}$ agent in terms of productivity and type $j \in \{n, s\}$ in terms of sophistication where n is represents 'naive' and s represents 'sophisticated'.

It is not possible to deceive the sophisticated agents, while the government would like to do so with the naive agents. Therefore, the government issues the following menu $\left\{\left(c_m^s, k_m^s, l_m^s\right); \left[\left(c_m^R, k_m^R\right), \left(c_m^I, k_m^I\right), l_m^n\right]\right\}_{m\in\{b,g\}}$. The naive agents will choose the option with the imaginary allocations, thinking that they will be consuming it, but will instead end up consuming the real allocations. The sophisticated agents will not be deceived by the presence of the imaginary allocations, so the government needs to provide enough incentives such that they won't be lured by the real allocations. In other words, the extent of redistribution for the naive agents is severely limited with the presence of sophisticated agents. Notice that it is not required that agents of the same skill level produce the same amount of output.

The incentive compatibility constraints for the productive sophisticated agents are as follows

$$u(c_g^s) - h(l_g^s) + w(k_g^s) \geq u(c_b^s) - h\left(\frac{\theta_b l_b^s}{\theta_g}\right) + w(k_b^s), \tag{14}$$

$$u(c_g^s) - h(l_g^s) + w(k_g^s) \geq u(c_m^R) - h\left(\frac{\theta_m l_m^n}{\theta_g}\right) + w(k_m^R), \forall m \in \{b, g\}. \tag{15}$$

Inequality (14) is the usual incentive compatibility constraint that appears in the Mirrlees setting, where the productive type would not want to mimic the unproductive type. Inequality (15) prevents the sophisticated individual from pretending to be naive. The same incentive compatibility constraints also appear for the unproductive sophisticated agents

$$u(c_b^s) - h(l_b^s) + w(k_b^s) \geq u(c_g^s) - h\left(\frac{\theta_g l_g^s}{\theta_b}\right) + w(k_g^s),$$

$$u(c_b^s) - h(l_b^s) + w(k_b^s) \geq u(c_m^R) - h\left(\frac{\theta_m l_m^n}{\theta_b}\right) + w(k_m^R), \forall m \in \{b, g\}.$$

The incentive compatibility constraints for the naive agents are the same as (8) and (9). The fooling constraints for the naive agents are the same as (10), (11), (12) and (13).

Similar to the previous sections, the government's ability to exploit the naive agent helps ease the pressure on the incentive compatibility constraints for the naive agents.

Lemma 6 *The government will fool the naive agents*

By Lemma 6, the productivity level of the naive agents can be elicited by the government without cost. However, the government must provide the sophisticated agents with information rent to prevent them from pretending to be naive. This concept is demonstrated in the following proposition, which characterizes the main result for diversely naive agents.

Proposition 3 *The optimal allocation* $\left\{(c_m^s, k_m^s, l_m^s); \left(c_m^R, k_m^R, l_m^n\right)\right\}_{m\in\{b,g\}}$ *has the following properties*

i. All types smooth consumption over time optimally. In essence, $u'(c_m^s) = w'(k_m^s)$ *and* $u'(c_m^R) = w'(k_m^R)$ *for all* $m \in \{b, g\}$.

ii. The high productivity agents consume more than the low productivity agents. In essence, $c_g^s > c_b^s, c_g^R > c_b^R$ *and* $k_g^s > k_b^s, k_g^R > k_b^R$.

iii. The sophisticated high productivity agents consume more than the naive high productivity agents: $c_g^s > c_g^R$ *and* $k_g^s > k_g^R$

iv. The sophisticated low productivity agents consume weakly less than the naive low productivity agents: $c_b^s \leq c_b^R$ *and* $k_b^s \leq k_b^R$

v. The low productivity agents produce too little: $u'(c_b^j) > \frac{1}{\theta_b} h'(l_b^j)$, *while for the high productivity agents:* $u'(c_g^j) = \frac{1}{\theta_g} h'(l_g^j)$, *for all* $j \in \{n, s\}$,

vi. The naive agents produce weakly more than the sophisticated agents of the same productivity level and strictly more for the high productivity agents. In essence, $y_m^n \geq y_m^s$ *for all* $m \in \{b, g\}$*, and the inequality is strict if* $m = g$*.*

The first part of Proposition 3 states that the government is able to help the agents overcome their self-control problem regardless of whether they are aware of the underlying present bias. For the naive agents, the government can implement a policy akin to the one in the previous section to encourage savings. The sophisticated agents are already aware of their temptations and are seeking a commitment device to help combat this problem, which the benevolent government provides.

Part (ii) of Proposition 3 is a result of the incentive compatibility constraints. The productive sophisticated agent is tempted to mimic a naive or sophisticated agent with low productivity. As is the case for most adverse selection problems, the optimal mechanism for screening is to decrease the consumption level of the agents with low productivity. Even though it is cost-less to screen the naive agents, the optimal mechanism would have to discourage the sophisticated agents from pretending to be naive. Therefore, it is not possible to achieve full redistribution in this environment. The same mechanism drives the result in part (v), which is a standard result from the optimal taxation literature.

Part (iii) and part (iv) of Proposition 3 compares the consumption level across cognitive abilities given the same productivity level. The sophisticated high productivity agents work less and consumes more than their naive counterpart because the government needs to provide them with an information rent to discourage pooling with the low productivity individuals, while the naive agents do not require such rent.

An interesting feature of Proposition 3 is parts (iv) and (vi) which demonstrates the existence of two possible equilibrium. One of the equilibrium, regime 1, is full separation, where each of the four agents consume and work different amounts. The other equilibrium, regime 2, has pooling at the bottom, where the high productivity agents are separated from the low productivity agents but separation in cognitive ability only happens for the high productivity types. In other words, for regime 2, the equilibrium has sophisticated and naive low productivity agents consuming and working the same amount.

There are two possible equilibrium regimes because one of the incentive compatibility constraints may or may not bind. The sophisticated agents are sentient and aware of the chicanery the government wishes to implement. As a result, the government may need to provide the sophisticated agents of both high and low productivity sufficient incentives to avoid pooling with the naive agents.

Figure 3 demonstrates the potential direction of misreports for each agent. A solid arrow pointing from type $(\theta, \hat{\beta})$ to type $(\theta', \hat{\beta}')$ means that type $(\theta, \hat{\beta})$ *is* indifferent between

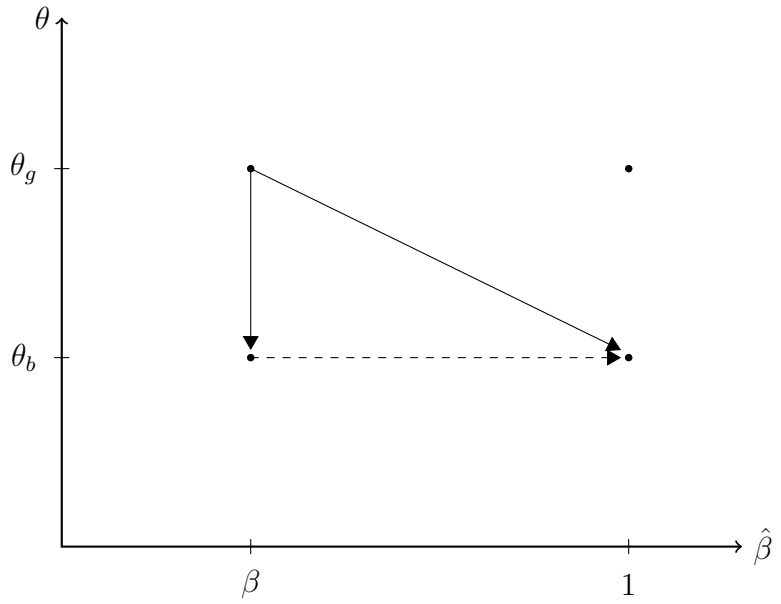

Figure 3: Binding Incentive Constraints

reporting truthfully and reporting to be type $(\theta', \hat{\beta}')$. A dotted arrow pointing from type $(\theta, \hat{\beta})$ to type $(\theta', \hat{\beta}')$ means that type $(\theta, \hat{\beta})$ *might be* indifferent between reporting truthfully and reporting to be type $(\theta', \hat{\beta}')$. Intuitively, it is the downward incentive compatibility constraints that are active for the result of Proposition 3. In other words, the government needs to prevent the sophisticated high productivity agents from misreporting to be a low productivity agent of either sophistication level. The government also needs to prevent the sophisticated low productivity agents from pretending to be naive low productivity agents.

Figure 3 shows that there are two possible equilibrium, which depends on whether the incentive constraint for preventing the sophisticated low productivity type from pretending to be the naive counterpart is binding. Figure 4 and figure 5 show the allocations of the two regimes on the consumption-output indifference curve. The horizontal axis represents the output and the vertical axis represents the utility from consumption $M = u(\cdot) + w(\cdot)$.

A sufficient condition for partial pooling is to assume that $h'''(\cdot) > 0$. This is because it is not optimal to require the naive low productivity agent to produce more, since the effort cost from doing so would be too formidable. Full separation can only occur if the change in the disutility from labor is sufficiently small and if π_b^n is small. This is because the naive unproductive type is producing more at the expense of his own utility, which wouldn't happen if a large enough population of agents are naive and unproductive.

If sophisticated agents are present, redistribution is limited and the first best outcome is not achievable. The naive agents do not receive information rents, so they are weakly worse off than their sophisticated counterparts. It is even possible for the sophisticated

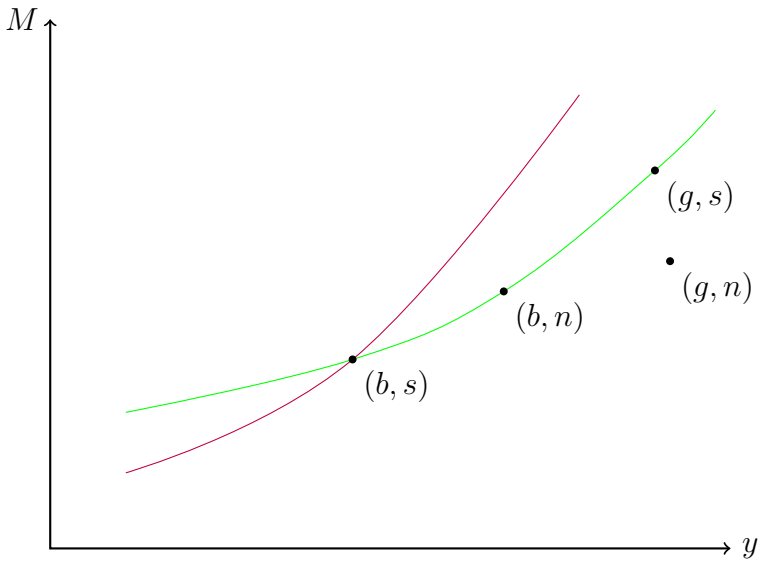

Figure 4: Regime 1

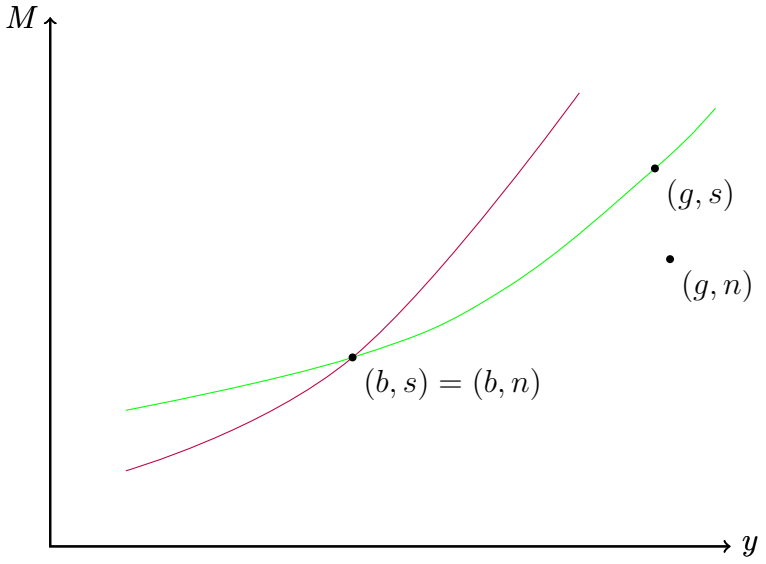

Figure 5: Regime 2

low productivity agents to receive information rents to discourage them from imitating a naive low productivity agent. This result shows that agents who hold the incorrect beliefs (naive) and are dogmatic (do not infer from the government's policies) are akin to inelastic individuals, so they should be the target of higher government taxes.

4.1 Implementation

[To be added]

5 Model with Government Uncertainty

In the previous sections, I have assumed that the government knows the present bias of the agents. This is an extreme assumption. Various studies have tried to estimate the quasi-hyperbolic discounting model and have arrived at vastly different estimations.[7] Models with time-inconsistent individuals have also demonstrated how the optimal policy is extremely sensitive to the values of the hyperbolic discount rates.[8] As a result, it is more natural to assume a government that is uncertain whether a taste change would occur or the degree of the present bias.

[To be added]

6 Optimal Policies with Restrictions

In the previous sections, the government was able to fool the agents with impunity. However, in the real world, there are several concerns against deceiving the agents. I will explore some of these concerns in this section. Namely, I would discuss the implications of implementing linear taxes, since the literature has adopted a positive view of utilizing linear taxes to as a way of introducing paternalistic measures, as in Krusell, Kuruscu and Smith [(2010)]. I also introduce political economy concerns, where the agents could potentially revolt or vote the incumbent out of office if they were deceived.

6.1 Linear Taxes

Following a recent trend in the literature, the government's policies are *minimal* in the sense that it chooses to indirectly affect the savings decisions of the agents through an direct linear tax (dependent on revealed productivity type) on consumption or savings. In other words, an important caveat is the decentralization of an agent's consumption-savings decision. As opposed to a more intrusive policy such as a minimum savings rule in Amador, Werning and Angeletos [(2006)].

For illustrative purposes, I will focus on a model with two productivity types $\Theta = \{\theta_b, \theta_g\}$ and fully naive agents, so the type of cognitive bias, magnitude or frequency, does not matter. I will first discuss the tax policy the government enacts to fix the intertemporal temptation

[7] Frederick, Loewenstein and O'Donoghue [(2002)] surveys a variety of studies that attempt to estimate the quasi-hyperbolic model.

[8] O'Donoghue and Rabin [(2006)] showed how the optimal sin taxes can differ wildly depending on the assumed value of β in the $\beta - \delta$ model. Laibson, Repetto and Tobacman [(1998)] showed how the welfare implications of a tax-deferred defined contribution retirement savings plan can vary with the assumed hyperbolic discount factor.

problem. The government would choose a linear tax on either consumption or savings, which is based on $V(c,k,l)$ to correct for the temptation problem. For type m agent, given any after-tax earnings Y_m, he would solve the following consumption-savings problem

$$\max_{c_m^R, k_m^R} u(c_m^R) + \beta w(k_m^R), \tag{16}$$

subject to

$$c_m^R + (1+\tau_m^k)k_m^R = Y_m,$$

where τ_m^k is the savings tax that depends on the reported type m. Notice that (16) takes into account the temptation the agents suffer after their earnings from labor have been realized. The agent chooses the consumption and savings level based on his budget constraint and the first order condition: $\left(1+\tau_m^k\right) u'(c_m^R) = \beta w'(k_m^R)$.

Hence, to correct for the present bias, the government can set the savings tax as

$$\tau_m^k = \beta \frac{w'(k_m^R)}{u'(c_m^R)} - 1, \forall \theta_m \in \Theta, \tag{17}$$

or alternatively, the consumption tax as

$$\tau_m^c = \frac{u'(c_m^R)}{\beta w'(k_m^R)} - 1, \forall \theta_m \in \Theta, \tag{18}$$

which are both evaluated using the real allocations. The optimal intertemporal taxes can then be derived using the primal approach by substituting out the taxes using (17) or (18). For a consumption tax, the implementability constraint is

$$\frac{u'\left(c_m^R\right)}{\beta w'\left(k_m^R\right)} c_m^R + k_m^R = Y_m, \forall \theta_m \in \Theta.$$

While for savings tax, the implementability constraint is

$$c_m^R + \frac{\beta w'\left(k_m^R\right)}{u'\left(c_m^R\right)} k_m^R = Y_m, \forall \theta_m \in \Theta.$$

Following Krusell, Kuruscu and Smith [(2010)], I will choose to focus on a policy that subsidizes savings.

Similarly, the government will also implement tax policies to create efficient labor provision when there is private information in production efficiency. Since labor decisions are made before the present bias sets in, the government must design a truth-telling mechanism based on $U(c,k,l)$ to separate the two types of workers.

With fully naive agents, the following incentive compatibility constraints are evaluated at the imaginary allocations, $\forall \theta_m, \theta_{m'} \in \Theta$,

$$u(c_m^I) - h(l_m) + w(k_m^I) \geq u(c_{m'}^I) - h\left(\frac{\theta_{m'} l_{m'}}{\theta_m}\right) + w(k_{m'}^I),$$

which requires each type to truthfully reveal their productivity type. The imaginary allocations are determined by the following consumption-savings problem

$$\max_{c_m^I, k_m^I} u(c_m^I) + w(k_m^I),$$

subject to

$$c_m^I + (1 + \tau_m^k) k_m^I = Y_m,$$

The agent chooses the consumption and savings level based on his budget constraint and the first order condition: $\left(1 + \tau_m^k\right) u'(c_m^I) = w'(k_m^I)$.

To summarize, the set of policy instruments available to the government is $\left\{T_m, \tau_m^k\right\}_{m \in \{b,g\}}$, where T_m is a non-linear type specific labor income tax and τ_m^k is the indirect savings tax defined in (17). The taxes are picked based on different utilities, and they interact with each other because both taxes are announced simultaneously and the government has full commitment. This is different from the fooling mechanisms discussed in the previous sections. More specifically, the government needs to take into account how the agents' perceived allocations, (c_m^I, k_m^I), and the realized allocations, (c_m^R, k_m^R), would affect each other when choosing the optimal taxes.

The government's main concern is whether the instruments used to ameliorate the temptation problem would affect the incentives to tell the truth before production. Since all tax policies are announced at the beginning and it is in force throughout the agents' lifetime, a linear subsidy attempting to help a naive agent's temptation problem can be misconstrued as an impediment to choosing the optimal consumption and savings bundle. This disagreement between the agent and the government could distort the agent's incentives to tell the truth at the reporting stage.

First, the following lemma shows the relationship between the imaginary allocation and the actual allocation that will be chosen.

Lemma 7 *For any agent with productivity θ_m, under truth-telling, the following relationship between (c_m^I, k_m^I) and (c_m^R, k_m^R) holds: for $\hat{\beta} > \beta$, $c_m^R > c_m^I$ and $k_m^R < k_m^I$, and for $\hat{\beta} < \beta$, $c_m^R < c_m^I$ and $k_m^R > k_m^I$.*

Lemma 7 highlights the dissonance between a naive agent's present-self and his future-

self. With $\beta < \hat{\beta} = 1$, the agent is unaware that he is incorrectly inflating his retirement savings and systematically understating his consumption compared to the reality. This is because he does not anticipate his time preference to change when he decides to consume and save. I will focus on the case of inadequate saving $\beta < 1$.

Before I present the government's problem, I can rewrite the incentive compatibility constraints in terms of the agent's indirect utility from the savings and consumption decision.[9] Also, the taxes are replaced with the real allocations that are eventually chosen by the agents for any given taxes. Let for all $m, m' \in \Theta$ and $m \neq m'$,

$$M_m(c_m^R, k_m^R, l_m) \equiv \max_{c_m^I, k_m^I} u(c_m^I) - h(l_m) + w(k_m^I)$$

subject to

$$c_m^I + \frac{\beta w'(k_m^R)}{u'(c_m^R)} k_m^I = c_m^R + \frac{\beta w'(k_m^R)}{u'(c_m^R)} k_m^R.$$

and

$$M_m(c_{m'}^R, k_{m'}^R, l_{m'}) \equiv \max_{c_{m'}^I, k_{m'}^I} u(c_{m'}^I) - h\left(\frac{\theta_{m'} l_{m'}}{\theta_m}\right) + w(k_{m'}^I)$$

subject to

$$c_m^I + \frac{\beta w'(k_{m'}^R)}{u'(c_{m'}^R)} k_m^I = c_{m'}^R + \frac{\beta w'(k_{m'}^R)}{u'(c_{m'}^R)} k_{m'}^R.$$

Note that $M_m(c_m^R, k_m^R, l_m)$ is the maximal utility a type m agent would receive if he reports his type truthfully, and $M_m(c_{m'}^R, k_{m'}^R, l_{m'})$ is the maximal utility a type m agent would receive if he misreports himself as $m' \neq m$. None of the incentive compatibility constraints can be ignored since the direction of the deviation would depend on how the agents react to tax wedges. In other words, the relative size of the substitution effect and the income effect would determine which incentive compatibility constraint is binding. The direction of the deviation is ambiguous with a linear savings tax.

I adopt the primal approach by solving for the optimal allocations. The utilitarian government's objective is presented below

$$\max p_g \left[u(c_g^R) - h(l_g) + w(k_g^R)\right] + p_b \left[u(c_b^R) - h(l_b) + w(k_b^R)\right],$$

subject to

$$p_g \left(\theta_g l_g - c_g^R - k_g^R\right) + p_b \left(\theta_b l_b - c_b^R - k_b^R\right) = 0,$$

$$M_g(c_g^R, k_g^R, l_g) \geq M_g(c_b^R, k_b^R, l_b),$$

[9] If agents are not fully naive, it would not be possible to replace the incentive compatibility constraints with the value functions, since they are not evaluated using the same discount factor.

$$M_b(c_b^R, k_b^R, l_b) \geq M_b(c_g^R, k_g^R, l_g).$$

First, notice that without intertemporal taxation, the optimal labor allocations in the Mirrlees setting is also incentive compatible in the setting with naiveté. This is because the naive agents believe that they are time consistent and discount at the correct discount rate. Therefore, they evaluate their imaginary consumption and savings allocation which corresponds to the optimal Mirrlees consumption and savings allocations. However, the naive agents would not choose to consume and save at the Mirrlees allocations when the temptation for present consumption hits. Therefore, even when the government is restricted to linear taxes, it is able to achieve a welfare at least as high as the welfare attained under sophistication.

6.1.1 Qualitative and Quantitative Analysis of Linear Taxes

Let λ_m and γ be the Lagrange multipliers on the incentive compatibility constraint for type m and the government budget constraint respectively. The first order conditions are

$$[p_g + (\lambda_g - \lambda_b)]\, u'\left(c_g^R\right) + (\lambda_g - \lambda_b)\, \Delta_g^u = \gamma p_g,$$

$$[p_b - (\lambda_g - \lambda_b)]\, u'\left(c_b^R\right) - (\lambda_g - \lambda_b)\, \Delta_b^u = \gamma p_b$$

$$[p_g + (\lambda_g - \lambda_b)]\, w'\left(k_g^R\right) + (\lambda_g - \lambda_b)\, \Delta_g^w = \gamma p_g,$$

$$[p_b - (\lambda_g - \lambda_b)]\, w'\left(k_b^R\right) - (\lambda_g - \lambda_b)\, \Delta_b^w = \gamma p_b.$$

$$(p_g + \lambda_g)\, h'(l_g) - \lambda_b \frac{\theta_g}{\theta_b} h'\left(\frac{\theta_g l_g}{\theta_b}\right) = \theta_g \gamma p_g,$$

$$(p_b + \lambda_b)\, h'(l_b) - \lambda_g \frac{\theta_b}{\theta_g} h'\left(\frac{\theta_b l_b}{\theta_g}\right) = \theta_b \gamma p_b,$$

where $\Delta_m^u = \left[u'\left(c_m^I\right) - u'\left(c_m^R\right)\right] + u'\left(c_m^I\right) \frac{u''\left(c_m^R\right)}{u'(c_m^R)} \left(c_m^R - c_m^I\right)$, and $\Delta_m^w = \left[w'\left(k_m^I\right) - w'\left(k_m^R\right)\right] + w'\left(k_m^I\right) \frac{w''\left(k_m^R\right)}{w'(k_m^R)} \left(k_m^R - k_m^I\right)$, for all $m \in \{b, g\}$.

From the first order conditions, it is evident that if $\beta = 1$, then the real allocations would equal the imaginary allocations. In that case, the solution would be the optimal allocations from the Mirrlees setting. Therefore, by varying the degree of temptation β, the Mirrlees setting is a special case of the model with temptation and cognitive limitations.

The terms Δ_m^u and Δ_m^w show how the real allocations affect the agents' perceived imaginary allocations. For example, Δ_m^u measures the difference between the change in an agent's perceived utility to a small increase in c_m^R and the change in his realized utility. To see this, notice that if the government wishes to increase c_m^R, it would affect the after-tax income and

the 'price' of saving. In other words,

$$\begin{aligned}\Delta_m^u &= \frac{\partial M_m\left(c_m^R, k_m^R, l_m\right)}{\partial c_m^R} - u'(c_m^R) \\ &= \frac{\partial M_m}{\partial Y_m}\frac{\partial Y_m}{\partial c_m^R} + \frac{\partial M_m}{\partial \tau_m^k}\frac{\partial \tau_m^k}{\partial c_m^R} - u'(c_m^R),\end{aligned}$$

where after-tax income can be expressed in terms of the realized allocations $Y_m = c_m^R + (1+\tau_m^k)k_m^R$. Similarly, Δ_m^w represents a similar change with respect to an increase in k_m^R. Therefore, if Δ_m^u or Δ_m^w are strictly positive, then the government can relax the incentive compatibility constraint at the expense of consumption smoothing.

Notice that with linear policies, an agent's perception is closely tied to the targeted real allocations. Unlike the previous section where the government is able to separate the redistribution problem from paternalistic goals, with linear policies, the government is only able to achieve full redistribution by sacrificing consumption smoothing. For example, the government can relax the incentive compatibility constraint by increasing k_g^R. I will proceed to analyze an example for linear subsidies.

6.1.2 Example: CRRA Utility

I consider an example where the agents have CRRA utility for both periods. In other words, I consider the following form of utility for the agents: $u(c) = \frac{c^{1-\sigma}}{1-\sigma}$ and $w(k) = \frac{k^{1-\rho}}{1-\rho}$.

With CRRA, for a type m agent,

$$\Delta_m^u = \frac{1}{(c_m^I)^\sigma}\left[\left(1 - \left(\frac{c_m^I}{c_m^R}\right)^\sigma\right) - \sigma\left(1 - \frac{c_m^I}{c_m^R}\right)\right],$$

and

$$\Delta_m^w = \frac{1}{(k_m^I)^\rho}\left[\left(1 - \left(\frac{k_m^I}{k_m^R}\right)^\rho\right) - \rho\left(1 - \frac{k_m^I}{k_m^R}\right)\right].$$

If the coefficient of relative risk aversion σ belong in the set $(0, 1]$, then $\Delta_m^u \geq 0$, so the targeted real allocation c_m^R can be increased to relax the incentive compatibility constraint. The optimal real allocation would look similar to the Mirrlees allocation, but with possible intertemporal distortions. Of particular interest, if $\sigma = \rho = 1$, then $\Delta_m^u = \Delta_m^w = 0$ for all $m \in \{b, g\}$. Therefore, by the first order conditions, for $u(\cdot) = w(\cdot) = log(\cdot)$, the optimal allocation is the Mirrlees second best allocation.

This could be of potential interest since Chetty [(2006)] used data on labor supply behavior to estimate the coefficient of relative risk aversion. He found that the mean implied value of the coefficient of relative risk aversion to be approximately 0.71, and his finding

is relatively robust to the specifications of the model. This is in contrast to other larger estimates obtained by studying the capital markets. However, the estimation obtained in Chetty [(2006)] is intimately related to the issue this paper wishes to address, because it focuses on the effects of tax policy on labor supply. Therefore, the analysis on the optimal allocations in this section can potentially be of practical use.

[Welfare comparison to be added]

6.2 Political Constraints

In a political economy, the incentives to be re-elected would constrain the set of implementable policies. Even for benevolent political candidates, if the primary goal is to win the election, political incentives would distort the choice of policies. This is especially true when elections are held after the onset of the present bias. The competition for votes could force the candidates to pander to the voters' desire for present consumption and undermine the implementation of optimal savings policies.[10]

To model the political competition, I will assume that the election is held after the present bias and before the agents make their intertemporal savings decisions. The political candidates announce their policies on savings. Among the candidates is the incumbent, who announces the tax policies on labor provision and savings. The incumbent is not allowed to backtrack on the savings policy during the election. In the present model, if agents share the same degree of present bias β, then political competition would force the candidates to announce policies that maximizes the ex-post utility.

Suppose all agents are fully naive and have heterogeneous degrees of present bias β, which is distributed according to $G(\beta|\theta_m)$ with bounded support $\left[\underline{\beta}, \overline{\beta}\right]$ and $\overline{\beta} < 1$. Similar to the tree cutting model of Lizzeri and Yariv [(2014)], during the election, candidates would announce a fraction x of post-income tax output Y_m to be saved for retirement. The preference over x is single-peaked, so the median voter theorem holds. In equilibrium, candidates would announce the same savings policy, x_b^* and x_g^*, such that

$$x_b^* = \arg\max_x u\left((1-x)Y_b\right) + \beta_b^M w\left(xY_b\right),$$

$$x_g^* = \arg\max_x u\left((1-x)Y_g\right) + \beta_g^M w\left(xY_g\right),$$

where β_b^M and β_g^M are the median present bias values for each productivity type. It is obvious that due to political competition, all agents under-save.

[10] The timing of elections has shown to be of crucial importance. Bisin, Lizzeri and Yariv [(2014)] showed how political candidates would exploit the voters' present bias and undo the incentives for private commitment when elections are held in tandem with the intertemporal decisions of the agents.

[To be added]

7 Discussions

In this section, I will address some concerns regarding the assumptions in the paper and some of the implications of the results, in particular, the message of paternalism this paper implies.

7.1 Naiveté and Non-common Priors

In the paper, I considered an economy populated by partially and diversely naive agents. This is in sharp contrast to the existing literature on time-inconsistent preferences, which usually adopts the view that the agent is sophisticated. A partially naive agent is not fully aware of his time-inconsistency, while a sophisticated agent is. The paper departs from the usual assumption in cognitive ability due to recent developments in psychology and behavioral economics.

DellaVigna and Malmendier [(2006)] studied gym membership data. They found those who chose to be members attended the gym so seldom and irregularly that they would have been better off going as non-members. This empirical phenomenon is difficult to explain with rational or even sophisticated agents. The literature has interpreted this result as evidence in support of naiveté. The gym members hold a false belief that their willingness to exercise in the present will persist in the future, which leads them to make an incorrect contracting choice. Many other papers have demonstrated such naiveté using empirical data, including an examination of the credit card market by Ausubel [(1999)] and Shui and Ausubel [(2005)]. Models of partial naiveté also help explain the impact of the status quo in 401(k) plan choices, which is called the default effect. Madrian and Shea [(2001)] have documented the default effect on contribution rates in 401(k)s.

More recently, there is also experimental data in support of naiveté. For example, Hey and Lotito [(2009)] have found that subjects display dynamically inconsistent behavior in-line with naiveté.

A common objection to the adoption of the partial naiveté assumption is that agents have the ability to learn. After repeated decision making, an agent should and is expected to learn about his behavioral bias and thus, become fully aware of his time-inconsistency. He may even correct it accordingly. However, on the issue of retirement, most people retire only once in their lifetime. Therefore, it is safe to assume that people are unable to learn about their time inconsistency when it comes to retirement decisions, and remain largely

unaware of their behavioral bias.

There is also evidence that people do a poor job of learning about their future preferences and thus remain ignorant of their time-inconsistency problem even after repeated decision making. The psychology literature has identified several possible forces that obstruct learning. For example, it is recognized that we tend to disregard information that run counter to our beliefs, while paying much closer to attention to information that could support our beliefs. This is called confirmation bias. Another related phenomenon documented is conservatism, which describes an updating bias where individuals gives too much credence to past observations and not enough weight to new information. Another psychological phenomenon that could obstruct learning is the fact that human memory often displays limitations, so information updating is not performed on the full set signals.[11] Despite such evidence, I provide a model where learning occurs and characterize when such learning might be incomplete.

I also implicitly assumed that while the agents were partially naive, the government can anticipate the change in discount factors correctly, which creates the conflict in beliefs. Though this difference may seem arbitrary, I believe this to be a reasonable assumption. The government has access to all agents' saving behavior in the economy, while the agent has limited knowledge of this. Also, the government employs researchers, such as experts at the Bureau of Labor Statistics, studying the savings behavior of its agents. Therefore, it is safe to assume that the government is better informed about the agents' systematically changing time preferences.

7.2 Alternative Welfare Criteria

The choice of the welfare criteria in a multi-selves model is often left to the modeler's own discretion. In line with most of the work in this literature, I chose maximizing the ex-ante utility of the agents as the government's welfare objective. This view is motivated by the fact that agents wish and plan to consume allocations according to their ex-ante utility, but are subject to the whims of their ex-post utility, which they see as falling into uncontrolled temptations. This is modeled by the fact that the agents use their ex-ante utility to evaluate the incentive compatibility constraints.

However, this does not preclude the government from placing strictly positive welfare weights on the ex-post utility. The motivation for it may be that the government hopes the agent could be more spontaneous and enjoy life while he or she is young. If the government

[11] Gottlieb [(2011)] studies a model of learning with confirmation bias and conservatism and finds that learning is never complete even in the limit. Wilson [(2014)] studies a model with limited memory which generates imperfect learning.

chooses to do so, it is still able to achieve the first best allocation under the new welfare criteria. In other words, the results of the paper do not change much if the welfare criteria is different. As a result, the main idea presented is robust to subjective judgment for the appropriate welfare criterion. This is because non-sophisticated agents are dogmatic in their beliefs, and are thus easily directed towards choosing the allocations that the government wishes to implement, whatever they might be.

Though the results of the paper do not change with regards to the welfare choice, I prefer using the ex-ante utility as the main welfare criteria. As was mentioned before, the ex-post utility reflects unreasoned and instinctive preferences that the agent inherently wishes to avoid. This interpretation is consistent with the Bernheim and Rangel [(2004)] interpretation of ex-post selves. Consequently, it is natural to evaluate welfare according to the ex-ante preferences.

Another reason for using the ex-ante utility in the welfare analysis is due to a technical complication that arises in the savings application if the welfare criteria included the ex-post preferences. To aggregate both the ex-ante preference and the ex-post preference, both need to be defined over the same domain. The ex-ante preferences are defined over the labor decision and the consumption decision for both periods. However, the ex-post preference is only the intertemporal consumption decision defined over the two periods, since the labor decision has already been made. Therefore, in principle, a welfare criteria that includes the ex-post utility is meaningless when it comes to evaluating the preference over labor decisions.

7.3 Paternalism

Extensive research has been made on the optimal redistributive policies that trades off equity concerns with the potential loss in efficiency. More recently, in light of developments in behavioral economics, an argument has been made for paternalistic policies that aim to aid individuals in overcoming their undesirable tendencies. Most have argued for paternalistic policies that limit the breach of sovereignty by examining mechanisms that would alter the choices of an individual with behavioral biases, while having little effect on individuals without such biases.[12] However, this examination has been done in isolation of other goals that the government may have. In other words, a systematic analysis on how paternalism interacts with other motives, such as redistribution, has not been discussed.

In this paper, the government has both redistributive and paternalistic goals. I abstract

[12]In addition to libertarian paternalism as prescribed by Sunstein and Thaler [(2008)], Camerer et al. [(2003)] have also defended the implementation of paternalistic policies provided that they bring large benefits to boundedly rational individuals while limiting their cost on rational individuals. They call this 'asymmetric paternalism' since it leaves rational agents unaffected.

from issues of sovereignty and examine the optimal policy without ulterior constraints. I show that the the first best outcome is achievable provided that exploiting the naiveté of individuals is acceptable. Though my analysis discusses both policy and welfare implications, it is not meant to be a normative analysis. The arguments for and against paternalism are equally compelling, but this paper is not meant to take a stand on either side. In fact, it could be used to argue for paternalism and for anti-paternalism. Those in favor of paternalism could interpret the results as a further validation of manipulating individuals, not only for their own good, but for increasing the social welfare. On the other hand, anti-paternalists could argue this paper shows that even a rational and benevolent government with paternalistic goals would be motivated to go too far in exploiting agents, and that the gains in social welfare come at an exorbitant price. For example, a soft paternalistic savings plan may succumb to the government's redistributive goals and trigger a slippery slope towards more intrusive paternalism, as described in Rizzo and Whitman [(2009)]. A strong case can also be made for the moral basis of deceiving individuals who are not aware of their biases. In fact, for deception to be sustainable, the paper recommends not educating individuals about their biases.

Though this paper does not add to the discussion of whether paternalism is desirable, I believe that it does show the importance and need for rigorous analysis of paternalistic policies. An uninhibited analysis of paternalism helps us understand the form of the optimal policies, which leads naturally to a discourse of whether these policies should be adopted or rejected based on moral or philosophical considerations.

8 Summary and Conclusion

In this paper, I examined the optimal policies for a government facing a population of non-sophisticated agents with hidden productivity. I showed that the first best welfare is attainable despite the presence of asymmetric information. This is because with non-sophisticated agents, the government can separate types by exploiting their inability to precisely forecast the eminent taste change in the future. The optimal policy requires nonlinear type-specific savings subsidies. Type specific subsidies can help the government separate the types, and the non-linearity helps deceive the agents. I also explored several settings where such a strong result would not hold.

The result presented would also apply to models of industrial organization. For example, it could be applied to a model of gym membership where consumers have heterogeneous marginal value of attending gym, but are not fully aware of their time inconsistency. The gym can fully price discriminate with a membership contract that is type specific and includes

heavily discounted usage prices in the future with an expensive alternative option. Consumers would be attracted by the discounts they would enjoy in the future, mis-predicting the fact that their tastes would change and would prefer the alternative option.

A serious issue that is not being addressed by the present model is the lack of learning by the agents. A dynamic model with non-dogmatic agents can potentially shed light on this issue. If agents are expected to learn about their present bias problem, the government must adjust their optimal policies each period to continue to deceive the agents. Once the agents learn about their bias, the government is no longer able to exploit them. The optimal path of policies would have to trade-off the immediate benefit of achieving the government's redistributive and paternalistic goals, with the long-run cost of taxing sophisticated agents. Another serious issue that was not rigorously addressed in this paper was the desirability of paternalistic regulations.

Appendix A: Proofs

Proof of Proposition 1, 2

Let μ_m^I (μ_m^R) be the Lagrange multiplier on the fooling constraint for productivity type θ_m to preferring the imaginary (real) allocation over the real (imaginary) allocation. Finally, let $\lambda(\theta_{m'};\theta_m)$ be the Lagrange multiplier for the incentive compatibility constraint on type θ_m misreporting to be $\theta_{m'}$.

Let us begin with magnitude naiveté, and analyze the first order conditions for the imaginary consumption, $\forall\theta_m \in \Theta$ and $\forall n \in N$,

$$\left\{\sum_{\theta_{m'}\in\Theta}\left[\lambda(\theta_{m'};\theta_m)-\lambda(\theta_m;\theta_{m'})\right]+\alpha\mu_m^I\right\}\frac{\partial U}{\partial c_{m,n}^I}=\left[\mu_m^R-(1-\alpha)\mu_m^I\right]\frac{\partial V}{\partial c_{m,n}^I}.$$

By Assumption 1 and the fact that $\lim_{c_n\to 0}\frac{\partial U}{\partial c_n}=+\infty$ and $\lim_{c_n\to 0}\frac{\partial V}{\partial c_n}=+\infty$, so consumption is strictly positive (non-negativity constraints never bind), the following is immediate

$$\sum_{\theta_{m'}\in\Theta}\left[\lambda(\theta_{m'};\theta_m)-\lambda(\theta_m;\theta_{m'})\right]+\alpha\mu_m^I=\mu_m^R-(1-\alpha)\mu_m^I=0.$$

Summing across all types yields us

$$\sum_{\theta_m\in\Theta}\sum_{\theta_{m'}\in\Theta}\left[\lambda(\theta_{m'};\theta_m)-\lambda(\theta_m;\theta_{m'})\right]=0 \tag{19}$$

This implies that $\alpha\sum_{m'}^{M}\mu_m^I=0$. As a result, the Kuhn-Tucker necessary conditions implies that $\mu_m^I=0$, which also gives us $\mu_m^R=0$.

For frequency naiveté, the first order conditions for the imaginary consumption, $\forall\theta_m\in\Theta$ and $\forall n\in N$,

$$\left\{\alpha\sum_{\theta_{m'}\in\Theta}\left[\lambda(\theta_{m'};\theta_m)-\lambda(\theta_m;\theta_{m'})\right]+\mu_m^I\right\}\frac{\partial U}{\partial c_{m,n}^I}=\mu_m^R\frac{\partial V}{\partial c_{m,n}^I}.$$

By Assumption 1, the following must hold $\alpha\sum_{\theta_{m'}\in\Theta}\left[\lambda(\theta_{m'};\theta_m)-\lambda(\theta_m;\theta_{m'})\right]+\mu_m^I=\mu_m^R=0$. Using a similar method as in (19), it follows that $\mu_m^I=\mu_m^R=0$ for all productivity types.

Since $\mu_m^I=\mu_m^R=0$, the first order conditions on the imaginary consumption for both types of naiveté and for all $\theta_m\in\Theta$ have the following property

$$\sum_{\theta_{m'}\in\Theta}\lambda(\theta_{m'};\theta_m)=\sum_{\theta_{m'}\in\Theta}\lambda(\theta_m;\theta_{m'}) \tag{20}$$

Consider the most productive agent θ_M and assume that there exists a type $\theta_{\hat{m}}$ such that $\lambda(\theta_{\hat{m}}, \theta_M) > 0$. By (20), there exists a type $\theta_{\tilde{m}}$ such that $\lambda(\theta_M, \theta_{\tilde{m}}) > 0$. In other words, if the most productive type is indifferent between truth-telling and pretending to be a less efficient type $\theta_{\hat{m}}$, then there is another type of agent $\theta_{\tilde{m}}$ that would be indifferent between truth-telling and pretending to be the most efficient type.

By Assumption 2, there exists another allocation for type θ_M with larger $c^I(\theta_M)$ and more labor $l(\theta_M)$ such that type θ_M strictly prefers it to the original one and type $\theta_{\tilde{m}}$ would never choose this new allocation. To see this, let $(y(\theta_M), c^I(\theta_M))$ and $(y(\theta_{\tilde{m}}), c^I(\theta_{\tilde{m}}))$ denote the original allocations, and $(y^*(\theta_M), c^{I*}(\theta_M))$ and $(y^*(\theta_{\tilde{m}}), c^{I*}(\theta_{\tilde{m}}))$ the new allocations. Choose good $n \in N$ such that its production depends on labor. Let $MRS(y,c)_{M,n}$ and $MRS(y,c)_{\tilde{m},n}$ be the marginal rate of substitution of the two types, θ_M and $\theta_{\tilde{m}}$, in y_n and c_n. The easiest way to construct the new allocations is to choose it such that $(y_n^*(\theta_M), c_n^{I*}(\theta_M)) = (y_n(\theta_M) + \epsilon, c_n^I(\theta_M) + \alpha\epsilon)$ and $(y_n^*(\theta_{\tilde{m}}), c_n^{I*}(\theta_{\tilde{m}})) = (y_n(\theta_{\tilde{m}}), c_n^I(\theta_{\tilde{m}}))$, where $MRS(y(\theta_M), c^I(\theta_M))_{M,n} < \alpha < MRS(y(\theta_{\tilde{m}}), c^I(\theta_{\tilde{m}}))_{\tilde{m},n}$ and ϵ is chosen to be sufficiently large so that type $\theta_{m'}$ is strictly worse off when he pretends to be type θ_M. Since the imaginary allocation does not enter the government's welfare criterion, the extra output of ϵ can then be redistributed, which raises the social welfare. Therefore, $\lambda(\theta_M, \theta_{m'}) = 0$ for all $\theta_{m'}$ which contradicts (20), so $\lambda(\theta_{\hat{m}}, \theta_M) = 0$ for all $\theta_{\hat{m}}$.

The same argument can be repeated for all lower productivity types. In essence, it is never optimal for $\lambda(\theta_m, \theta_{m'}) > 0$ when $\theta_{m'} < \theta_m$, so by (20), it is also not optimal for $\lambda(\theta_m, \theta_{m'}) > 0$ when $\theta_{m'} > \theta_m$. Therefore, all Lagrange multipliers for all incentive compatibility constraints are non-positive. This proves Proposition 2, and since the argument does not depend on α, it also proves Proposition 1. ■

Proof of Corollary 3, 4

By Proposition 1 and Proposition 2, the government can implement the first best allocation. Let $\left\{\left(c^R(\theta_m), l(\theta_m)\right)\right\}_{\theta_m \in \Theta}$ be the first best allocation. Suppose the government does not fool the agents, then by definition, for all $\theta_m \in \Theta$, $\left(c^R(\theta_m), l(\theta_m)\right) = \left(c^I(\theta_m), l(\theta_m)\right)$, which violates the incentive compatibility constraint for some types. It follows that the government must implement a fooling mechanism to achieve the first best allocation. ■

Proof of Proposition 3:

Let $\lambda_{i,q}^{j,r}$ denote the Lagrange multipliers on the incentive compatibility constraints, where a productivity level i and sophistication level j agent is discouraged from pretending to be a productivity level q and sophistication level r agent. Let γ be the Lagrange multiplier for the government budget constraint.

By Lemma 6, the incentive compatibility constraints for the naive agents can be ignored. The first order conditions for the sophisticated high productivity type agents are as follows

$$\left(\pi_g^s + \lambda_{g,b}^{s,s} + \lambda_{g,b}^{s,n} + \lambda_{g,g}^{s,n} - \lambda_{b,g}^{s,s}\right) u'(c_g^s) = \gamma \pi_g^s, \tag{21}$$

$$\left(\pi_g^s + \lambda_{g,b}^{s,s} + \lambda_{g,b}^{s,n} + \lambda_{g,g}^{s,n} - \lambda_{b,g}^{s,s}\right) w'(k_g^s) = \gamma \pi_g^s, \tag{22}$$

$$\left(\pi_g^s + \lambda_{g,b}^{s,s} + \lambda_{g,b}^{s,n} + \lambda_{g,g}^{s,n}\right) \frac{1}{\theta_g} h'(l_g^s) - \lambda_{b,g}^{s,s} \frac{1}{\theta_b} h'\left(\frac{\theta_g l_g^s}{\theta_b}\right) = \gamma \pi_g^s. \tag{23}$$

The first order conditions for the sophisticated low productivity type agents are

$$\left(\pi_b^s + \lambda_{b,g}^{s,s} + \lambda_{b,g}^{s,n} + \lambda_{b,b}^{s,n} - \lambda_{g,b}^{s,s}\right) u'(c_b^s) = \gamma \pi_b^s, \tag{24}$$

$$\left(\pi_b^s + \lambda_{b,g}^{s,s} + \lambda_{b,g}^{s,n} + \lambda_{b,b}^{s,n} - \lambda_{g,b}^{s,s}\right) w'(k_b^s) = \gamma \pi_b^s, \tag{25}$$

$$\left(\pi_b^s + \lambda_{b,g}^{s,s} + \lambda_{b,g}^{s,n} + \lambda_{b,b}^{s,n}\right) \frac{1}{\theta_b} h'(l_b^s) - \lambda_{g,b}^{s,s} \frac{1}{\theta_g} h'\left(\frac{\theta_b l_b^s}{\theta_g}\right) = \gamma \pi_b^s. \tag{26}$$

The first order conditions for the naive high productivity type agents are

$$\left(\pi_g^n - \lambda_{g,g}^{s,n} - \lambda_{b,g}^{s,n}\right) u'(c_g^R) = \gamma \pi_g^n, \tag{27}$$

$$\left(\pi_g^n - \lambda_{g,g}^{s,n} - \lambda_{b,g}^{s,n}\right) w'(k_g^R) = \gamma \pi_g^n, \tag{28}$$

$$\left(\pi_g^n - \lambda_{g,g}^{s,n}\right) \frac{1}{\theta_g} h'(l_g^n) - \lambda_{b,g}^{s,n} \frac{1}{\theta_b} h'\left(\frac{\theta_g l_g^n}{\theta_b}\right) = \gamma \pi_g^n. \tag{29}$$

Finally, the first order conditions for the naive low productivity type agents are

$$\left(\pi_b^n - \lambda_{b,b}^{s,n} - \lambda_{g,b}^{s,n}\right) u'(c_b^R) = \gamma \pi_b^n, \tag{30}$$

$$\left(\pi_b^n - \lambda_{b,b}^{s,n} - \lambda_{g,b}^{s,n}\right) w'(k_b^R) = \gamma \pi_b^n, \tag{31}$$

$$\left(\pi_b^n - \lambda_{b,b}^{s,n}\right) \frac{1}{\theta_b} h'(l_b^n) - \lambda_{g,b}^{s,n} \frac{1}{\theta_g} h'\left(\frac{\theta_b l_b^n}{\theta_g}\right) = \gamma \pi_b^n. \tag{32}$$

Part (i) of the proposition immediate follows from the first order conditions.

The analysis will now proceed by checking the slackness of the remaining incentive compatibility constraints. The proof will proceed via a series of lemmas.

Lemma 8 *If* $\lambda_{b,g}^{s,s} = 0$, *then* $\lambda_{g,g}^{s,n} = \lambda_{b,g}^{s,n} = 0$.

Proof If $\lambda_{b,g}^{s,s} = 0$, then from the first order conditions (23) and (29), $l_g^n \geq l_g^s$, and for the

consumption level, from (21), (22), (27) and (28), $c_g^R \leq c_g^s$ and $k_g^R \leq k_g^s$. This would mean that the utility for the sophisticated high productivity type is higher than the utility for the naive high productivity type, so $\lambda_{g,g}^{s,n} = \lambda_{b,g}^{s,n} = 0$. ■

Suppose that $\lambda_{b,g}^{s,s} = 0$, its validity will be checked later. By Lemma 8, this implies that $\lambda_{g,g}^{s,n} = \lambda_{b,g}^{s,n} = 0$.

Lemma 9 *At least two of the three Lagrange multipliers on the incentive compatibility constraints are strictly positive.*

Proof It is obvious that at least one Lagrange multiplier on the incentive compatibility constraint must be strictly positive.

First, assume that only $\lambda_{b,b}^{s,n} > 0$, then by (21), (22), (24) and (25), $c_b^s > c_g^s$ and $k_b^s > k_g^s$, but from (23) and (26), it is implied that $l_g^s > l_b^s$. This violates the incentive compatibility constraint since the sophisticated high productivity type would rather pretend to be the sophisticated low productivity type.

Next, assume that only $\lambda_{g,b}^{s,s} > 0$, then from the first order conditions (24), (25), (30) and (8), $c_b^R > c_b^s$ and $k_b^R > k_b^s$, but $l_b^n < l_b^s$ from (26) and (8). The sophisticated low productivity type would rather pretend to be the naive low productivity type, so it is not incentive compatible.

Finally, assume that only $\lambda_{g,b}^{s,n} > 0$. The following relationships must hold

$$\begin{aligned} u(c_g^s) - h(l_g^s) + w(k_g^s) &= u(c_b^R) - h\left(\frac{\theta_b l_b^n}{\theta_g}\right) + w(k_b^R) \\ &\geq u(c_b^s) - h\left(\frac{\theta_b l_b^s}{\theta_g}\right) + w(k_b^s), \end{aligned}$$

and

$$u(c_b^R) - h(l_b^n) + w(k_b^R) \leq u(c_b^s) - h(l_b^s) + w(k_b^s).$$

This gives us

$$h\left(\frac{\theta_b l_b^s}{\theta_g}\right) - h(l_b^s) \geq h\left(\frac{\theta_b l_b^n}{\theta_g}\right) - h(l_b^n),$$

and by the strict convexity of $h(\cdot)$ and the fact that $\theta_b < \theta_g$, it follows that $l_b^n \geq l_b^s$. Furthermore, since the sophisticated high productivity type would rather mimic the naive low productivity type than the sophisticated low productivity type, so we must have $c_b^R \geq c_b^s$ and $k_b^R \geq k_b^s$. However, if only $\lambda_{g,b}^{s,n} > 0$, then by the first order conditions (24), (25), (30) and (8), $c_b^R < c_b^s$ and $k_b^R < k_b^s$, which is a contradiction. ■

Lemma 10 *If* $\lambda_{b,b}^{s,n} > 0$, *then* $\lambda_{g,b}^{s,s} > 0$ *and* $\lambda_{g,b}^{s,n} > 0$.

Proof By Lemma 9, if $\lambda_{b,b}^{s,n} > 0$, then either $\lambda_{g,b}^{s,s} > 0$ or $\lambda_{g,b}^{s,n} > 0$. I will proceed by contradiction and assume that only one of the high productivity type incentive compatibility constraint binds.

Suppose $\lambda_{g,b}^{s,n} > 0$ and $\lambda_{g,b}^{s,s} = 0$, then

$$\left[u(c_b^R) - h\left(\frac{\theta_b l_b^n}{\theta_g}\right) + w(k_b^R)\right] - \left[u(c_b^s) - h\left(\frac{\theta_b l_b^s}{\theta_g}\right) + w(k_b^s)\right] \geq 0.$$

Since it was assumed that $\lambda_{b,b}^{s,n} > 0$, then

$$[u(c_b^s) + w(k_b^s)] - \left[u(c_b^R) + w(k_b^R)\right] = h(l_b^s) - h(l_b^n).$$

This gives us

$$h\left(\frac{\theta_b l_b^s}{\theta_g}\right) - h(l_b^s) \geq h\left(\frac{\theta_b l_b^n}{\theta_g}\right) - h(l_b^n),$$

and by the strict convexity of $h(\cdot)$ and the fact that $\theta_b < \theta_g$, it follows that $l_b^n \geq l_b^s$. This implies that $c_b^R \geq c_b^s$ and $k_b^R \geq k_b^s$. However, if $\lambda_{g,b}^{s,n} > 0$ and $\lambda_{g,b}^{s,s} = 0$, then the first order conditions (24), (25), (30) and (8) yield $u'(c_b^R) > u'(c_b^s)$. This is a contradiction.

Now suppose $\lambda_{g,b}^{s,n} = 0$ and $\lambda_{g,b}^{s,s} > 0$, then a similar analysis would yield $l_b^n \leq l_b^s, c_b^R \leq c_b^s$ and $k_b^R \leq k_b^s$. From the first order conditions (30) and (8), it is clear that $u'(c_b^R) = \frac{1}{\theta_b} h'(l_b^n)$. It follows that $u'(c_b^s) \leq \frac{1}{\theta_b} h'(l_b^s)$. From the first order conditions (24) and (26),

$$\left(\pi_b^s + \lambda_{b,b}^{s,n}\right)\left[u'(c_b^s) - \frac{1}{\theta_b} h'(l_b^s)\right] = \lambda_{g,b}^{s,s}\left[u'(c_b^s) - \frac{1}{\theta_g} h'\left(\frac{\theta_b l_b^s}{\theta_g}\right)\right],$$

which gives the following strict relationship $u'(c_b^s) < \frac{1}{\theta_b} h'(l_b^s)$. It follows that $\lambda_{g,b}^{s,s} > \pi_b^s + \lambda_{b,b}^{s,n}$. By the first order condition (24), $\gamma \pi_b^s < 0$ immediately follows, which is a contradiction. ■

By Lemma 9 and Lemma 10, there are two possible cases to consider: all three remaining incentive compatibility constraints are binding, and only the incentive compatibility constraints for the productive type are binding. For the first case, when all three are binding, it is immediate from the incentive compatibility constraints that $c_b^s = c_b^R$, $k_b^s = k_b^R$ and $l_b^s = l_b^n$. Therefore, this yields a partial pooling equilibrium. For the second case, when only the incentive compatibility constraints for the productive type are binding, it is immediate from the incentive constraints that $c_b^s < c_b^R$, $k_b^s < k_b^R$ and $l_b^s < l_b^n$.

The rest of the results for Proposition 3 follows immediately from the first order conditions. It is also easy to check that the sophisticated low productivity type agent would strictly prefer truth-telling over pretending to be a high productivity agent for both cases. This completes the proof. ■

Proof of Lemma 7:

For any type m, the imaginary allocation is related to the real allocation by the marginal rate of intertemporal substitution $\frac{u'(c_m^R)}{\beta w'(k_m^R)} = \hat{\beta}\frac{u'(c_m^I)}{w'(k_m^I)}$. By the budget constraint, I can express the relationship in terms of consumption

$$\frac{u'\left[y_m - \left(1+\tau^k\right)k_m^R\right]}{\beta w'\left(k_m^R\right)} = \frac{u'\left[y_m - \left(1+\tau^k\right)k_m^I\right]}{w'\left(k_m^I\right)}. \tag{33}$$

Let $h(x) = \frac{u'\left[y_m - \left(1+\tau^k\right)x\right]}{w'(x)}$, and since the period utility functions are strictly increasing and strictly concave, then we have $h'(x) > 0$. With $\beta < \hat{\beta}$, for (33) to hold, we have $c_m^R > c_m^I$ and $k_m^R < k_m^I$. ■

Appendix B: Quasi-hyperbolic Discounting Model

I will consider a quasi-hyperbolic discounting model with three periods. In contrast to the model presented in the paper, the agents will work for two periods and retire at the third and final period. The within period timing will remain the same as in the paper: agents will work before the present bias occurs and then make consumption savings decision. I will focus on the magnitude naiveté case. Following Laibson [(1997)] and O'Donoghue and Rabin [(2001)], the utility of the agents is represented as follows

$$\begin{aligned}
&U_1\left(c_1, c_2, k, l_1, l_2\right) = u(c_1) - h(l_1) + \hat{\beta}_1\delta\left[u(c_2) - h(l_2) + \delta w(k)\right], \\
&U_2\left(c_2, k, l_2\right) = u(c_2) - h(l_2) + \hat{\beta}_2\delta w(k), \\
&U_3\left(c_3\right) = w(c_3).
\end{aligned}$$

Notice that I allow the partially naive agents to learn from their mistakes and update their beliefs on β, so unless the agents are dogmatic, I allow for $\hat{\beta}_1 \neq \hat{\beta}_2$. Learning does not occur if the agents start off sophisticated.

The learning process is not explicitly modeled. The only restriction is if $\hat{\beta}_1 > \beta$, then $\beta \leq \hat{\beta}_2 < \hat{\beta}_1$. The agents can be partially naive for both periods, or be sophisticated at the second period. Therefore, there are two cases to consider: the partial learning ($\hat{\beta}_2 \neq \beta$) case and the full learning ($\hat{\beta}_2 = \beta$) case. It is plausible to imagine that agents learn about their biases outside of the model as well. As a result, I will further assume that the learning process is independent of the government policy. For the full learning case, this assumption rules out the possibility of the government choosing to deceive in the second period instead

of the first. I will focus on the case with non-sophisticated agents, $1 \geq \hat{\beta}_1 > \beta$. Without loss of generality, let $\delta = 1$.

For simplicity, let $\Theta = \{\theta_b, \theta_g\}$ and let the initial distribution be $1 > \Pr(\theta_{m,1}) = \pi_m > 0$, with transition probability $1 > \Pr(\theta_{m',2}|\theta_{m,1}) = \pi_{m,m'} > 0$. The agents only differ in their productivity. They share the same underlying present bias β and initial belief $\hat{\beta}_1$. They also share the same learning process, so $\hat{\beta}_2$ is the same for all agents. The analysis will begin by discussing the partial learning case.

For the partial learning case, the real allocations are

$$\left\{ \left(c_1^R(\theta_m^1), l_1(\theta_m^1)\right); \left(c_2^R(\theta_{m'}^2; \theta_m^1), l_2(\theta_{m'}^2; \theta_m^1);\right) k^R(\theta_m^1, \theta_{m'}^2), \right\}_{m,m' \in \{b,g\}},$$

and the imaginary allocations are

$$\left\{c_1^I(\theta_m^1), c_2^I(\theta_{m'}^2; \theta_m^1), k^I(\theta_m^1, \theta_{m'}^2)\right\}_{m,m' \in \{b,g\}}.$$

In the second period, at the reporting stage, any type $(\theta_m^1, \theta_{m'}^2)$ agent faces similar incentive compatibility constraints and fooling constraints as (8), (9),(10), (11),(12) and (13). Since for partial learning $\hat{\beta}_2 > \beta$, the imaginary allocations can be designed such that the incentive compatibility constraints are non-binding, the government can fool the agents and achieve the desired redistribution without any distortions.

For the first period, in addition to the first period imaginary allocations, the government deceives the partially naive agents with an imaginary continuation value $B^I(\theta_m^1)$ for any reported first period type θ_m^1. After the agents supply their labor, the present bias appears and would instead choose $c_1^R(\theta_m^1)$ and a continuation value of

$$B(\theta_m^1) = \sum_{\theta_{m'}^2 \in \Theta} \pi_{m,m'} \left[u\left(c_2^I(\theta_{m'}^2; \theta_m^1)\right) - h\left(l_2(\theta_{m'}^2; \theta_m^1)\right) + w\left(k^I(\theta_m^1, \theta_{m'}^2)\right)\right],$$

which is the 'chosen' continuation value for the θ_m^1 agent in the first period.

The type m agent faces the following incentive compatibility constraint in the first period, for any $\theta_{\hat{m}}^1 \in \Theta$,

$$u\left(c_1^I(\theta_m^1)\right) - h\left(l_1(\theta_m^1)\right) + B^I(\theta_m^1) \geq u\left(c_1^I(\theta_{\hat{m}}^1)\right) - h\left(\frac{\theta_{\hat{m}}^1 l_1(\theta_{\hat{m}}^1)}{\theta_m}\right) + B^I(\theta_{\hat{m}}^1).$$

and the following fooling constraints

$$u\left(c_1^I(\theta_m^1)\right) + \hat{\beta}_1 B^I(\theta_m^1) \geq u\left(c_1^R(\theta_m^1)\right) + \hat{\beta}_1 B(\theta_m^1),$$

$$u\left(c_1^R(\theta_m^1)\right) + \beta B(\theta_m^1) \geq u\left(c_1^I(\theta_m^1)\right) + \beta B^I(\theta_m^1).$$

Therefore, with the appropriate imaginary first period consumption and continuation value, the government is able to deceive the agents in the first period and achieve any redistribution.

For the full learning case, since the agents are sophisticated in the second period, the government is unable to deceive them with second period imaginary allocations. Hence, the government is only able to fool the agents in the first period. In the second period, the incentive compatibility constraint for the efficient agent is binding for any reported first period type, so there is distortion in the second period allocations and the government is unable to achieve the first best welfare. However, for the first period, the government can deceive the agents in a similar way as the partial learning case and attain perfect insurance across productivity types.

References

Amador, Manuel, Ivan Werning and George-Marios Angeletos, "Commitment vs. Flexibility," *Econometrica*, 2006, 74 (2), 365-396.

Ausubel, Lawrence, "Adverse Selection in the Credit Card Market," *Unpublished*.

Bassi, Matteo, "Mirrlees Meets Laibson: Optimal Income Taxation with Bounded Rationality," *CSEF Working Paper*, No. 266.

Beaudry, Paul, Charles Blackorby and Dezso Szalay, "Taxes and Employment Subsidies in Optimal Redistribution Programs," *American Economic Review*, 2009, 99(1), 216-242.

Bernheim, Douglas and Antonio Rangel, "Addiction and Cue-Triggered Decision Processes," *American Economic Review*, 2004, 94(5), 1558-1590.

Bisin, Alberto, Alessandro Lizzeri and Leeat Yariv, "Government Policy with Time Inconsistent Voters," *American Economic Review*, forthcoming.

Camerer, Colin, Samuel Issacharoff, George Loewenstein, Ted O'Donoghue and Matthew Rabin, "Regulation for Conservatives: Behavioral Economics and the Case for Asymmetric Paternalism," *University of Pennsylvania Law Review*, 2003, 151, 1211-1254.

Cremer, Helmuth, Pierre Pestieau and Jean-Charles Rochet, "Direct Versus Indirect Taxation: The Design of the Tax Structure Revisited," *International Economic Review*, 2001, 42, 781-799.

Cremer, Helmuth, Pierre Pestieau and Jean-Charles Rochet, "Capital Income Taxation When Inherited Wealth is Not Observable," *Journal of Public Economics*, 2003, 87, 2475-2490.

Chetty, Raj, "A New Method of Estimating Risk Aversion," *American Economic Review*, 2006, 96(5), 1821-1834.

Dellavigna, Stefano, "Psychology and Economics: Evidence from the Field," *Journal of Economic Literature*, 2009, 47(2), 315-372.

DellaVigna, Stefano and Ulrike Malmendier, "Paying Not to Go to the Gym," *American Economic Review*, 2006, 96(3), 694-719.

Diamond, Peter and Johannes Spinnewijn, "Capital Income Taxes with Heterogeneous Discount Rates," *American Economic Journal: Economic Policy*, 2011, 3, 52-76.

Eliaz, Kfir and Ran Spiegler, "Contracting with Diversely Naive Agents," *Review of Economic Studies*, 2006, 73, 689-714.

Frederick, Shane, George Loewenstein and Ted O'Donoghue, "Time Discounting and Time Preference: A Critical Review," *Journal of Economic Literature*, 2002, 40, 351-401.

Golosov, Mikhail, Narayana Kocherlakota and Aleh Tsyvinski, "Optimal Indirect and Capital Taxation," *Review of Economic Studies*, 2003, 70, 569-587.

Gottlieb, Daniel, "Will You Never Learn? Self Deception and Biases in Information Processing," *Working Paper*

Gruber, Jonathan and Botond Koszegi, "Tax Incidence when Individuals are Time Inconsistent: the Case of Cigarette Excise Taxes," *Journal of Public Economics*, 2004, 88, 1959-1987.

Guo, Jang-Ting and Alan Krause, "Dynamic Nonlinear Income Taxation with Quasi-Hyperbolic Discounting and No Commitment," *Journal of Economic Behavior and Organization*, 2015, 109, 101-119.

Heidhues, Paul and Botond Koszegi, "Exploiting naiveté about Self-Control in the Credit Market," *American Economic Review*, 2010, 100(5), 2279-2303.

Hey, John and Gianna Lotito, "Naive, Resolute or Sophisticated? A Study of Dynamic Decision Making," *Journal of Risk and Uncertainty*, 2009, 38, 1-25.

Krusell, Per, Burhanettin Kuruscu and Anthony Smith Jr., "Temptation and Taxation," *Econometrica*, 2010, 78(6), 2063-2084.

Laibson, David, "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 1997, 112(2), 443-477.

Laibson, David, Andrea Repetto and Jeremy Tobacman, "Self-Control and Saving for Retirement," *Brookings Papers on Economic Activity*, 1998, 1, 91-196.

Loewenstein, George, Ted O'Donoghue and Matthew Rabin, "Projection Bias in Predicting Future Utility," *Quarterly Journal of Economics*, 118, 1209-1248.

Lizzeri, Alessandro and Leeat Yariv, "Collective Self-Control," *Working Paper*, 2014.

Madrian, Brigitte and Dennis Shea, "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior," *Quarterly Journal of Economics*, 2001, 116(4), 1149-1187.

Mirrlees, James, "An Exploration in the Theory of Optimal Income Taxation," *Review of Economic Studies*, 1971, 38, 175-208.

O'Donoghue, Ted and Matthew Rabin, "Choice and Procrastination," *Quarterly Journal of Economics*, 2001, 116, 121-160.

O'Donoghue, Ted and Matthew Rabin, "Studying Optimal Paternalism, Illustrated by a Model of Sin Taxes," *American Economic Review Papers and Proceedings*, 2003, 93(2), 186-191.

O'Donoghue, Ted and Matthew Rabin, "Optimal Sin Taxes," *Journal of Public Economics*, 2006, 90, 1825-1849.

Rizzo, Mario and Douglas Glen Whitman, "Little Brother is Watching You: New Paternalism on the Slippery Slopes," *Arizona Law Review*, 2009, 51(3), 685-739.

Shui, Haiyan and Lawrence Ausubel, "Time Inconsistency in the Credit Market," *Working Paper*.

Spiegler, Rani, "Bounded Rationality and Industrial Organization," *Oxford University Press*, 2011.

Sunstein, Cass and Richard Thaler, "Libertarian Paternalism," *American Economic Review Papers and Proceedings*, 2003, 93(2), 175-179.

Thaler, Richard and Cass Sunstein, "Nudge: Improving Decisions About Health, Wealth, and Happiness," *Penguin Group, New York*, 2008.

Thaler, Richard and Shlomo Benartzi, "Save More Tomorrow: Using Behavioral Economics to Increase Employee Savings," *Journal of Political Economy*, 2004, 112, S164-S187.

Wilson, Andrea, "Bounded Memory and Biases in Information Processing," *Econometrica*, 2014, 82(6), 2257-2294.