Optimal Paternalistic Savings Policies*

Christian Moser† Pedro Olea de Souza e Silva‡

August 14, 2026

Abstract

We study optimal savings policies when there is a dual concern about under-saving for retirement and income inequality. Agents differ in time preferences and earnings ability, both unobservable to a planner with paternalistic and redistributive motives. We characterize the solution to this two-dimensional screening problem and provide a decentralization using realistic policy instruments: forced savings at low incomes—similar to Social Security—but a choice between savings accounts with different subsidies and caps at high incomes—like 401(k) and IRA accounts in the US. Offering more choice in savings at higher incomes facilitates redistribution. Relative to the current US retirement system, we find large welfare gains from increasing mandatory savings and limiting savings choice at low incomes.

Keywords: Optimal Taxation, Multidimensional Screening, Preference Heterogeneity, Redistribution, Social Security, Retirement Plans

JEL classification: H21, E62, H55

*We are grateful to Mike Golosov for invaluable advice and encouragement throughout this project. We thank Mark Aguiar, Youssef Benzarti, Nobu Kiyotaki, Dirk Krueger, Felix Kubler, Rasmus Lentz, Richard Rogerson, Chris Sleet, Karl Schmedders, Aleh Tsyvinski, and Sevin Yeltekin for helpful comments and suggestions. We also thank seminar participants at Princeton University, the Maxwell School at Syracuse University, the Federal Reserve Bank of Atlanta, FGV-São Paulo (EESP), FGV-Rio (EPGE), PUC-Rio, Insper, the University of São Paulo, Columbia University, the University of Connecticut, as well as attendants at the Conference of German Economists Abroad, the Whitebox Advisors Graduate Student Conference at Yale, the Stanford SITE Workshop on Psychology & Economics, the Stanford SITE Workshop on Computational Methods for Dynamic Economies and Games, the Asia Meeting of the Econometric Society, the EEA Meetings, the National Tax Association's Annual Conference, the NBER SI Public Economics workshop, and the Federal Reserve Bank of Richmond for their feedback. Moser also thanks the Federal Reserve Bank of Minneapolis and the Heller-Hurwicz Economics Institute at the University of Minnesota for their generous hospitality during a significant share of the period of work on this project. The views and opinions expressed are those of the authors and do not necessarily reflect the views or opinions of any organization with whom the authors have been, are, or may be affiliated with in the future. Any errors are our own.

†Columbia University, NBER, CEPR, CESifo, and IZA@LISER. Email: c.moser@columbia.edu.

‡OpenAI.

1 Introduction

A shared feature of many modern welfare states is limited choice in savings for retirement.[1] These systems commonly force contributions toward old-age benefits on the basis of a paternalistic motive to induce adequate savings for retirement, particularly among low-income groups (Diamond, 1977; Kotlikoff et al., 1982; Feldstein, 1985). Rationales for the paternalistic motive derive both from the behavioral sciences and also from a purely neoclassical economics tradition. On one hand, individuals may make mistakes when choosing under incomplete information and uncertainty (Tversky and Kahneman, 1974), or they may suffer from time-inconsistent decision making over the life-cycle (Laibson, 1997). On the other hand, altruism and lack of government commitment give rise to an externality problem labeled the Samaritan's Dilemma, leading individuals to rationally under-save in anticipation of free-riding on public funds during retirement (Buchanan, 1975; Prescott, 2004; Sleet and Yeltekin, 2006). In both environments, individuals' preferences are characterized by *present bias* and paternalistic savings policies may be welfare-improving.

We study optimal retirement savings policies when there is a paternalistic motive to overcome individuals' present bias problems. The central question we ask is: how much choice in savings should be optimally offered throughout the income distribution? To address this question, we integrate a paternalistic savings motive into an optimal taxation framework, allowing us to study the problem of savings adequacy jointly with the issue of income inequality. Our key insight is that there exists a trade-off between paternalism and redistribution. As a result, the optimal policy enforces high savings rates at low incomes but offers a choice between various subsidized savings options at high incomes. Qualitatively, the optimal policies in our framework resemble many real-world retirement savings systems, including Social Security and various subsidized savings accounts in the US. Quantitatively, however, we find large welfare gains relative to current US policies from increasing mandatory savings and limiting savings choice at low incomes.

In our theoretical framework, the interaction between two ingredients gives rise to a novel trade-off in optimal savings policy design. The first ingredient, motivated by recent experimental evidence (Montiel Olea and Strzalecki, 2014), is heterogeneity in individuals' present bias. The

[1] Government-mandated old-age benefits were administered in ancient Rome to prevent revolts by impoverished army veterans (Choi, 2015). In the modern world, German chancellor Otto von Bismarck instituted the Old Age and Disability Insurance Law of 1889 to guarantee adequate incomes for retired workers (Kotlikoff, 1996). The US Old-Age, Survivors, and Disability Insurance program, or Social Security in short, signed into law in 1935 under President Roosevelt to ameliorate the extent of poverty among retirees, is nowadays the nation's largest federal government social policy, with 884 billion US dollars in transfers to 60 million beneficiaries in 2016 (Social Security Administration, 2017).

second ingredient is heterogeneity in earnings ability as in Mirrlees (1971), Diamond (1998), and Saez (2001). A paternalistic and redistributive planner defines the efficient savings rate according to a single time preference and attaches different welfare weights across ability types.[2] The planner picks a consumption and labor allocation to maximize welfare subject to incentive compatibility and a resource constraint. The theoretical analysis of this problem is complex since, as is well known, multi-dimensional screening problems lead to failure of the first-order approach that the optimal taxation literature usually relies on (Golosov et al., 2003, 2016). We exploit the paternalistic formulation to provide a partial characterization of this problem under weak regularity conditions. Our main theoretical result highlights the trade-off between paternalism and redistribution in the second-best economy: low-ability agents are bunched at an inefficiently high savings rate, while high-ability agents are separated by time preferences at lower savings rates. Intuitively, the planner offers choice in savings as a carrot and stick to incentivize work effort at high ability levels, thereby facilitating redistribution.

This theoretical characterization is useful because the optimal allocation can be decentralized as a competitive equilibrium given three realistic policy instruments: first, mandated old-age benefits as a function of income; second, a finite number of retirement accounts with different income-dependent subsidy rates and contribution limits; and third, a non-linear labor income tax. Intuitively, for high enough forced savings, low-income as well as impatient high-income agents will be constrained and rely only on mandated old-age benefits, whereas more patient high-income agents choose to sequentially exhaust the limits on subsidized retirement accounts. Qualitatively, this set of policy instruments resembles real-world retirement savings systems, such as Social Security plus 401(k) and various individual retirement arrangement (IRA) accounts in the US.

We apply this framework to quantitatively study the current US retirement savings and tax-transfer system vis-à-vis optimal policies in our model. This is a non-trivial task because failure of the linear independence constraint qualification in multi-dimensional screening problems renders numerical optimization routines unstable (Judd and Su, 2006). To overcome this problem, we develop a broadly applicable algorithm that efficiently solves high-dimensional non-linear optimization problems by finding the smallest set of binding constraints at the optimum. We first calibrate a positive version of our model to infer the joint distribution of time preferences and earn-

[2]That the planner respects a single time preference can be motivated by adopting an individual's perspective before the realization of a present bias shock, as in Amador et al. (2006). That welfare weights depend only on ability reflects a desire for income redistribution or insurance across ability types independent from the present bias shock.

ings ability using microdata on life-cycle income and wealth accumulation from the Health and Retirement Study (HRS) and the Current Population Survey (CPS). Identification of individual preference heterogeneity comes from differences in wealth at retirement conditional on life-time income. We then use our algorithm to solve the normative model and recover social preferences by extending the inverse-optimum approach (Bourguignon and Spadaro, 2012) to our setting. Finally, we combine the normative model with the inferred worker type distribution and social preferences in order to analyze optimal savings policies and quantify welfare gains from reforms to the current US system.

We present three main results from our quantitative analysis. First, we find substantial empirical heterogeneity in present bias and hence in implied optimal savings rates throughout the income distribution. Our calibration recovers annualized discount rates ranging from 0.905 to 0.999 between the 10th and the 90th percentile of the distribution, and a mild positive correlation of 0.10 between discount rates and income. In spite of this heterogeneity, the optimal savings rate at the bottom of the income distribution is uniformly set to 20 percent, even though the first-best rate is around to 16.5 percent. In contrast, savings rates for individuals earning USD 200,000 vary substantially between 15 and 21 percent.

Second, we discuss welfare implications of reforms to the current US savings and tax system as we uncover a tension between its two components. On one hand, the US tax-transfer system is best justified through welfare weights that are less redistributive than utilitarian (i.e. put more weight on high ability levels). On the other hand, our model rationalizes the large dispersion in savings rates at high incomes in the US as welfare weights that are more redistributive than utilitarian (i.e. put more weight on low ability levels). This is because the only reason a planner offers choice in savings is to facilitate redistribution. Hence, the current system is off the Pareto frontier, with 17.5 percent of consumption-equivalent welfare gains available from increasing mandatory savings and limiting savings choice, particularly at low incomes.

Third, we discuss implications for optimal savings instruments in our decentralization. Optimal contribution limits on retirement accounts are approximately affine in earnings. Individuals with annual incomes up to USD 65,000 receive only Social Security payments. Above that threshold, optimal savings vehicles include a "subsidized account" with a contribution limit of 1.8 percent of income, and a "tax-preferred account" with a limit of 3.7 percent. Further accounts have caps close to zero. Hence, a small number of accounts is sufficient to approximate the optimal

savings schedule. Optimal subsidy rates on retirement accounts in our decentralization are progressive. The "subsidized account" features a 30 percent subsidy that phases out to zero at around USD 15,000 in annual income before steadily increasing to a 25 percent tax rate at USD 200,000 in earnings. The second "tax-preferred account" is taxed at a rate that increases from 20 to 40 percent over the same income range. Finally, the tax on a regular savings account without cap is optimally set to approximately 45 percent.

Our main insight is more general than the application to savings policies. We characterize optimal choice architecture across income groups when private and social preferences disagree and tax revenues are valued. This formulation nests many behavioral and neoclassical problems. We discuss implications for Pigouvian taxation and quantity restrictions in their context.

Related literature. This paper contributes to three strands of the literature.[3] The first strand is concerned with the optimal taxation of capital. The classical result by Atkinson and Stiglitz (1972) implies that with agreement in preferences between the planner and agents only income, but not savings, should be distorted for redistributive purposes. Also relying on preference agreement is the zero long-run capital taxation proposition by Judd (1985) and Chamley (1986), subsequently revisited by Atkeson et al. (1999), Lansing (1999), Phelan and Stacchetti (2001), Hassler et al. (2008), Saez (2013), and Straub and Werning (2014). In our framework, paternalism provides an alternative motive for capital taxes or subsidies. Closely related to our work, Saez (2002), Diamond and Spinnewijn (2011), and Golosov et al. (2013) consider heterogeneous time preferences without paternalism and show that the correlation between discount factors and earnings ability matters for the optimal degree of capital taxation. Hosseini and Shourideh (2017) study optimal retirement policy reforms with heterogeneous mortality rates and time preferences. Relative to their work, a novel aspect of our paper is to consider the interaction between paternalism and redistribution with heterogeneity in both time preferences and earnings ability. In this setting, we find that the optimal dispersion of marginal capital tax rates is larger at high incomes.

The second literature that we relate to is the field of behavioral public finance, much of which has focused on optimal taxation without heterogeneity in behavioral biases and redistribution. For instance, O'Donoghue and Rabin (2003, 2006) and Gruber and Köszegi (2004) consider the incidence of linear consumption taxes when certain goods are either over-consumed (e.g. cigarettes)

[3] In Section 2.4, we further discuss our theoretical findings in light of some of the most related results in the literature.

or under-consumed (e.g. retirement savings). Farhi and Werning (2007, 2010), Pavoni and Yazici (2016), and Phelan and Rustichini (2016) study optimal estate taxation in a model with a single level of present bias. Lockwood and Taubinsky (2017) allow for non-linear labor earnings taxes and a linear tax on the "sin good." Amador et al. (2006) study an optimal delegation problem with a common level of present bias and find a minimum savings rule to be optimal.[4] In related work, Chetty et al. (2009), Beshears et al. (2015), and Farhi and Gabaix (2015) consider optimal policy design in the presence of behavioral agents but without redistribution. By allowing for transfers in such an environment, we highlight a novel trade-off due to the interaction between paternalism and redistribution. As a result, the optimal policy features over-saving at low incomes and differentially distorted savings decisions at high incomes. Our paper also complements recent work by Yu (2016) and Lockwood (2016), who focus on implications of present bias for income taxation under a redistributive motive. In contrast, our focus is on characterizing optimal savings policies.

The third strand of related work studies multi-dimensional screening problems. Rochet (1987), McAfee and McMillan (1988), Armstrong (1996), Rochet and Choné (1998), and Armstrong and Rochet (1999) emphasize challenges in the analysis of optimal contracts with higher-dimensional unobserved heterogeneity. Some important contributions in the field of public finance have made further progress. Kleven et al. (2009) analyze the optimal taxation of couples, while Rothschild and Scheuer (2013, 2015, 2016) characterize optimal income taxes under multi-dimensional skill heterogeneity. We contribute to this literature in two ways. First, we provide a partial characterization of the solution to a two-dimensional screening problem under the assumption of paternalism. Second, we develop a numerical algorithm that efficiently solves more general two-dimensional screening problems, making it potentially useful in a variety of other applications.

Outline. The paper is organized as follows. Section 2 characterizes the optimal savings problem with two-dimensional unobserved heterogeneity. Section 3 provides a decentralization of the optimal allocation using realistic policy instruments. Section 4 describes the numerical algorithm used to solve the model and calibrates it to US microdata in order to evaluate current retirement savings and tax policies. Section 5 generalizes our main theoretical result and discusses applications to behavioral and neoclassical problems. Finally, Section 6 concludes.

[4]Similar setups have been studied in the context of monetary policy (Athey et al., 2005), sovereign debt dynamics (Aguiar and Amador, 2011), fiscal rules (Halac and Yared, 2014), the market for commitment devices (Galperti, 2015), and parent-child relations (Doepke and Zilibotti, 2017).

2 Characterizing optimal paternalistic savings policies

2.1 Model setup

This section presents our benchmark model for the analysis of optimal savings policies when there is a dual concern about under-saving for retirement and income inequality. We present a two-period life-cycle model with two-dimensional unobserved heterogeneity in present bias and earnings ability.[5] Following the optimal taxation literature, we characterize the second-best allocation of resources in a mechanism design formulation of the problem without restricting ourselves to a any specific policy instruments.

A unit mass of agents live for two periods indexed by t, work and retirement, with common discount rate δ. Agents differ in two unobservable attributes. The first attribute is earnings ability, denoted $\theta \in \Theta = \{\theta_1, \ldots, \theta_N\}$, where $0 \leq \theta_1 < \ldots < \theta_N < +\infty$. The second attribute is the degree of *present bias*, denoted $\beta \in B = \{\beta_1, \ldots, \beta_M\}$ where $0 < \beta_1 < \ldots < \beta_M = 1$.[6] We understand β as a reduced-form placeholder for the disagreement between the planner and agents at the time of the savings decision, which may arise from a behavioral bias (Laibson, 1997) or from an externality problem absent government commitment (Sleet and Yeltekin, 2006). We do not impose restrictions on the distribution over agents' types, $\pi(\theta, \beta)$, other than assuming full support. In particular, our analysis does not rely on any particular correlation between θ and β.

Utility is defined over consumption c_t for $t = 1, 2$ and income y. The planner evaluates *experienced utility* of type (θ, β) according to

$$V(c_1, c_2, y; \theta) = u(c_1) - \frac{v(y)}{\theta} + \delta u(c_2)$$

where $u'(\cdot) > 0$, $u'(0) = +\infty$, $u''(\cdot) < 0$ and $v(0) = 0$, $v'(0) = 0$, $v'(\cdot), v''(\cdot) > 0$ for $c_1, c_2, y \geq 0$. Note that $V(c_1, c_2, y; \cdot)$ does not directly depend on β. At the time of choosing their savings, however, agents evaluate *decision utility* according to

$$U(c_1, c_2, y; \theta, \beta) = u(c_1) - \frac{v(y)}{\theta} + \beta\delta u(c_2)$$

[5] The essence of our theory is conveyed in a simple two-period model. In Appendix B, we characterize a multi-period life-cycle model with heterogeneity in hyperbolic discount factors.

[6] In our quantitative analysis in Section 4, we relax the assumption of $\beta \leq 1$ when estimating the distribution of present bias from microdata on life-time income and wealth accumulation. Furthermore, Section 5 presents a generalized model that allows for both excessive and insufficient action-taking.

Our preferred interpretation of this setup is that agents and the planner share a common evaluation of utility at time 0, but disagreement arises at time 1 due to heterogeneous degrees of present bias captured by the additional discount factor β (Amador et al., 2006). Finally, a storage technology transfers resources between periods at gross rate of return R.

Following Mirrlees (1971), we assume that the planner observes consumption and labor income but not agents' types directly, and designs the game to be played by agents in the economy. Although this game could take an arbitrary form, the Revelation Principle guarantees that it is sufficient to consider incentive compatible direct mechanisms, in which agents' payoffs depend only on their reported type. We call such this assignment rule an allocation and denote it $\mathcal{A} = \{c_1(\theta,\beta), c_2(\theta,\beta), y(\theta,\beta)\}_{(\theta,\beta)\in\Theta\times B}$. We characterize properties of the optimal allocation before showing how it can be decentralized using a set of realistic policy instruments.

An allocation satisfies *incentive compatibility (IC)* if using agents' decision utility we have

$$(\theta,\beta) = \arg\max_{(\theta',\beta')} U\left(c_1(\theta',\beta'), c_2(\theta',\beta'), y(\theta',\beta'); \theta, \beta\right) \quad \forall (\theta,\beta) \in \Theta \times B \tag{1}$$

An incentive compatible allocation can be implemented with agents truthfully reporting their types as an equilibrium strategy in the direct mechanism. An allocation is *feasible* if it satisfies

$$\sum_{(\theta,\beta)\in\Theta\times B} \pi(\theta,\beta)\left[y(\theta,\beta) - c_1(\theta,\beta) - \frac{c_2(\theta,\beta)}{R}\right] \geq 0 \tag{2}$$

A feasible allocation allows for transfers across types but restricts the planner's net budget balance to be weakly positive. We define *welfare* as agents' experienced utilities aggregated as

$$W\left(\{c_1(\theta,\beta), c_2(\theta,\beta), y(\theta,\beta)\}_{(\theta,\beta)\in\Theta\times B}\right) = \sum_{(\theta,\beta)\in\Theta\times B} \pi(\theta,\beta)\,\lambda(\theta)\,V(\theta,\beta) \tag{3}$$

where $\lambda(\theta) \geq 0$ are Pareto weights, normalized such that $\sum_{\theta,\beta}\pi(\theta,\beta)\lambda(\theta) = 1$. Consistent with our interpretation of paternalism, we assume welfare weights depend only on θ but not on β. In this environment, we define efficiency with respect to the planner's preference.

Definition 1. Given Pareto weights $\{\lambda(\theta)\}_{\theta\in\Theta}$, the planner's problem is to choose a *second-best* or *constrained efficient* allocation $\mathcal{A}^{**}$ that maximizes welfare (3) subject to IC (1) and feasibility (2). We say an allocation $\mathcal{A}^{*}$ is *first-best* or *efficient* if it maximizes welfare (3) subject to feasibility (2).

2.2 Example with 2×2 types

It is instructive to illustrate the model mechanics in a simple environment. Assume for now two levels of earnings ability and two levels of present bias. For simplicity, let $\beta \in \{\beta_L, \beta_H\}$ with $\beta_L < \beta_H = 1$, let $\theta \in \{\theta_L, \theta_H\}$ with $0 = \theta_L < \theta_H$, set $R\delta = 1$, and suppose the planner is weakly more redistributive than utilitarian, $\lambda(\theta_L) \geq \lambda(\theta_H)$.[7] Clearly, the first-best allocation features savings at a rate satisfying the planner's Euler equation, which implies $c_1 = c_2$, independent of β. We proceed step-wise in providing a full characterization of the second-best allocation.

Bunching at low ability. Because present bias levels do not enter the planner's objective, the first-best allocation treats identically agents with common θ but different β. As $u'(0) = +\infty$ and high-ability agents can work while low-ability individuals cannot, only the former has strictly positive labor income. Since income is observable, low-ability agents cannot pretend to work and hence the relevant IC constraints in θ-space are the ones from high ability to low ability: $u(c_1(\theta_H, \beta)) - v(y_1(\theta_H, \beta))/\theta_H + \beta\delta u(c_2(\theta_H, \beta)) \geq u(c_1(\theta_L, \beta')) + \beta\delta u(c_2(\theta_L, \beta'))$ for levels of present bias $\beta, \beta' \in \{\beta_L, \beta_H\}$. Assigning average utilities to low-ability agents trivially preserves IC among them. Because the the previous IC constraint is linear in utility levels $u(c_1(\theta_L, \beta'))$ and $u(c_2(\theta_L, \beta'))$ on the right-hand side, this also preserves IC between high and low ability levels: $u(c_1(\theta_H, \beta)) - v(y_1(\theta_H, \beta))/\theta_H + \beta\delta u(c_2(\theta_H, \beta)) \geq \bar{u}_1(\theta_L) + \beta\delta\bar{u}_2(\theta_L)$ for $\beta, \beta' \in \{\beta_L, \beta_H\}$, where $\bar{u}_t(\theta_L) = \sum_{\beta'} \pi(\beta' | \theta_L) u(c_t(\theta_L, \beta'))$. Therefore, it is incentive compatible for the planner to allocate $\bar{c}_t(\theta_L) = u^{-1}(\bar{u}_t(\theta_L))$ to all low-ability agents. Furthermore, this perturbation leaves welfare unchanged. However, strict concavity of u implies that such an allocation is strictly less costly, $\bar{c}_1(\theta_L) + \bar{c}_2(\theta_L)/R < \sum_{\beta'} \pi(\beta' | \theta_L)[c_1(\theta_L, \beta') + c_2(\theta_L, \beta')/R]$ whenever $c_t(\theta_L, \beta_L) \neq c_t(\theta_L, \beta_H)$ for some $t \in \{0, 1\}$.

In summary, offering the same allocation to low-ability types preserves IC, leaves welfare unchanged, but saves resources. We conclude that the planner optimally bunches low-ability agents: $(c_1(\theta_L, \beta), c_2(\theta_L, \beta)) = (c_1(\theta_L), c_2(\theta_L))$ for $\beta \in \{\beta_L, \beta_H\}$.

Separation at high ability. Should high-ability agents also be bunched? By way of contradiction, suppose that $(c_1(\theta_H, \beta), c_2(\theta_H, \beta), y(\theta_H, \beta)) = (c_1(\theta_H), c_2(\theta_H), y(\theta_H))$ for $\beta \in \{\beta_L, \beta_H\}$. There are two cases to consider.

[7] With $\theta_L = 0$ we mean the limiting case of low-ability agents not working, $y_L = 0$, and their disutility from work being $v(0)/0 = 0$ by L'Hôpital's rule.

In the first case, high-ability agents are bunched with $c_1(\theta_H) > c_2(\theta_H)$, illustrated as point A in Figure 1(a). At this point, the indifference curve of β_L-types is steeper as impatient types require relatively more period 2 consumption to compensate a given change in period 1 consumption. While continuing to offer allocation A, the planner can target patient agents by offering allocation B. Since points on the 45-degree-line minimize the cost of providing a given utility level, allocation B lies in the interior of the budget set. At high ability, β_H-types are indifferent between the allocations, while β_L-types strictly prefer A over B. In summary, offering allocation B in addition to allocation A preserves IC, leaves welfare unchanged, but saves resources—a contradiction.

In the second case, high-ability agents are bunched with $c_2(\theta_H) \geq c_1(\theta_H)$. Clearly, allocations with $c_2(\theta_H) > c_1(\theta_H)$ are dominated by one with $c_2(\theta_H) = c_1(\theta_H)$, illustrated as point D in Figure 1(b). By offering an additional allocation E with higher period 1 consumption, the planner can target impatient agents. Moving them toward this allocation, their IC constraints become slack and welfare decreases. In the second-best solution, the planner can then make (θ_H, β_L)-types work more and use those extra resources for redistribution while preserving IC. The first-order welfare gain from transfers to θ_L-types strictly exceeds the second-order welfare loss from allowing the deviation by (θ_H, β_L)-types whenever allocations D and E are close enough. In summary, this perturbation improves welfare while preserving IC and feasibility—a contradiction.

Combining both cases, we conclude that high-ability types are optimally separated by present bias: $(c_1(\theta_H, \beta_L), c_2(\theta_H, \beta_L)) \neq (c_1(\theta_H, \beta_H), c_2(\theta_H, \beta_H))$.[8]

Implied savings rates. Low-ability agents must be bunched such that $c_2(\theta_L) \geq c_1(\theta_L)$, or else moving them from point A to B in Figure 1(a) would leave welfare unchanged, preserve IC, but save resources. Suppose now that $c_2(\theta_L) = c_1(\theta_L)$, shown as point F in Figure 1(c). A similar argument as before shows that moving low-ability types in the direction of point G induces a second-order welfare loss but relaxes the IC constraint of (θ_H, β_L)-types to a first order, enabling net welfare gains from increased redistribution. It follows that the second-best solution to the planner's problem features $c_2(\theta_L) > c_1(\theta_L)$.

Inspection of Figure 1(a) also establishes that $c_2(\theta_H, \beta_H) \geq c_1(\theta_H, \beta_H)$ must be optimal. Fur-

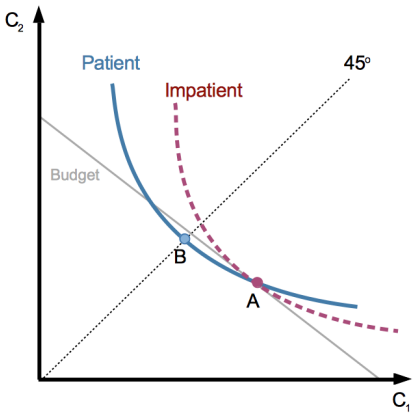
[8] All θ_H-types optimally have the same earnings. To see this, by way of contradiction and without loss of generality suppose that $y(\theta_H, \beta_L) < y(\theta_H, \beta_H)$. Then one could offer $\overline{y}(\theta_H) = \sum_{\beta'} \pi(\beta' | \theta_L) u(c_t(\theta_H, \beta'))$ to both θ_H-types, keeping resources constant. Due to convexity of v, (θ_H, β_L)-types can be compensated for working more by transferring resources to them from (θ_H, β_H)-types in a way to keep her welfare at the previous level, while saving resources—a contradiction. Hence, high-ability types share the same income: $y(\theta_H, \beta) = y(\theta_H, \beta) = y(\theta_H)$ for $\beta \in \{\beta_L, \beta_H\}$.

thermore, the only reason the planner would distort savings of (θ_H, β_H)-types is if their allocation were envied by (θ_H, β_L)-types. In Appendix A.1, we show that this cannot be the case and discuss relevant IC constraints at the solution to this problem. We conclude that savings of patient high-ability agents are optimally undistorted, $c_1(\theta_H, \beta_H) = c_2(\theta_H, \beta_H)$. Together with our earlier result on separation at high ability, this implies that (θ_H, β_L)-types in period 1 are allowed to over-consume, $c_1(\theta_H, \beta_L) > c_2(\theta_H, \beta_L)$.

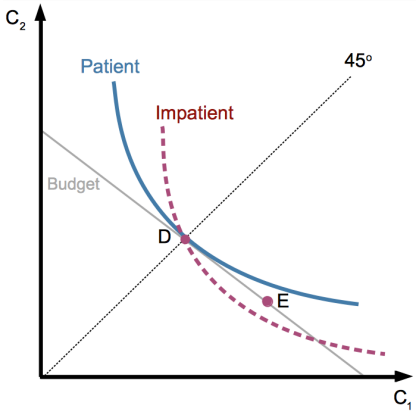
Summary. Intuitively, the planner offers choice as a screening device to identify high-ability types. At low ability, enforcing uniform savings is costless and setting the rate above the efficient level deters deviations by impatient high-ability types pretending to have low ability. At high ability, the cost of enforcing a given savings rate varies with individual levels of present bias. Consequently, savings by patient high-ability types are optimally left undistorted, while impatient high-ability types are allowed to over-consume in period 1.

Figure 1. Consumption perturbations

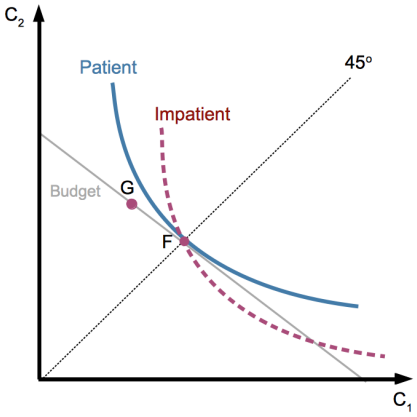
(a) Separation at high ability, case 1



(b) Separation at high ability, case 2



(c) Over-saving at low ability



2.3 General results

We now turn back to our benchmark model with more general heterogeneity in earnings ability and present bias. The usual approach in one-dimensional screening problems is to rely on the Spence-Mirrlees single-crossing condition and a monotonicity property of allocations in types to reduce the set of relevant IC constraints to only local constraints between bordering types. It is well-known that this approach does not extend to problems with multi-dimensional types (Rochet

and Stole, 2003; Battaglini and Lamba, 2015). We exploit the paternalistic formulation to provide a partial characterization of this problem for the top and bottom of the ability distribution.[9]

Bunching and separation. Our main result characterizes the extent to which the optimal allocation differentiates between different β-types throughout the ability distribution given a utilitarian or more redistributive planner.

Theorem 1. *Assume $\lambda'(\theta) \leq 0$ and fix $\{\theta_2, \ldots, \theta_{N-1}\}$. Then there exist scalars $\underline{\theta} > 0$ and $\overline{\theta} < +\infty$ such that at the solution to the planner's problem:*

1. *If $\theta_1 < \underline{\theta}$, then all types $\{(\theta_1, \beta) : \beta \in B\}$ are bunched, i.e. for $t = 1, 2$ and all $\beta \in B$:*

$$(c_t(\theta_1, \beta), y(\theta_1, \beta)) = (c_t(\theta_1), y(\theta_1))$$

2. *If $\theta_N > \overline{\theta}$, then types $\{(\theta_N, \beta) : \beta \in B\}$ are separated in their consumption, i.e. for some $\beta, \beta' \in B$:*

$$(c_1(\theta_N, \beta), c_2(\theta_N, \beta)) \neq (c_1(\theta_N, \beta'), c_2(\theta_N, \beta'))$$

Proof. See Appendix A.2.3. □

Intuitively, there is non-trivial interaction between heterogeneity in present bias and the redistributive motive. On one hand, the planner wants agents to save at a uniform rate given by the planner's Euler equation, $u'(c_1(\theta, \beta)) = R\delta u'(c_2(\theta, \beta))$. On the other hand, the planner wants to tailor consumption and labor allocations to each high-ability type separately so as to maximize redistribution toward lower ability types, which is valued under a redistributive motive.

The first part of Theorem 1 states that the planner finds it optimal to bunch the lowest-ability types regardless of their present bias level β. Clearly, such bunching is a feature of the first-best allocation. What we show is that at low enough ability levels it is approximately costless to enforce this feature, as the disutility of labor makes it costly for these agents to deviate toward higher ability levels. Since low-ability types are net transfer recipients the planner can enforce

[9] The assumption of a paternalistic planner reduces the dimensionality of the objective function—but not of the IC and feasibility constraints—and allows for a partial characterization of the solution to the planner's problem. Using numerical methods, we confirm in section 4 that those properties extend to the interior of the ability distribution.

an arbitrary savings level as long as this preserves higher ability types' downward-binding IC constraints. Consequently, bunching is optimal at the bottom of the ability distribution.[10]

The second part of Theorem 1 states that the highest-ability types are optimally separated across present bias levels β. From the agents' points of view, disagreement in time preferences between each other and with the planner imply differential distortions given any one allocation for those types. Although the planner would like them to adhere to the same consumption-savings schedule, at high ability levels the welfare gain from enforcing uniform savings is outweighed by the resource cost from lower output as a result of savings distortions. Instead, the planner can extract more resources from the highest-ability types when their downward-binding IC constraints are relaxed by offering savings options tailored to their individual levels of present bias. Given that Pareto weights are decreasing in ability, this can improve welfare by transferring extra resources toward lower ability types. As a result, the planner optimally allows for different allocations among the highest-ability types to facilitate redistribution.

Optimal savings distortions. We now characterize the nature of optimal savings distortions. To this end, it is useful to define two wedges that represent distortions at the solution to the planner's problem. First, we define the *decision wedge* to capture distortions in agents' view:

$$\tau^D(\theta,\beta) = 1 - \frac{u'(c_1(\theta,\beta))}{R\beta\delta u'(c_2(\theta,\beta))}$$

The decision wedge captures deviations from agents' Euler equations relating consumption between periods 1 and 2. Under laissez-faire, agents would choose their preferred savings rate and $\tau^D(\theta,\beta) = 0$. A negative decision wedge is akin to a positive implied tax on current consumption, which is associated with a higher savings rate than agents would pick in laissez-faire. Note that for any given allocation, the decision wedge differs across β-types.

Second, we define the *efficiency wedge* to capture distortions in the planner's view:

$$\tau^E(\theta,\beta) = 1 - \frac{u'(c_1(\theta,\beta))}{R\delta u'(c_2(\theta,\beta))}$$

The efficiency wedge measures deviations from the planner's Euler equation and thus the wel-

[10] In Appendix B, we present an interpretation of β as a hyperbolic discount factor in a dynamic environment and show that the planner optimally provides full insurance to lowest-ability types against their future time inconsistency.

fare losses relative to the efficient marginal rate of substitution. The first-best economy features $\tau^E(\theta, \beta) = 0$ for all types. A positive efficiency wedge means that agents over-consume in period 1 relative to period 2 when compared to the first-best savings rate. The decision wedge coincides with the efficiency wedge for $\beta = 1$. For all other β-types we have $\tau^D(\theta, \beta) < \tau^E(\theta, \beta)$, indicating that the planner wants agents to save at a higher rate for all $\beta < 1$.

By signing decision wedges and efficiency wedges across types at the solution to the planner's problem, we learn about the nature of distortions and inefficiencies that characterize the second-best allocation. The following result provides such a characterization given a utilitarian or more redistributive planner.

Theorem 2. *Assume $\lambda'(\theta) \leq 0$ and fix $\{\theta_2, \ldots, \theta_{N-1}\}$. Then there exist scalars $\underline{\theta} > 0$ and $\overline{\theta} < +\infty$ such that at the solution to the planner's problem:*

1. *If $\theta_1 < \underline{\theta}$, then:*
   - *$\tau^D(\theta_1, \beta) < 0$ for $\beta < 1$ and $\tau^D(\theta_1, \beta_M) \leq 0$;*
   - *$\tau^E(\theta_1, \beta) = \tau^E(\theta_1) \leq 0$ for all β;*
2. *If $\theta_N > \overline{\theta}$, then:*
   - *$\tau^D(\theta_N, \beta_M) = \tau^E(\theta_N, \beta_M) \leq 0$;*
   - *$\tau^E(\theta_N, \beta_1) > 0$.*

Proof. See Appendix A.2.4. □

Intuitively, Theorem 2 shows that savings distortions optimally vary throughout the income distribution. The interaction between paternalism and redistribution is again key to understanding this result. Dispersion in savings distortions is optimally used as an additional screening device when there is present bias heterogeneity. While the planner would like to correct savings throughout the ability distribution, a given savings distortion is more costly at higher ability levels. Thus the trade-off between paternalism and redistribution determines optimal decision wedges and efficiency wedges throughout ability distribution.

The first part of Theorem 2 states that the lowest-ability types experience an implied savings subsidy that is strictly positive for all β-types except for the most patient type whose subsidy

is weakly positive.[11] Furthermore, these implied savings subsidies are strong enough to move the lowest-ability types weakly above the first-best savings rate. As in our example with 2×2 types, the planner optimally uses high savings at the bottom as a screening device to encourage work effort at higher ability levels. This is optimal because the associated welfare costs from such distortions are of second-order while they relax downward-binding IC constraints to a first order. Furthermore, at low enough ability levels, the welfare gains from enforcing a high savings rate outweigh the resource cost of discouraging work effort at these levels. Consequently, a high savings rate is optimal among low-ability types.

The second part of Theorem 2 shows that top-ability types' savings rates vary between the planner's and agents' preferred rates. Agents with $\beta < 1$ face a strictly positive implied savings subsidy, while the implied subsidy for the most patient type is weakly positive. Along our previous intuition, the planner cares about correcting all agents' savings decisions, hence induces them to save more than they would in laissez-faire. But bringing their savings up to the efficient level is too costly, hence a strictly positive efficiency wedge remains for all but the most patient agent, whose efficiency wedge is weakly positive. For high ability levels, welfare losses due to inefficient savings are outweighed by the additional resources extracted from them as their IC constraints with respect to low ability types are relaxed. As a result, the planner optimally allows some of the top-ability types to save less than the first-best rate, though more than their preferred rate.

2.4 Comparison to most related results in the literature

We have argued that our main theoretical results arise from the interaction between present bias heterogeneity and the redistributive motive. To understand the forces in our model, it is instructive to relate our model to three influential results in the literature.

First, the intermediate goods taxation result of Atkinson and Stiglitz (1972) states that in a static environment without preference disagreement, intertemporal consumption decisions optimally remain undistorted even in the presence of a redistributive motive. Their result can be illustrated in our simple example with 2×2 types and $\beta = 1$ for all agents. In that model, the only dimension of heterogeneity is the level of earnings ability $\theta \in \{\theta_L, \theta_H\}$ where $\theta_L = 0 < \theta_H$. The only relevant IC constraint is then $u\left(c_1\left(\theta_H\right)\right) - v\left(y\left(\theta_H\right)\right)/\theta_H + \delta u\left(c_2\left(\theta_H\right)\right) \geq u\left(c_1\left(\theta_L\right)\right) +$

[11] In numerical simulations, we find a strictly positive decision wedge for (θ_1, β_M)-types and a strictly positive efficiency wedge for all (θ_1, β)-types to be robust features of the solution to the planner's problem.

$\delta u\left(c_2\left(\theta_L\right)\right)$. Since everyone agrees on the value of consumption over time, we can rewrite the planner's problem in its dual form as a resource cost minimization problem for each ability type: $\min_{c_1,c_2}\left\{c_1+\frac{1}{R}c_2\right\}$ s.t. $u\left(c_1\right)+\delta u\left(c_2\right)=\bar{U}\left(\theta\right)$, where $\bar{U}\left(\theta\right)$ depends on the optimal transfers across ability types taking into account IC. Taking first-order conditions of the above problem, we get $u'\left(c_1\left(\theta\right)\right)=R\delta u'\left(c_1\left(\theta\right)\right)$ and therefore $\tau^D\left(\theta_L\right)=\tau^D\left(\theta_H\right)=\tau^E\left(\theta_L\right)=\tau^E\left(\theta_H\right)=0$. Hence, redistribution without paternalism leads to undistorted savings.[12]

The second related result is that of Farhi and Werning (2010) who find the optimal efficiency wedge is monotonically increasing in ability under redistribution and a constant level of present bias. Their environment resembles our example with $\beta_L=\beta_H=\beta<1$ and $\theta\in\{\theta_L,\theta_H\}$ where $\theta_L=0<\theta_H$. The IC constraint is then $u\left(c_1\left(\theta_H\right)\right)-v\left(y\left(\theta_H\right)\right)/\theta_H+\beta\delta u\left(c_2\left(\theta_H\right)\right)\geq u\left(c_1\left(\theta_L\right)\right)+\beta\delta u\left(c_2\left(\theta_L\right)\right)$. Given $\lambda\left(\theta_L\right)\geq\lambda\left(\theta_H\right)$ and $0=y\left(\theta_L\right)<y\left(\theta_H\right)$, the IC constraint must bind at the solution, so the planner would like to redistribute more toward θ_L-types. The planner can improve upon the efficient savings rate by increasing θ_L-types' savings rate and decreasing type θ_H-types' savings rate. Both perturbations incur a second-order welfare loss but strictly relax the IC constraint, which facilitates redistribution and leads to a first-order net welfare gain. As a result, at the optimum, θ_L-type agents strictly over-save, $\tau^E\left(\theta_L\right)<0$, while θ_H-type agents strictly under-save, $\tau^E\left(\theta_H\right)>0$. Hence, the interaction between redistribution and a constant degree of paternalism gives rise to efficiency wedges that are strictly increasing in ability.

Third, the result in Amador et al. (2006) in a model without redistribution but with heterogeneity in present bias is also relevant to our analysis. They find that the optimal policy in this environment takes the form of a minimum savings threshold, which leaves patient agents' savings undistorted. Although for different reasons, the optimal policy emerging from our framework also entails greater dispersion in savings at higher ability levels. A unique feature of our environment relative to theirs is that at low ability bunching occurs above the first-best savings rate, while implied savings rates are differentially distorted at high ability.

Our model combines the two ingredients of redistribution and present bias heterogeneity. The forces in Atkinson and Stiglitz (1972), Farhi and Werning (2010), and Amador et al. (2006) are also present in our model and partially characterized by Theorems 1 and 2. As in Atkinson and

[12]Savings may of course be distorted for other reasons not present in our benchmark model, such as insurance in an incomplete markets environment with uninsurable income risk (Golosov et al., 2007). In Appendix B, we consider such a dynamic environment and extend our main results to this setting. The main insight emerging from this analysis is that the planner's inverse Euler equation replaces the static Euler equation in our formulation above.

Stiglitz (1972), the savings of (θ_N, β_M)-types are undistorted. As in Farhi and Werning (2010), the efficiency wedge is decreasing in ability. And as in Amador et al. (2006), low-ability types are optimally bunched regardless of present bias levels. Our model bridges these seminal results in the optimal taxation literature and brings to bear additional richness due to the interaction between redistribution and heterogeneity in present bias. In choosing the menu of allocations, the planner optimally offers a degenerate choice set to low-ability agents, but a tailored menu of differentially distorted choices to high-ability agents (Theorem 1), with lower savings distortions toward higher ability levels (Theorem 2).

3 Decentralization using realistic retirement savings policies

The previous section characterized features of the optimal allocation in an environment with present bias heterogeneity and redistribution. Next, we study the implications of these findings for the design of realistic policies. To this end, we equip a government with three instruments that resemble many real-world retirement savings systems.

The first instrument is a tax-financed old-age transfers as a function of life-time income, which agents cannot borrow against, such as Social Security in the US.[13] The second instrument is comprised of a finite set of retirement savings accounts with subsidies and contribution limits that depend on income. We interpret one of these accounts as a regular savings account with no cap. Mapping this into the real-world policies, we have in mind the multitude of direct contribution plans such as 401(k) and IRA accounts that feature a combination of tax-incentivized employer matching, tax-preferred treatment, and contribution limits that depend on income, in addition to a personal investment account. The third instrument is a non-linear income tax that depends on the set of retirement savings account used by the agent.

Given their earnings ability θ and one of M present bias levels β, agents take as given the savings and tax system and decide on how much to work, y, as well as the division of their net income between consumption and the set of available retirement savings accounts. They receive mandated old-age benefits $b(y)$, which we think of as their first (forced) savings account. In

[13] Indeed, the use of Social Security payment streams as collateral on loans is prohibited by federal law under Title II of the Social Security Act, Sec. 207. [42 USC. 407] (a). See also Feldstein and Liebman (2002) for a discussion of theoretical and empirical issues related to the Social Security program.

addition, they may save in retirement savings accounts indexed by $m = 2, \ldots, M$ with net rates of return $(1 - \tau_m(y))R$ and contribution limits $\bar{a}_m(y)$.[14] Retirement savings accounts are sorted in the generosity of their subsidy rates, $1 - \tau_2(y) \geq \ldots \geq 1 - \tau_M(y) \geq 1$, so that agents use account m only after exhausting contribution limits on the more generous accounts $2, \ldots, m-1$. While working, agents face a non-linear income tax $T_{M_0}(y)$, where M_0 denotes the least generous retirement savings account used by the agent.[15] The problem that an agent of type (θ, β) solves is summarized as follows:

$$\begin{aligned} \max_{c_1, c_2, y, M_0} \; & u(c_1) - \theta v(y) + \beta\delta u(c_2) \\ \text{s.t.} \quad & c_1 + \sum_{m=2}^{M_0} a_m = y - T_{M_0}(y) \\ & c_2 = b(y) + R \sum_{m=2}^{M_0} (1 - \tau_m(y)) a_m \\ & 0 \leq a_m \leq \bar{a}_m(y) \\ & M_0 = 1 + \sum_{m=2}^{M} \mathbf{1}[a_m > 0] \end{aligned} \tag{4}$$

Definition 2. A *competitive equilibrium with retirement savings policies* is a feasible allocation $\mathcal{A} = \{c_1(\theta, \beta), c_2(\theta, \beta), y(\theta, \beta)\}_{(\theta,\beta) \in \Theta \times B}$ that, given a set of retirement savings and tax-transfer policies $\left(\{\bar{a}_m(\cdot)\}_{m=2}^{M}, \{\tau_m(\cdot)\}_{m=2}^{M}, \{T_m(\cdot)\}_{m=1}^{M}, b(\cdot)\right)$, solves agents' problem (4).

In the spirit of Ramsey (1927), the planner takes as given agents' maximizing behavior and picks parameters on retirement savings and tax-transfer policies that yields the maximum welfare given social preferences δ and $\{\lambda(\theta)\}_{\theta \in \Theta}$. Such retirement savings policy instruments with realistic features are sufficient to implement the optimal allocation from the planner's problem.

Proposition 1. *The solution to the planner's problem can be decentralized as a competitive equilibrium with retirement savings policies. Assume $\lambda'(\theta) \leq 0$ and fix $\{\theta_2, \ldots, \theta_{N-1}\}$. Then there exist scalars $\underline{\theta} > 0$ and $\overline{\theta} < +\infty$ such that if $\theta_1 < \underline{\theta}$ and $\theta_N > \overline{\theta}$ then the decentralization satisfies:*

[14] We could extend our model by allowing for $\beta_M > 1$, which would require an additional minimum participation threshold for retirement savings accounts. Indeed, low-income individuals in the US are less likely to have access to 401(k) accounts with employer-matched contributions (Financial Engines, 2015).

[15] That income taxes depend on savings accounts use is not an unrealistic feature, given that contributions to 401(k) or IRA accounts, both Roth and regular, have differential tax treatments in the US. The ability to condition income taxes on savings account usage is a qualitatively important feature of the optimal policy at high incomes, as the planner wants to levy higher taxes on present-biased high-ability agents relative to more patient agents at the same ability level.

1. *Types θ_1 receive only old-age benefits and use none of the optional retirement savings accounts;*

2. *Types θ_N use retirement savings accounts in addition to receiving old-age benefits.*

Proof. Follows directly from the proofs of Theorems 1 and 2. □

The intuition behind the decentralization described above is straight-forward. The two key features that the optimal policy toolset needs to replicate are: first, bunching of low-income agents at a uniform, high savings rate; and second, strict separation in savings rates at higher incomes. The planner must replicate these features by picking policy parameters appropriately. The old-age benefits schedule must be generous enough relative to low incomes so as to constrain agents to be at a corner in their savings decision, unwilling to put additional funds in any of the retirement savings accounts. Old-age benefits must also be small enough relative to high incomes so as to allow those agents to self-select into the available retirement savings accounts. Agents with high θ will successively exhaust the contribution limits on the retirement savings account, starting with the most generous account and using the account with the next most generous subsidy rate after that.

We conclude that the optimal policy tools qualitatively resemble many real-world retirement savings systems, which feature forced savings at low incomes and a choice between multiple subsidized savings accounts toward higher income levels.

4 Quantitative exercise

This section evaluates through the lens of our normative model current US retirement savings policies in four steps. First, we develop a broadly applicable computational algorithm that allows us to numerically solve our framework with general two-dimensional heterogeneity. Second, we use a positive version of our consumption-savings model to recover the joint distribution of earnings ability and time preferences from microdata on life-time income and wealth accumulation under current US policies. Third, we adapt the inverse-optimum approach (Bourguignon and Spadaro, 2012) to select social preferences that most closely rationalize current US savings and tax policies. Finally, we use the calibrated normative framework to describe optimal savings policies and quantify welfare gains from reforms to the current system.

4.1 Numerical algorithm for solving multidimensional screening problems

Finding numerical solutions to two-dimensional screening problems has been recognized to be a difficult task (Rochet and Choné, 1998). This is partly because many of the techniques that one-dimensional optimization problems commonly rely on fail in a multi-dimensional context. Specifically, the first-order approach (Rogerson et al., 1985) does not extend seamlessly to multi-dimensional settings and thus global IC constraints may bind. This is not just a question of computational intensity. As Judd and Su (2006) point out, when the number of binding IC constraints at the optimum exceeds the number of choice variables then the linear independence constraint qualification (LICQ) fails, rendering Karush-Kuhn-Tucker conditions and other Lagrangian optimization routines unstable.[16] In a two-period environment with $N = N_\theta \times N_\beta$ types, which implies $N^2 - N$ global IC constraints and $3N$ choice variables, failure of the LICQ may occur with as few as $N = 6$ types and in practice occurs commonly for larger N. Consequently, standard optimization routines such as `fsolve` or `fmincon` in MATLAB fail to deliver reliable solutions.

We contribute to this literature a stable and computationally efficient numerical algorithm to solve a general class of multidimensional nonlinear optimization problems.[17] Our algorithm finds the smallest set of IC constraints sufficient to solve the global program. To this end, we first "convexify" our problem in utility space. We then initiate the algorithm by selecting a small set of IC constraints in addition to feasibility, which allows us to efficiently solve a relaxed program. Subsequently, we check global IC before iteratively adding and dropping constraints according to a stochastic rule based on the ranking of violations for excluded constraints, and Lagrange multipliers for included constraints at the solution to the relaxed program. Once we find a solution that satisfies global IC and feasibility, convexity of the problem guarantees that this be the globally unique solution. In theory, our algorithm converges to the global optimum with probability one, although convergence take finite but arbitrarily long time. In practice, we find that the algorithm converges quickly even for large-scale problems.[18] Appendix C gives further details of the com-

[16] Formally, the LICQ states that the gradients of the binding constraints at the solution are linearly independent. The LICQ is a sufficient condition for convergence in many numerical optimization algorithms and in many applications necessary for convergence or at least reasonable speed thereof.

[17] Our algorithm relies neither on the paternalistic nor the redistributive formulation of our problem and can be readily extended to more general settings.

[18] In all parameterizations of our problem, we find that a small fraction of all global constraints bind at the solution, allowing us to solve the problem using optimization routines that are robust to mild LICQ failures. The same solution may not obtain efficiently when attempting to solve the problem subject to the complete set of global IC constraints.

putational algorithm and demonstrates its application to a large-scale variant of our problem with $6\times1{,}000$ types, 18,000 choice variables, and 35,994,001 constraints.

4.2 Calibrating the joint distribution of earnings ability and time preferences

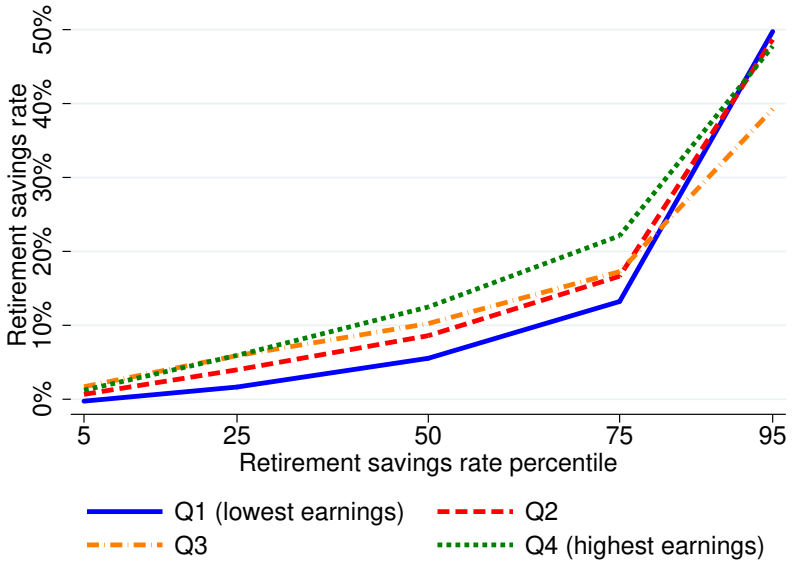
4.2.1 Calibration strategy

Identification. A key ingredient in our quantitative analysis of optimal savings policies is the joint distribution of earnings ability and time preferences. To identify this, we use data on *retirement savings rate*s, defined as the ratio of wealth at retirement to life-time earnings, reported in Engen et al. (2005) using the University of Michigan Health and Retirement Study (HRS). As will become clear shortly, through the lens of our model, this statistic is informative about an individual's propensity to save conditional on an earnings history.

The data show considerable heterogeneity in retirement savings rates both within and between life-time earnings groups. Figure 2 plots retirement savings rates as a function of savings rate percentiles (x-axis) and life-time earnings quartiles (colored lines). Three points are noteworthy. First, across income groups a substantial share of the population accumulate negligible net financial assets throughout their working life, with over one quarter of individuals entering retirement with less than five percent of life-time earnings. Second, there is substantial variation in retirement savings rates within income groups, ranging from close to zero to over 40 percent between the fifth and 95th percentiles of the savings rate distribution. Third, retirement savings rates show a mildly positive covariance with life-time earnings quartiles. For example, less than 20 percent of the highest earnings quartile individuals show at most a five percent retirement savings rate, while the same fraction is around 50 percent for the lowest earnings quartile.

It should be noted that our identification of time preferences is not free from potential criticism. For example, Aguiar and Hurst (2005) note that mismeasurement of home production can explain parts of the observed consumption drop upon retirement. A level shift in retirement wealth is not per se a problem for our analysis as our focus lies on heterogeneity. Potentially more problematic is dispersion in the reliance on unmeasured home production, which would tend to lead us to overestimate the variation in present bias conditional on earnings ability. However, we can allow for sizable degrees of measurement error in retirement wealth without changing our main conclu-

Figure 2. Distribution of retirement savings rates by income quartiles



Note: Each line represents the distribution of retirement savings rates within one life-time earnings quartile. Retirement savings rates are defined as the ratio of non-Social Security wealth at retirement to life-time earnings. Life-time earnings are computed given reported earnings history and estimates from Khitatrakun et al. (2000). Non-Social Security wealth includes all liquid wealth, deposits in retirement accounts, defined benefit plans, business equity, other real estate equity, and half of the primary home value. Source: Engen et al. (2005) using the 1992 HRS sample of households.

sions. Observed variation in retirement savings rates may also be due to efficient motives outside of our model, such as heterogeneous longevity risk (Pijoan-Mas and Ríos-Rull, 2014), long-term care risk (Ameriks et al., 2015), or dynastic precautionary savings (Boar, 2017). We can address this issue in two ways. First, the richness of the HRS data allows us to partially alleviate such concerns by controlling for a myriad of covariates including spousal characteristics, health status, inheritance values, retirement age, life expectancy, and degrees of risk aversion. Second, it would be straight-forward to extend the model to explicitly incorporate other dimensions of heterogeneity such as differences in life expectancy as a deterministic function of (θ, β).

In support of our interpretation of retirement savings rates as reflecting present bias, the HRS data show that households with below-median retirement savings rates are 73 percent more likely to have thought "hardly at all" about retirement and 25 percent less likely to have thought "a lot" about retirement. Hence the nature of the decision process differs across savings groups, in line with findings in the psychological science literature linking hyperbolic time discounting to lower cognitive skills (Burks et al., 2009).

Parameters. How does our model target the joint distribution of earnings ability and present bias? To this end, we parameterize household preferences by

$$U\left(c_1, c_2, y; \theta, \beta\right) = \frac{c_1^{1-1/\sigma} - 1}{1 - 1/\sigma} - \frac{\ell^{1+1/\gamma}}{1 + 1/\gamma} + \psi \frac{c_2^{1-1/\sigma} - 1}{1 - 1/\sigma}$$

where labor supply is $\ell = y/\theta$. We adopt as exogenous parameters a standard value for the intertemporal elasticity of substitution, $\sigma = 0.5$, and a Frisch elasticity of labor supply of $\gamma = 1$, which falls in the middle of macro- and micro-estimates in the literature (Rogerson, 1988; Chetty, 2012). As alluded to above, the effective discount factor $\psi = \beta\delta$ is identified off heterogeneity in empirical retirement savings rates conditional on life-time earnings. While β and δ are hard to identify separately in the data, in the next subsection we infer a value of δ as the discount factor embedded in current US retirement savings policies. In line with Amador et al. (2006), we view β in the two-period model as standing in for present bias in the fully dynamic model.

Before calibrating key model parameters, we approximate current US tax-transfer and retirement savings policies in our model. To this end, we model the current US tax-transfer system using a parsimonious approximation proposed by Feldstein (1969) and used in related work by Persson (1983), Benabou (2000), Heathcote et al. (2014), and Heathcote and Tsujiyama (2015). In this formulation, net transfers T depend on taxable income Y according to

$$T\left(Y; \lambda, \tau\right) = Y - \lambda Y^{1-\tau} \tag{5}$$

We adopt Heathcote et al. (2014)'s estimates of the level parameter λ and the progressivity parameter τ in equation (5) using Panel Study of Income Dynamics (PSID) data for 2002–2006. They find that $\tau = 0.151$ provides the best fit to the current US tax-transfer system, and $\lambda = 0.836$ balances the government budget. Furthermore, we model the current retirement savings system as a combination of old-age benefits and a number of savings vehicle subject to different income-specific subsidy rates and contribution limits. We model Social Security taxes and transfers using the 2014 income tax rate, a USD 118,500 earnings exemption threshold, and a replacement rate schedule of old-age transfers as a function of life-time income. We also integrate three different savings accounts: first, a 401(k) account allowing voluntary tax-deferred contributions up to $18,000 plus 50

percent employer-sponsored contribution matching (Financial Engines, 2015); second, an IRA account allowing for voluntary tax-deferred contributions up to $5,500; and third, a regular savings account with a real annual rate of return of $R - 1 = 3.44\%$ (Gourinchas and Parker, 2002).

Taking current policies as given, we calibrate the joint distribution of (θ, ψ) to match the joint distribution of earnings and retirement savings rates in the data. Specifically, we first map percentiles of the distribution of average life-time earnings between age 25 to 65 from the 2013 March Current Population Survey (CPS), denoted $\{\tilde{y}_i\}_i$, into the model earnings ability distribution. The distribution of life-time earnings in the CPS broadly conforms with that in the HRS but provides us with more precise estimates of earnings percentiles. We pick the distribution of earnings ability $\{\theta_i\}_i$ to approximate the earnings distribution by setting $\theta_i = \tilde{y}_i^{1+1/\gamma}$. We then map empirical retirement savings rates from the HRS data into model savings rates $\{s_i\}_i$, where $s_i = (c_{2,i}/R) \,/\, (c_{1,i} + c_{2,i}/R)$. To match retirement savings rates across income groups, we let the marginal distribution of effective discount factors be $\psi \sim Beta\,(a, b)$ with shape parameters $a > 0$ and $b > 0$ over ten discrete grid points. The Beta distribution is convenient for our purposes as it allows for asymmetry and it is bounded in $[0, 1]$. Importantly, we allow in our calibration for the possibility that $\beta > 1$, implying that agents are more patient than the planner.[19] We then define the joint distribution of earnings ability and present bias as the Gaussian copula between the marginal distributions of θ and ψ with correlation parameter ω. In practice, our calibration targets the 25th, 50th, and 75th percentiles of the distribution of retirement savings rates across quartiles of the lifetime earnings distribution, yielding a total of 12 targets.

4.2.2 Calibration results

Table 1 shows the results from our calibration exercise. Ability parameters $\{\theta_i\}_i$ match an earnings distribution with mean 56,164, while the 10th and 90th percentiles are 13,521 and 112,170, respectively, all reported in 2013 US dollars. The mean effective discount factor is $\psi_{annual} = 0.985$.[20] Our calibration also points to considerable heterogeneity in discount factors, but with 90 percent of mass between 0.905 and 0.999 in annualized terms. Finally, the Gaussian copula correla-

[19] The proposition that government is more myopic than its citizens is a common tenet in the political economy and international finance literatures (Aguiar and Amador, 2011; Halac and Yared, 2015).

[20] We annualize the discount factor ψ_{annual} assuming 40 periods of working life and 20 periods of retirement such that $\psi = \psi_{annual}^{40}\left(1 - \psi_{annual}^{21}\right) / \left(1 - \psi_{annual}^{41}\right)$. Our estimate corresponds to a mean compound private discount factor of $\psi = 0.322$ between work and retirement periods.

tion parameter between discount factors and earnings ability is estimated to be slightly negative, $\omega = -0.107$. But importantly this leads to a mildly positive correlation of 0.100 between discount factors and earnings, as individuals with a higher discount factor endogenously supply more labor. Overall, our present bias estimates fall within a range of empirically plausible estimates, suggesting that this parameter choice provides a reasonable basis for our optimal policy analysis.[21]

Table 1. Calibration results for joint distribution of earnings ability and discount factors

Description	Values
PANEL A. CALIBRATED PARAMETERS	
$\{\theta_i\}_i$	CPS earnings percentiles
a	0.855
b	1.795
ω	-0.107
PANEL B. IMPLIED MOMENTS	
Earnings y, mean	56,164
Earnings y, 10th percentile	13,521
Earnings y, 90th percentile	112,170
Annualized discount factor ψ_{annual}, mean	0.985
Annualized discount factor ψ_{annual}, 10th percentile	0.905
Annualized discount factor ψ_{annual}, 90th percentile	0.999
$Corr\left(y, \psi_{annual}\right)$	0.100

Note: The ability distribution $\{\theta_i\}_i$ is picked to match CPS earnings percentiles, reported in 2013 US dollars. Discount factors are distributed $\psi \sim Beta\left(a, b\right)$ and reported as an annualized discount factor ψ_{annual}, assuming 40 periods of working life and 20 periods of retirement, where $\psi = \psi_{annual}^{40}\left(1 - \psi_{annual}^{21}\right) / \left(1 - \psi_{annual}^{41}\right)$. The joint distribution is a Gaussian copula between marginal distributions of ability and discount factors with correlation parameter ω.

Table 2 reports the model fit vis-à-vis the data. Because we tie our hands with a sparse parameterization, the model cannot perfectly match the empirical discount factor distribution, although its overall shape is captured well. In our model as in the data, there are large differences in savings rates within life-time earnings quartiles, which our model generates through dispersion in present bias conditional on earnings ability. Our calibrated model also matches the positive gradient of savings rates across life-time earnings quartiles, captured by the correlation parameter ω. The

[21] While reduced form in nature, our estimates broadly conform with findings from a range of different settings. See Augenblick et al. (2015) and Beshears et al. (2015) for laboratory experiments; Ashraf et al. (2006), Tanaka et al. (2010), Jones and Mahajan (2015), and Kaur et al. (2015) for field experiments; and Laibson et al. (1998, 2017) for natural field data estimates of present bias. The positive gradient of estimated discount factors in income we find is in line with empirical correlations between life-time income and savings rates (Dynan et al., 2004) as well as with present bias estimates in Paserman (2008), Meier and Sprenger (2015), and Lockwood (2016). See also De Nardi and Fella (2017) for a comprehensive overview of dynamic quantitative models linking wealth heterogeneity to preference heterogeneity.

large amount of dispersion in empirical retirement savings rates leads us to estimate significant heterogeneity in present bias, both within and across income groups.

Table 2. Calibration fit to retirement savings rates by life-time earnings quartile

	Q1 (lowest)		Q2		Q3		Q4 (highest)	
Savings rate percentile	Data	Model	Data	Model	Data	Model	Data	Model
25	0.0165	0.0163	0.0398	0.0367	0.0590	0.0451	0.0593	0.0653
50	0.0554	0.0899	0.0860	0.1021	0.1024	0.1083	0.1248	0.1254
75	0.1322	0.1431	0.1664	0.1712	0.1726	0.1765	0.2211	0.1839

Note: Objective in calibration is to minimize L^2 norm between model and data statistics. Retirement savings rates are defined as the ratio of non-Social Security wealth at retirement to life-time earnings. Life-time earnings are computed given reported earnings history and estimates from Khitatrakun et al. (2000). Non-Social Security wealth includes all liquid wealth, deposits in retirement accounts, defined benefit plans, business equity, other real estate equity, and half of the primary home value. Source: Engen et al. (2005) using the 1992 HRS sample of households.

4.3 Inferring social preferences using the inverse-optimum approach

In this subsection, we compare current US retirement savings and tax policies with optimal policies arising from our normative model. In theory, we could feed any social preferences, consisting of a set of Pareto weights $\{\lambda(\theta_i)\}_i$ and discount factor δ, into our model for this policy analysis. In practice, a growing strand of the public finance literature uses the inverse optimum approach (Bourguignon and Spadaro, 2012; Heathcote and Tsujiyama, 2015; Lockwood and Weinzierl, 2016), which selects social preferences that most closely rationalize current real-world policies, as a natural starting point.[22] In line with this approach, we measure the distance of current policies from the Pareto frontier implied by our normative model, thus minimizing the welfare gains available from reforms to the current system. In a separate exercise, we repeat our analysis through the lens of a utilitarian planner, thus providing another popular benchmark in the literature.

Following Heathcote and Tsujiyama (2015), we parameterize Pareto weights across ability levels as $\lambda(\theta) = \exp(-\alpha\theta) / \left(\sum_{\theta',\beta'} \pi(\theta',\beta') \exp(-\alpha\theta')\right)$, where $\alpha \in \mathbb{R}$ indexes the government's redistributive motive. If $\alpha = 0$ then the government is utilitarian, while higher α imply a greater taste for redistribution. We estimate social preferences (α, δ) by solving our normative model over a grid of such duplets and then picking the combination that minimizes the L^2 norm between

[22] See Stantcheva (2016) for a discussion of some of the strengths and drawbacks of this approach.

allocations in the normative and positive versions of our model:[23]

$$(\alpha, \delta) = \arg\min_{(\tilde{\alpha}, \tilde{\delta})} \sum_{\theta, \beta} \pi(\theta, \beta) \left[\left(c_1^n(\theta, \beta; \tilde{\alpha}, \tilde{\delta}) - c_1^p(\theta, \beta) \right)^2 + \left(c_2^n(\theta, \beta; \tilde{\alpha}, \tilde{\delta}) - c_2^p(\theta, \beta) \right)^2 + \left(y^n(\theta, \beta; \tilde{\alpha}, \tilde{\delta}) - y^p(\theta, \beta) \right)^2 \right]$$

where superscript n denotes our normative model's optimal allocation as a function of $(\tilde{\alpha}, \tilde{\delta})$, and superscript p denotes allocations from our calibrated positive model given current US policies.

Table 3 summarizes our estimation results. We find that $\alpha = -0.601$ and $\delta_{annual} = 0.974$, with associated compound discount factor $\delta = 0.224$ between working life and retirement, lead the optimal allocation from our normative model to best approximate the allocation given current policies in our calibrated positive model. Together with our calibration for ψ this implies that the level of present bias, $\beta_{annual} = \psi_{annual}/\delta_{annual}$, has mean $\mathbb{E}[\beta_{annual}] = 1.011$, corresponding to a compound mean present bias level of $\mathbb{E}[\beta] = 1.436$ between two periods. Consistent with recent experimental evidence by Montiel Olea and Strzalecki (2014), we allow for some agents to be overly patient at the time of their savings decision and our estimates imply that these agents make up 57 percent of the population. At the same time, 43 percent of the population discount the future at a higher rate than the planner, indicating that these agents would save less than they do under current policies.

Table 3. Social preferences estimated using inverse optimum approach

Description	Values
PANEL A. ESTIMATED PARAMETERS	
Pareto weight curvature α	-0.601
Social discount factor δ	0.224
PANEL B. IMPLIED MOMENTS	
Pareto weight $\lambda(\theta)$, mean of θ	1.000
Pareto weight $\lambda(\theta)$, 10th percentile of θ	0.836
Pareto weight $\lambda(\theta)$, 90th percentile of θ	1.762
Present bias β, mean	1.436
Present bias β, 10th percentile	0.073
Present bias β, 90th percentile	2.219

Note: The redistributive parameter α guides the gradient of Pareto weights across ability levels according to $\lambda(\theta) = \exp(-\alpha\theta) / \left(\sum_{\theta', \beta'} \pi(\theta', \beta') \exp(-\alpha\theta') \right)$, normalized such that $\lambda(\mathbb{E}\theta) = 1.000$. The discount factor used to calculate social welfare $V(\cdot)$ is given by δ.

[23] As a robustness check we also searched for social preferences that minimize the consumption-equivalent welfare gain associated with moving from current to optimal policies, yielding qualitatively similar results.

4.4 Optimal policies and reforms to the current system

In this section, we use the calibrated normative model to compare optimal savings policies versus the current US retirement savings system.

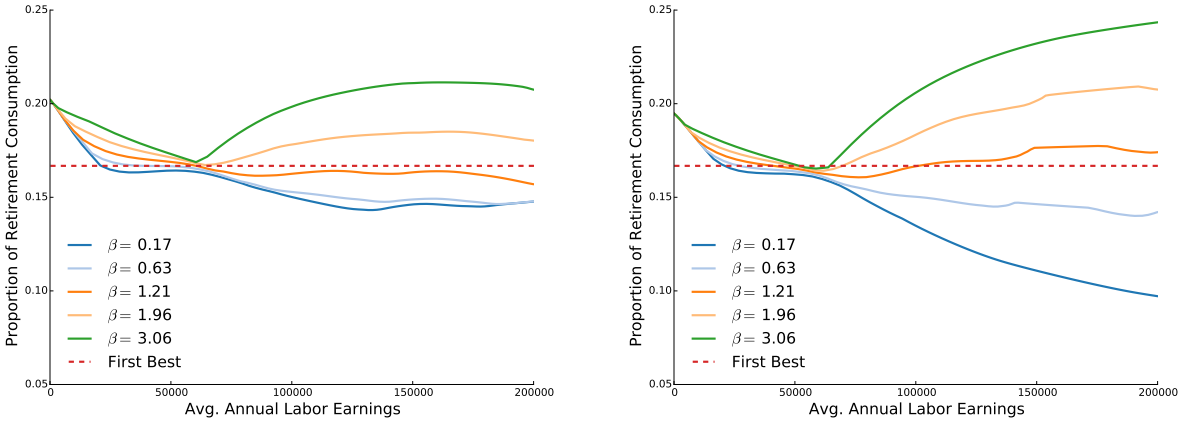
Redistributive motive and optimal dispersion in savings rates. How much choice in savings is optimally offered throughout the income distribution? Figure 3(a) plots as a function of income (x-axis) and present bias levels (colored lines) the optimal retirement savings rate, $s = (c_2/R) \,/\, (c_1 + c_2/R)$, from our calibrated normative model. As in Theorem 1 of our theoretical characterization, low earnings feature a uniform savings rate, while at high earnings optimal savings rates vary widely across present bias levels. As in Theorem 2, at incomes close to zero individuals optimally save at a 20 percent rate, which significantly exceeds percent the first-best savings rate of 16.6 percent. In contrast, savings rates for individuals earning USD 200,000 vary substantially across types, between 15 and 21 percent. Compared to the empirical savings rates in Figure 2, which span a wide array of savings rates in each income quartile, the optimal savings rates are uniformly higher and less dispersed, particularly at low incomes.

How is this pattern influenced by the planner's redistributive preferences? Figure 3(b) plots optimal retirement savings rates for a planner with utilitarian (i.e. more redistributive) preferences than in the benchmark while keeping all other parameters fixed. At low earnings, the level and dispersion of optimal retirement savings rates is similar to our benchmark calibration. At high earnings, however, greater taste for redistribution implies considerably more dispersion in optimal savings rates. For example, the range of optimal savings rate at high incomes increases substantially, varying between 10 and 25 percent. The planner uses flexibility in savings rates at high earnings to extract more resources for redistribution toward the bottom. Conversely, the more the planner cares about low-ability individuals, the lower the welfare losses from high-ability types deviating from the preferred savings rate. As a result, the optimal dispersion of savings rates at higher incomes is increasing in the planner's redistributive taste α. For higher α, the optimal savings rates more closely approximate empirical savings rates at high incomes, although low-income individuals still save too little relative to the social optimum.

Figure 3. Optimal retirement savings rates fan out under more redistributive welfare function

(a) Benchmark

(b) Utilitarian planner



Note: Proportion of retirement consumption is defined as $s = (c_2/R) \,/\, (c_1 + c_2/R)$. The horizontal axis shows annualized earnings during working life, while the vertical axis shows the optimal proportion of retirement consumption. Each line represent the allocation of agents with a given level of β. Benchmark is the main calibration described in the text, using less redistributive than utilitarian welfare weights. Utilitarian is attaching equal weights across ability types and using the same calibrated discount factor δ as in the benchmark calibration.

Quantitative analysis of optimal savings accounts. What are the quantitative features of optimal savings accounts, and how do they compare to the current US retirement savings system? Figure 4 plots the caps on retirement savings accounts arising from the benchmark normative model. Individuals with annual incomes up to USD 65,000 receive only Social Security payments. Above that threshold, optimal savings vehicles include a "subsidized account" with a contribution limit of around 1.8 percent of income, and a "tax-preferred account" with a limit of around 3.7 percent. Further accounts have caps close to zero. Hence, a small number of accounts is sufficient to approximate the optimal savings schedule.

Figure 5 plots optimal tax (if positive, or subsidy if negative) rates on savings across retirement savings accounts in the decentralization. Optimal subsidy rates are progressive (i.e. lower subsidies, or higher taxes, at higher incomes). The "subsidized account" features a 30 percent subsidy that phases out to zero around USD 15,000 in annual income before steadily increasing to a 25 percent tax rate at USD 200,000 in earnings. The second "tax-preferred account" is taxed at a rate that increases from 20 to 40 percent over the same income range. Finally, the tax on a regular savings account is optimally set around 45 percent. The presence of taxes on savings indicates that the planner in our benchmark normative model thinks that some individuals want to save too much.

Figure 4. Optimal contribution limits on retirement savings accounts



Note: Contribution limits are defined as upper bounds on savings accounts in the decentralization. The horizontal axis shows annualized earnings during working life, while the vertical axis shows contribution limits for each of three different retirement savings accounts. Each line represents one of three retirement savings accounts.

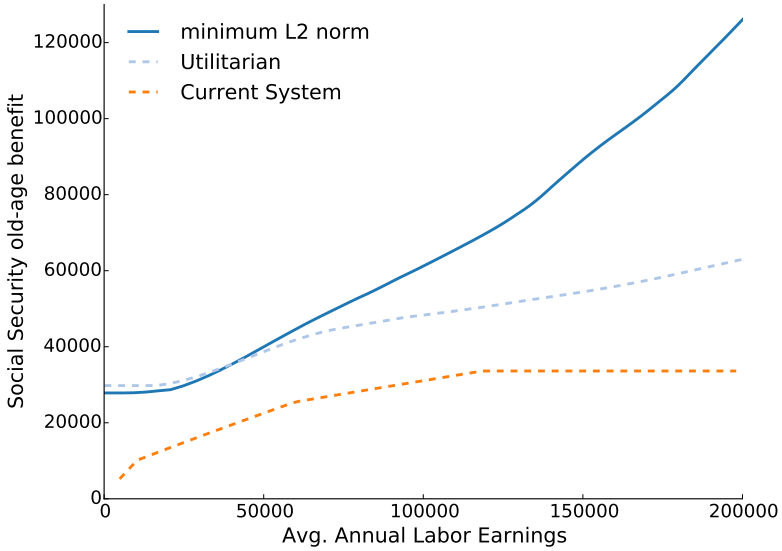
Figure 5. Optimal savings taxes on retirement savings accounts



Note: Savings tax (if positive) or subsidy (if negative) rates on retirement savings accounts. The horizontal axis shows annualized earnings during working life, while the vertical axis shows the savings tax rate. Each line represents one of the three retirement savings accounts and the regular savings account.

We now turn to the analysis of Social Security old-age benefits. Figure 6 compares current US Social Security old-age benefits (orange dashed line) as a function of annualized earnings (x-axis) with the normative model implications under the benchmark social preferences ($\alpha = -0.601$, solid blue line) and under utilitarian social preferences ($\alpha = 0$, dashed blue line). In our benchmark calibration, the planner cares relatively more about high-ability individuals and consequently would like them to save at much higher levels than currently embedded in the current Social Security benefits schedule. This is because under this parameterization there is little desire for transfers toward lower income levels and since lower incomes themselves generate relatively few resources that could be transferred upward, the planner's objective is primarily to get high-income individuals to save adequately for retirement. Utilitarian social preferences approximate the shape of old-age benefits reasonably well, but call for uniformly higher old-age benefits, that is a lower planner's discount factor.[24]

Figure 6. Fit of normative model versus positive model: old-age benefits as a function of income



Note: Annualized Social Security old-age benefits (y-axis) by annualized labor earnings level (x-axis). Each line represents the retirement benefits in a different system: current US Social Security old-age benefits (orange dashed line), normative model with benchmark social preferences ($\alpha = -0.601$, solid blue line), and normative model with utilitarian social preferences ($\alpha = 0$, dashed blue line). US Social Security old-age benefits are approximated using the Social Security Administration's Quick Calculator, taking into account contribution limits and decreasing replacement rates at higher earnings levels.

[24]Our estimated social discount factor of $\delta = 0.224$ was picked to approximate the consumption-savings allocation given real-world policies, not to match specific aspects of the current system.

Welfare gains from reforms. Our framework lends itself to analyzing jointly the current US retirement savings and tax-transfer system. We find large welfare gains from reforms to the current system as we uncover a tension between its two components. This tension arises because there exist no social preferences (α, δ) that jointly rationalize both systems.

On one hand, the US tax-transfer system is best justified through welfare weights that are less redistributive than utilitarian, i.e. put more weight on high ability levels. Indeed, Heathcote and Tsujiyama (2015) find that the US tax-transfer system is best rationalized by a value of $\theta = -0.566$, which is more redistributive than a laissez-faire planner ($\theta = -1$) but less redistributive than the utilitarian benchmark ($\theta = 0$). Those authors also show that the optimal system approximates well the current US tax-transfer system, with small welfare gains available from reforms.

On the other hand, an evaluation of the US retirement savings system through the lens of our model requires more redistributive than utilitarian welfare weights in order to match the shape of the Social Security benefits schedule (Figure 6), and the vast amount of choice in savings offered to high-income individuals (Figures 2 and 3(a)). To illustrate this divergence, we re-estimate social preferences to match the current US retirement savings system. We find that a more redistributive than utilitarian planner ($\alpha = 0.150$) with low social discount factor ($\delta = 0.067$, or $\delta_{annual} = 0.940$) closely approximates the current system of Social Security and retirement savings accounts on its own.[25] Again, the intuition for this result is that the only reason a planner would offer choice in savings is to facilitate redistribution of resources toward lower incomes.

Hence, independent of social preferences, the current system is off the Pareto frontier. Our welfare calculations should be taken with a grain of salt, as they measure the divergence between the planner's and agents' preferences over and above the classical measures in welfare analysis (Lucas, Jr., 1987; Krusell and Smith, 1999; Krusell et al., 2009), making our estimates not directly comparable to those calculations. With this qualification in mind, we find large divergence between allocations under the current policy system and optimal allocations from the calibrated normative model, amounting to 17.5 percent consumption-equivalent welfare gains in our benchmark calibration.[26] This welfare distance metric is even larger when adopting social preferences

[25] Relative to Figure 6, these social preferences shift down the Social Security old-age benefits scheduled plotted by the blue dashed line. Relative to Figure 3(b), these social preferences result in slightly higher dispersion in savings rates at top incomes, comparable to the empirical savings rates across income groups shown in Figure 2.

[26] We compute consumption-equivalent welfare gains as a uniform change in consumption during working life and retirement, holding fixed labor supply.

to match either one of the retirement savings or tax-transfer systems individually. Hence, there is a tension in reconciling the current US system through the lens of our model that allows us to jointly evaluate existing retirement savings and tax-transfer policies.

How much does a planner value simple reforms to the current system? To answer this question, we consider the following realistic policy instruments: a tax-transfer schedule with parameters λ and τ as in equation (5); uncapped social security contributions and benefits with a constant replacement rate $\gamma > 0$ at all earnings levels; a single retirement savings account with a minimum earnings eligibility threshold $y^{min} \geq 0$, a contribution limit as a share of labor income $a^{max}(y) = \bar{a}y$ for $\bar{a} \in [0, 1]$, and a government-sponsored matching rate of 50 percent; and a regular savings account subject to regular income taxes. Given our benchmark social preferences, we then solve the Ramsey problem of a planner equipped with these simple policy instruments.

We find that with these simple instruments the government is able to obtain sizable consumption-equivalent welfare gains of 7.0 percent, or 40 percent of the full optimum. In this exercise, the optimal policy involves significantly higher social security benefits ($\gamma = 1.94$) funded by more progressive income taxation ($\lambda = 0.74$ and $\tau = 0.41$). Individuals with incomes above y^{min} =34,775 US dollars start using the retirement savings account, which is capped at 0.6 percent of income. In essence, given our estimated Pareto weights, the government should increase old-age consumption and reduce savings choice throughout the earnings distribution, relative to the current US retirement system.

5 Generalization to other behavioral and neoclassical problems

Our main insights extend naturally into a variety of behavioral and neoclassical environments featuring two key ingredients. The first ingredient is heterogeneous disagreement between agents and the planner. Following Mullainathan et al. (2012), we summarize this as *preference wedges* between agents' and the planner's evaluations of the returns to some action. In a behavioral context, preference wedges may arise due to harm that agents inflict on themselves when they suffer from psychological biases leading to deviations from the rational choice paradigm. In a purely neoclassical context, preference wedges may arise when externalities lead one agent's actions to have unpriced effects on others. While both environments share the presence of a paternalistic

motive, the extent of disagreement between agents and the planner varies due to differences in the strength of behavioral biases, or due to differences in the extent to which economic context shapes individuals' incentives to engage in the production of externalities.

The second ingredient is the desire of the planner to collect revenues. In this context, the classical motive stressed by the public finance literature is redistribution. But proceeds from tax collection may also be used to finance public spending that is valued in the government's utility function. We will discuss below the case of redistribution, although our insights extend directly to other motives for revenue generation.

In a general environment combining these two ingredients, there is a trade-off between paternalism and redistribution. Our general result is that optimal policies involve a quantity restrictions at low earnings, whereas at high earnings individuals are given a choice between distorted options.

5.1 General environment

A continuum of consumers are characterized by earnings ability $\theta \in \Theta = \{\theta_1, \ldots, \theta_N\} \subseteq \mathbb{R}_+$ and their strength of temptation $\alpha \in A = \{\alpha_1, \ldots, \alpha_M\} \subseteq \mathbb{R}$ with $0 \in A$. They take an action $a = [0, 1]$ with associated unit cost p and income-equivalent benefit b that depends on a such that $b'(a) > 0$ and $b''(a) < 0$. We impose $b'(0) > p$ and $b'(1) < p$ to guarantee an interior solution.

A planner evaluates welfare according to agents experienced utility, which depends on consumption c and earnings y according to

$$\mathcal{U}^E(\theta) = u(c) - \frac{v(y)}{\theta}$$

where $u'(\cdot) > 0$, $u''(\cdot) < 0$, $v'(\cdot) > 0$,$v''(\cdot) > 0$, and $\lim_{y \to 0} v'(y) = 0$. Note that experienced utility depends on the action a only through its effect on consumption. Meanwhile, agents act according to their decision utility

$$\mathcal{U}^D(\theta, \alpha) = u(c) + \alpha\varepsilon(a) - \frac{v(y)}{\theta}$$

which depends on consumption c, earnings y and the payoff to a generic action a. Here, $\alpha\varepsilon(a)$ is a preference wedge that consists of two elements: first, the strength of temptation α, which we

allow to vary across individuals and be either be positive (leading to excessive action taking) or negative (leading to insufficient action taking); and second, the temptation utility $\varepsilon(a)$ such that $\varepsilon'(\cdot) > 0$, $\varepsilon''(\cdot) \leq 0$ and $b\left(\varepsilon^{-1}(\cdot)\right)$ is weakly concave.

For $\alpha = 0$ we have $\mathcal{U}^E = \mathcal{U}^D$ and there is agreement between the government and the agents. For $\alpha \neq 0$, an agent's private cost (or gain) from action a differs from the social cost (or gain) from that action. We assume agents types are unobservable and distributed according to $\pi(\theta, \alpha)$ with full support. An allocation $\{a(\theta,\alpha), c(\theta,\alpha), y(\theta,\alpha)\}_{(\theta,\alpha)\in\Theta\times A}$ is resource compatible if

$$\sum_{\theta,\alpha} \pi(\theta,\alpha)\left[y(\theta,\alpha) - c(\theta,\alpha) - pa(\theta,\alpha) + b(a(\theta,\alpha))\right] \geq 0$$

In this abstract formulation, the first-best level of the action a is given by, $a^* = b'^{-1}(p) \in (0,1)$, which does not vary with agents' type and is independent of redistribution.[27]

Consider first the laissez-faire economy without government intervention. Agents then choose (a, c, y) to maximize $\mathcal{U}^D(\theta,\alpha)$ subject to the budget constraint $y - c - pa + b(a) \geq 0$. Then agents with $\alpha \neq 0$ will choose a laissez-faire level of action $a^{LF}(\theta,\alpha) \neq a^*$ and $a^{LF}(\theta,\alpha) \neq a^{LF}(\theta,\alpha')$ for $\alpha \neq \alpha'$. Next, we consider optimal government intervention, starting with the case of no redistribution. In this case, it is straight-forward to show that the socially optimal policy is a quantity restriction of $a(\theta,\alpha) = a^*$, allowing agents to choose income y and consumption c to satisfy $y(\theta,\alpha) - c(\theta,\alpha) - pa^* + b(a^*) \geq 0$. In this case, since the government's preferred action is that of the laissez-faire economy with zero preference wedges, then the first best-level of the action does not depend on agent's types.

Finally, we consider the general case with redistribution under utilitarian or more redistributive welfare weights.

Theorem 3. *Assume $\lambda'(\theta) \leq 0$ and fix $\{\theta_2, \ldots, \theta_{N-1}\}$. Then there exist scalars $\underline{\theta} > 0$ and $\overline{\theta} < +\infty$ such that at the solution to the planner's problem:*

1. *If $\theta_1 < \underline{\theta}$, then all types $\{(\theta_1,\alpha) : \alpha \in A\}$ are bunched, i.e. for all $\alpha \in B$:*

$$(a(\theta_1,\alpha), c(\theta_1,\alpha), y(\theta_1,\alpha)) = (a(\theta_1), c(\theta_1), y(\theta_1))$$

[27] In the context of our two-period savings model in Section 2, the action corresponded to picking a savings rate $s = (c_2/R)/(c_1 + c_2/R)$.

2. *If $\theta_N > \overline{\theta}$, then types $\{(\theta_N, \alpha) : \alpha \in A\}$ are separated in their action and consumption, i.e. for some $\alpha, \alpha' \in A$:*

$$(a(\theta_N, \alpha), c(\theta_N, \alpha)) \neq (a(\theta_N, \alpha'), c(\theta_N, \alpha'))$$

Proof. See Appendix D. □

Theorem 3 extends our previous result on bunching and separation in the optimal savings problem (Theorem 1) to this general framework with preference wedges. Specifically, our main result extends to the case when there is temptation to take insufficient action ($\alpha < 0$), or temptation to take excessive action ($\alpha > 0$), or both at once.[28] As a corollary, a simple Pigouvian tax cannot achieve the social optimum. Instead, the efficient policy requires differential distortions throughout the income distribution, with a menu of choices offered at high income levels.

5.2 Applications to behavioral and neoclassical environments

We briefly discuss give applications of this general framework to both behavioral and neoclassical environments. In all of the following setups, we consider the problem of a planner with both paternalistic and redistributive preferences.

Example 1: Inattention and sales taxes (Chetty et al., 2009; Goldin and Homonoff, 2013; Goldin, 2015). Agents may purchase a units of a good with benefit $b(a) = a^{1-\gamma}/(1-\gamma)$ for some $\gamma > 0$. The unit cost $p + t$ consists of the gross price p plus sales tax t. But, to varying degrees, agents are inattentive to the sales tax, acting as if the net price of the good were $p + (1-\alpha)t$ for $\alpha \in [0, 1]$. Then $a^B(\alpha) > a^*$ for all $\alpha > 0$ and $\varepsilon(a) = at$ is the behavioral wedge.

Our theoretical result above implies that such inattention should be directly addressed for low-income shoppers (e.g. in basic grocery stores) but to a lesser extent for high-income shoppers (e.g. in luxury goods stores).

Example 2: Overconfidence and financial regulations (Malmendier and Tate, 2005; Scheinkman and Xiong, 2003). Let $b(a) = \mathbb{E}[aX] - Var[aX]$ be the mean-variance utility of an individual investing in a units of a risky asset X with unit price p. The true population parameters guiding the random return X are $\mathbb{E}[X] = \mu$ and $Var[X] = \sigma^2$. But individuals with varying degrees

[28] Note that the distortions we characterized for the savings framework in Theorem 2 depend on details of the environment in the generic preference wedge formulation.

overconfidence underestimate the riskiness of returns according to $Var^B[X] = (1-\alpha)\sigma^2$ for $\alpha \in [0,1]$. The preference wedge associated with overconfidence is then $\varepsilon(a) = -a^2\sigma^2$.

Suppose the government levies a financial transaction tax and uses the proceeds for public spending. An application of our result to this setting implies that the government wants to correct investment decisions of low-ability agents but differentially distort those of high-ability types.

Example 3: Habit formation and corrective policies (Bernheim and Rangel, 2004; Guo and Krause, 2011; Koehne and Kuhn, 2015). Let $b(a) = a^{1-\gamma}/(1-\gamma)$ with $\gamma > 0$ be the true utility from taking action a today. Behavioral agents experience additional disutility $\alpha\left(a - a^R\right)^2/2$, with the heterogeneous habit formation parameter α indexing the strength of the penalty from taking an action today that differs from the reference point $a^R \in [0,1)$. Agents choose a to maximize $\kappa a^\gamma - \alpha\left(a - a^R\right)^2/2$. If the unit price of the action is $p > 1$, then the socially optimal action is given by $a^* = p^{-1/\gamma}$, while behavioral agents with $\alpha > 0$ would pick $a^B(\alpha,\theta) < a^*$ in laissez-faire. The behavioral wedge in this environment is defined as $\varepsilon(a) = -\left(a - a^R\right)^2/2$.

Given both paternalistic and redistributive motives, low-ability types optimally take a uniform action a above the first-best level, while high-ability agents deviate toward their habit to varying degrees. In particular, it is never optimal to completely correct habits of high-ability types.

Example 4: Smoking and drug policies (Gruber and Köszegi, 2004; O'Donoghue and Rabin, 2006). Agents' action $a \in [0,1]$ represents cigarette consumption, associated with unit price $p > 0$ and social benefit $b(a) = 0$ so that from the government's perspective, the optimal level of consumption is $a^* = 0$. However, some individuals would like to consume cigarettes as they experience immediate gratification from smoking according to $\alpha\sqrt{a}$. Here, the preference wedges is simply $\varepsilon(a) = \sqrt{a}$ and $\alpha \geq 0$ guides agents' strength of their temptation.

The optimal drug policy is more restrictive towards smokers at lower earnings levels. Such a policy can be implemented by introducing a voucher-based system for drug usage with voucher cost assignments decreasing with individual earnings and high enough unit costs so that low-income individuals effectively do not smoke.

Example 5: Fuel efficiency and environmental policies (Sallee, 2011; Allcott et al., 2014; Golosov et al., 2014). In the context of environmental policies, let agents' action $a \in [0,1]$ represent the fuel efficiency of a purchased purchase. Let the socially optimal level of energy efficiency be $a^* \in [0,1]$.

Suppose the cost of a vehicle with energy efficiency $a \in [0,1]$ is pa. The social benefit a vehicle of type a is $b(a)$, which represents monetary benefits from economic activity net of pollution costs. Agents differ in their willingness to damage the environment, indexed by $\alpha \in [0, \overline{\alpha}]$, where $0 < \overline{\alpha} < +\infty$. The private benefit from purchasing a vehicle of type a is $(1+\alpha)\, b(a)$. Not all agents fully internalize the effects of pollution when purchasing a vehicle so that $\alpha > 0$ for some agents. The preference wedge in this environment is $\varepsilon(a) = \alpha b(a)$.

Without redistribution, a simple quantity restriction of $a(\theta, \alpha) = a^*$ for all individuals is socially optimal. In contrast, with redistribution, optimal dispersion in fuel efficiency varies along the income spectrum. At low earnings levels, the optimal policy induces agents to purchase the same energy efficiency level. Such a policy can be implemented with income tax rebates on vehicle purchases that depend both on the desired level of energy efficiency of the car and on individual earnings. At higher earnings levels, the government allows agents to enjoy energy inefficient vehicles in exchange for lower tax rebates. Hence, the government is willing to trade-off a higher level of externalities from high earnings agents in exchange for increased redistribution.

6 Conclusion

In this paper, we develop a normative theory of paternalistic policies. Our main insight is that the optimal policy restricts choice at low incomes but offers various distorted choices at higher incomes. Intuitively, the planner offers choice as a carrot and stick to incentivize work effort among high-ability individuals, thereby facilitating redistribution. We apply this insight to the study of optimal retirement savings systems. The optimal policy can be implemented through forced savings at low incomes—similar to Social Security—but a choice between savings accounts with different subsidies and caps at high incomes—like 401(k) and IRA accounts in the US.

Quantitatively, our calibrated model implies significant variation in the mean level as well as the dispersion of optimal savings rates throughout the income distribution. Relative to the current US retirement savings and tax-transfer system, we find large welfare gains from increasing mandatory savings and limiting savings choice, particularly at low incomes. We find this is due to a tension between redistributive preferences embedded in the current retirement savings system versus tax-transfer policies in the US. A small number of realistic retirement savings accounts with progressive subsidies and linear caps in income approximate well the optimal policy.

The theoretical insights and numerical solution method we develop in this paper open up the door to studying a wide class of multi-dimensional screening problems used in public finance, contract theory, and industrial organization. Our work points to two interesting avenues for future research. First, it would be interesting to explore to what extent other instances of fiscal, monetary, and social policies can be rationalized with a paternalistic motive. Second, while our current paper explores the implications of a given degree of paternalism for optimal policy design, future work could employ our framework to back out the implied degree of paternalism embedded in different policies within and across countries.

References

Aguiar, Mark and Erik Hurst, "Consumption Versus Expenditure," *Journal of Political Economy*, oct 2005, *113* (5), 919–948.

_ **and Manuel Amador**, "Growth in the Shadow of Expropriation," *Quarterly Journal of Economics*, 2011, *126* (2).

Allcott, Hunt, Sendhil Mullainathan, and Dmitry Taubinsky, "Energy Policy with Externalities and Internalities," *Journal of Public Economics*, 2014, *112*, 72–88.

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

Ameriks, John, Joseph Briggs, Andrew Caplin, Matthew Shapiro, and Christopher Tonetti, "Long-Term Care Utility and Late in Life Saving," Technical Report, National Bureau of Economic Research, Cambridge, MA feb 2015.

Armstrong, Mark, "Multiproduct Nonlinear Pricing," *Econometrica*, 1996, pp. 51–75.

_ **and Jean charles Rochet**, "Multi-dimensional screening: A User's Guide," *European Economic Review*, 1999, *43* (4), 959–979.

Ashraf, N., D. Karlan, and W. Yin, "Tying Odysseus to the Mast: Evidence From a Commitment Savings Product in the Philippines," *The Quarterly Journal of Economics*, may 2006, *121* (2), 635–672.

Athey, Susan, Andrew Atkeson, and Patrick J. Kehoe, "The Optimal Degree of Discretion in Monetary Policy," *Econometrica*, sep 2005, *73* (5), 1431–1475.

Atkeson, Andrew, Varadarajan Chari, and Patrick Kehoe, "Taxing Capital Income: A Bad Idea," *Quarterly Review*, 1999, (Sum), 3–17.

Atkinson, Anthony B and Joseph E Stiglitz, "The Structure of Indirect Taxation and Economic Efficiency," *Journal of Public Economics*, 1972, *1* (1), 97–119.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger, "Working over Time: Dynamic Inconsistency in Real Effort Tasks," *The Quarterly Journal of Economics*, aug 2015, *130* (3), 1067–1115.

Battaglini, Marco and Rohit Lamba, "Optimal Dynamic Contracting: The First-Order Approach and Beyond," *Economic Theory Center Working Paper*, 2015, (46-2012).

Benabou, Roland, "Unequal Societies: Income Distribution and the Social Contract," *American Economic Review*, 2000, pp. 96–129.

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

Beshears, John, James J Choi, Christopher Harris, David Laibson, Brigitte C Madrian, and Jung Sakong, "Self Control and Commitment: Can Decreasing the Liquidity of a Savings Account Increase Deposits?," Technical Report, National Bureau of Economic Research 2015.

Boar, Corina, "Dynastic Precautionary Savings," *University of Rochester Working Paper*, 2017.

Bourguignon, François and Amedeo Spadaro, "Tax-Benefit Revealed Social Preferences," *The Journal of Economic Inequality*, mar 2012, *10* (1), 75–108.

Buchanan, J. M., "The Samaritan's Dilemma," in E.S. Phelps, ed., *Altruism, Morality and Economic Theory*, New York: Russel Sage Foundation, 1975, pp. 71–85.

Burks, Stephen V, Jeffrey P Carpenter, Lorenz Goette, and Aldo Rustichini, "Cognitive Skills Affect Economic Preferences, Strategic Behavior, and Job Attachment," *Proceedings of the National Academy of Sciences of the United States of America*, may 2009, *106* (19), 7745–50.

Chamley, Christophe, "Optimal Taxation of Capital Income in General Equilibrium with Infinite Lives," *Econometrica*, may 1986, *54* (3), 607.

Chetty, Raj, "Bounds on Elasticities With Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply," *Econometrica*, 2012, *80* (3), 969–1018.

_ , Adam Looney, and Kory Kroft, "Salience and Taxation: Theory and Evidence," *American Economic Review*, aug 2009, *99* (4), 1145–1177.

Choi, James J., "Contributions to Defined Contribution Pension Plans," *Annu. Rev. Financ. Econ.*, 2015, *7*, 161–178.

De Nardi, Mariacristina and Giulio Fella, "Saving and Wealth Inequality," 2017.

Diamond, P. A. and J. A. Mirrlees, "A Model of Social Insurance with Variable Retirement," *Working papers*, 1977.

Diamond, Peter A, "A Framework for Social Security Analysis," *Journal of Public Economics*, 1977, *8* (3), 275–298.

_ , "Optimal Income Taxation: An Example With a U-shaped Pattern of Optimal Marginal Tax Rates," *American Economic Review*, 1998, pp. 83–95.

Diamond, Peter and Johannes Spinnewijn, "Capital Income Taxes with Heterogeneous Discount Rates," *American Economic Journal: Economic Policy*, nov 2011, *3* (4), 52–76.

Doepke, Matthias and Fabrizio Zilibotti, "Parenting with Style: Altruism and Paternalism in Intergenerational Preference Transmission," *Working Paper*, 2017.

Dynan, Karen E., Jonathan Skinner, and Stephen P. Zeldes, "Do the Rich Save More?," *Journal of Political Economy*, apr 2004, *112* (2), 397–444.

Engen, Eric M, William G. Gale, and Cori E. Uccello, "Lifetime Earnings, Social Security Benefits, and the Adequacy of Retirement Wealth Accumulation," *Social Security Bulletin*, 2005, *66* (1), 38–57.

Esteban, Susanna and Eiichi Miyagawa, "Optimal Menu of Menus with Self-Control Preferences," *Working Paper*, 2004.

Farhi, E. and I. Werning, "Insurance and Taxation over the Life Cycle," *The Review of Economic Studies*, apr 2013, *80* (2), 596–635.

Farhi, Emmanuel and Iván Werning, "Inequality and Social Discounting," *Journal of Political Economy*, 2007, *115* (3), 365–402.

_ **and** _ , "Progressive Estate Taxation," *The Quarterly Journal of Economics*, 2010, *125* (2), 635–673.

_ **and** _ , "Capital Taxation: Quantitative Explorations of the Inverse Euler Equation," *Journal of Political Economy*, jun 2012, *120* (3), 398–445.

_ **and Xavier Gabaix**, "Optimal Taxation with Behavioral Agents," Technical Report, National Bureau of Economic Research 2015.

Feldstein, Martin, "The Optimal Level of Social Security Benefits," *The Quarterly Journal of Economics*, 1985, *100* (2), 303–320.

_ **and Jeffrey B. Liebman**, "Social Security," *Handbook of Public Economics*, 2002, *4*, 2245–2324.

Feldstein, Martin S, "The Effects of Taxation on Risk Taking," *The Journal of Political Economy*, 1969, pp. 755–764.

Fernandes, Ana and Christopher Phelan, "A Recursive Formulation for Repeated Agency with History Dependence," *Journal of Economic Theory*, 2000, *91* (2), 223–247.

Financial Engines, "Missing Out: How much Employer 401(k) Matching Contributions Do Employees Leave on the Table?," Technical Report, Financial Engines 2015.

Galperti, Simone, "Commitment, Flexibility, and Optimal Screening of Time Inconsistency," *Econometrica*, 2015, *83* (4), 1425–1465.

Goldin, Jacob, "Optimal Tax Salience," *Journal of Public Economics*, 2015, *131*, 115–123.

_ **and Tatiana Homonoff**, "Smoke Gets in Your Eyes: Cigarette Tax Salience and Regressivity," *American Economic Journal: Economic Policy*, feb 2013, *5* (1), 302–336.

Golosov, Mikhail, Aleh Tsyvinski, and Ivan Werning, "New Dynamic Public Finance: A User's Guide," in "NBER Macroeconomics Annual 2006, Volume 21," MIT Press, 2007, pp. 317–388.

_ **and** _ , "Designing Optimal Disability Insurance: A Case for Asset Testing," *Journal of Political Economy*, apr 2006, *114* (2), 257–279.

_ **, John Hassler, Per Krusell, and Aleh Tsyvinski**, "Optimal Taxes on Fossil Fuel in General Equilibrium," *Econometrica*, 2014, *82* (1), 41–88.

_ **, Maxim Troshkin, Aleh Tsyvinski, and Matthew Weinzierl**, "Preference Heterogeneity and Optimal Capital Income Taxation," *Journal of Public Economics*, jan 2013, *97*, 160–175.

_ **,** _ **, and** _ , "Redistribution and Social Insurance," *American Economic Review*, feb 2016, *106* (2), 359–386.

_ **, Narayana Kocherlakota, and Aleh Tsyvinski**, "Optimal Indirect and Capital Taxation," *Review of Economic studies*, 2003, pp. 569–587.

Gourinchas, Pierre-Olivier and Jonathan A Parker, "Consumption Over he Life Cycle," *Econometrica*, 2002, *70* (1), 47–89.

Gruber, Jonathan and Botond Köszegi, "Tax Incidence When Individuals Are Time-Inconsistent: The Case of Cigarette Excise Taxes," *Journal of Public Economics*, aug 2004, *88*, 1959–1987.

Guo, Jang-Ting and Alan Krause, "Optimal Nonlinear Income Taxation with Habit Formation," *Journal of Public Economic Theory*, jun 2011, *13* (3), 463–480.

Halac, Marina and Pierre Yared, "Fiscal Rules and Discretion Under Persistent Shocks," *Econometrica*, 2014, *82* (5), 1557–1614.

_ and _ , "Fiscal Rules and Discretion in a World Economy," *NBER Working Papers*, 2015.

Hassler, John, Per Krusell, Kjetil Storesletten, and Fabrizio Zilibotti, "On the Optimal Timing of Capital Taxes," *Journal of Monetary Economics*, 2008, *55* (4), 692–709.

Heathcote, Jonathan and Hitoshi Tsujiyama, "Optimal Income Taxation: Mirrlees Meets Ramsey," 2015.

_ , Kjetil Storesletten, and Giovanni L Violante, "Optimal tax progressivity: An Analytical Framework," Technical Report, National Bureau of Economic Research 2014.

Hosseini, Roozbeh and Ali Shourideh, "Retirement Financing: An Optimal Reform Approach," *Working Paper*, 2017.

Jones, Damon and Aprajit Mahajan, "Time-Inconsistency and Saving: Experimental Evidence from Low-Income Tax Filers," Working Paper 21272, National Bureau of Economic Research jun 2015.

Judd, Kenneth and Che-Lin Su, "Optimal Income Taxation with Multidimensional Taxpayer Types," in "Computing in Economics and Finance," Vol. 471 2006.

Judd, Kenneth L., "Short-Run Analysis of Fiscal Policy in a Simple Perfect Foresight Model," *Journal of Political Economy*, apr 1985, *93* (2), 298–319.

Kapička, Marek, "Efficient Allocations in Dynamic Private Information Economies with Persistent Shocks: A First-Order Approach," *The Review of Economic Studies*, 2013, p. rds045.

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan, "Self-Control at Work," *Journal of Political Economy*, dec 2015, *123* (6), 1227–1277.

Khitatrakun, Surachai, Yuichi Kitamura, and John Karl Scholz, "Pensions and Wealth: New Evidence from the Health and Retirement Study," *University of Wisconsin Working Paper*, 2000.

Kleven, Henrik Jacobsen, Claus Thustrup Kreiner, and Emmanuel Saez, "The Optimal Income Taxation of Couples," *Econometrica*, 2009, *77* (2), 537–560.

Koehne, Sebastian and Moritz Kuhn, "Optimal Taxation in a Habit Formation Economy," *Journal of Public Economics*, 2015, *122*, 31–39.

Kotlikoff, Laurence J., "Privatizing Social Security at Home and Abroad," *American Economic Review*, may 1996, *86* (2), 368–372.

_ , Avia Spivak, and Lawrence H. Summers, "The Adequacy of Savings," *American Economic Review*, 1982, *72* (5), 1056–1069.

Krusell, Per and Anthony A Smith, "On the Welfare Effects of Eliminating Business Cycles," *Review of Economic Dynamics*, jan 1999, *2* (1), 245–272.

_ , Toshihiko Mukoyama, Aysegül Sahin, and Anthony A. Smith, "Revisiting the welfare effects of eliminating business cycles," *Review of Economic Dynamics*, jul 2009, *12* (3), 393–404.

Laibson, David, "Golden Eggs and Hyperbolic Discounting," *The Quarterly Journal of Economics*, 1997, *112* (2), 443–478.

_ , Andrea Repetto, and Jeremy Tobacman, "Self-Control and Saving for Retirement," *Brookings Papers on Economic Activity*, 1998, *29* (1), 91–196.

_ , Peter Maxted, Andrea Repetto, and Jeremy Tobacman, "Estimating Discount Functions with

Consumption Choices over the Lifecycle," *Working Paper*, 2017.

Lansing, Kevin J., "Optimal Redistributive Capital Taxation in a Neoclassical Growth Model," *Journal of Public Economics*, sep 1999, *73* (3), 423–453.

Lockwood, Benjamin B., "Optimal Income Taxation with Present Bias," *Working Paper*, 2016.

_ **and Dmitry Taubinsky**, "Regressive Sin Taxes," Technical Report, Working Paper 2017.

_ **and Matthew Weinzierl**, "Positive and Normative Judgments Implicit in U.S. Tax Policy, and the Costs of Unequal Growth and Recessions," *Journal of Monetary Economics*, feb 2016, *77*, 30–47.

Lucas, Jr., Robert E., *Models of Business Cycles*, New York: Basil Blackwell, 1987.

Malmendier, Ulrike and Geoffrey Tate, "CEO Overconfidence and Corporate Investment," *The Journal of Finance*, dec 2005, *60* (6), 2661–2700.

McAfee, R Preston and John McMillan, "Multidimensional Incentive Compatibility and Mechanism Design," *Journal of Economic Theory*, 1988, *46* (2), 335–354.

Meier, Stephan and Charles D. Sprenger, "Temporal Stability of Time Preferences," *Review of Economics and Statistics*, may 2015, *97* (2), 273–286.

Mirrlees, James A, "An Exploration in the Theory of Optimum Income Taxation," *The review of economic studies*, 1971, pp. 175–208.

Montiel Olea, José Luis and Tomasz Strzalecki, "Axiomatization and measurement of quasi-hyperbolic discounting," *The Quarterly Journal of Economics*, 2014, *129* (3), 1449–1499.

Mullainathan, Sendhil, Joshua Schwartzstein, and William J. Congdon, "A Reduced-Form Approach to Behavioral Public Finance," *Annual Review of Economics*, jul 2012, *4* (1), 511–540.

O'Donoghue, Ted and Matthew Rabin, "Studying Optimal Paternalism, Illustrated by a Model of Sin Taxes," *American Economic Review*, 2003, pp. 186–191.

_ **and** _ , "Optimal Sin Taxes," *Journal of Public Economics*, nov 2006, *90* (10-11), 1825–1849.

Paserman, M. Daniele, "Job Search and Hyperbolic Discounting: Structural Estimation and Policy Evaluation," *The Economic Journal*, aug 2008, *118* (531), 1418–1452.

Pavoni, Nicola and Hakki Yazici, "Intergenerational Disagreement and Optimal Taxation of Parental Transfers," *The Review of Economic Studies*, jul 2016, *23*, rdw036.

Persson, Mats, "The Distribution of Abilities and the Progressive Income Tax," *Journal of Public Economics*, 1983, *22* (1), 73–88.

Phelan, Christopher and Aldo Rustichini, "Pareto Efficiency and Identity," *Working Paper*, 2016.

_ **and Ennio Stacchetti**, "Sequential Equilibria in a Ramsey Tax Model," *Econometrica*, nov 2001, *69* (6), 1491–1518.

Pijoan-Mas, Josep and José-Víctor Ríos-Rull, "Heterogeneity in Expected Longevities," *Demography*, dec 2014, *51* (6), 2075–2102.

Prescott, Edward C., "It's Irrational to Save," *Wall Street Journal*, dec 2004.

Ramsey, F. P., "A Contribution to the Theory of Taxation," *The Economic Journal*, mar 1927, *37* (145), 47.

Rochet, Jean-Charles, "A Necessary and Sufficient Condition for Rationalizability in a Quasi-Linear Context," *Journal of Mathematical Economics*, jan 1987, *16* (2), 191–200.

_ **and Lars A. Stole**, "The Economics of Multidimensional Screening," in Mathias Dewatripont,

Lars Peter Hansen, and Stephen J. Turnovsky, eds., *Advances in Economics and Econometrics*, Cambridge: Cambridge University Press, 2003, pp. 150–197.

_ **and Philippe Choné**, "Ironing, Sweeping, and Multidimensional Screening," *Econometrica*, 1998, pp. 783–826.

Rogerson, Richard, "Indivisible Labor, Lotteries and Equilibrium," *Journal of Monetary Economics*, jan 1988, *21* (1), 3–16.

Rogerson, William P, Rogerson, and William P, "The First-Order Approach to Principal-Agent Problems," *Econometrica*, 1985, *53* (6), 1357–67.

Rothschild, Casey and Florian Scheuer, "Redistributive Taxation in the Roy Model," *The Quarterly Journal of Economics*, 2013, *128* (2), 623–668.

_ **and** _ , "A Theory of Income Taxation under Multidimensional Skill Heterogeneity," *CESifo Working Paper Series*, 2015.

_ **and** _ , "Optimal Taxation with Rent-Seeking," *The Review of Economic Studies*, jul 2016, *83* (3), 1225–1262.

Saez, Emmanuel, "Using Elasticities to Derive Optimal Income Tax Rates.," *Review of Economic Studies*, 2001, *68* (1), 205.

_ , "The Desirability of Commodity Taxation under Non-Linear Income Taxation and Heterogeneous Tastes," *Journal of Public Economics*, 2002, *83* (2), 217–230.

_ , "Optimal Progressive Capital Income Taxes in the Infinite Horizon Model," *Journal of Public Economics*, 2013, *97*, 61–74.

Sallee, James M., "The Taxation of Fuel Economy," *Tax Policy and the Economy*, sep 2011, *25* (1), 1–38.

Scheinkman, Jose A. and Wei Xiong, "Overconfidence and Speculative Bubbles," *Journal of Political Economy*, dec 2003, *111* (6), 1183–1220.

Sleet, Christopher and Sevin Yeltekin, "Credibility and Endogenous Societal Discounting," *Review of Economic Dynamics*, jul 2006, *9* (3), 410–437.

Social Security Administration, "Annual Statistical Supplement, 2016 - Summary of OASDI Benefits in Current-Payment Status (5.A)," 2017.

Stantcheva, Stefanie, "Optimal Taxation and Human Capital Policies over the Life Cycle," Technical Report, National Bureau of Economic Research 2015.

_ , "Comment on 'Positive and Normative Judgments Implicit in U.S. Tax Policy and the Costs of Unequal Growth and Recessions' by Benjamin Lockwood and Matthew Weinzierl," *Journal of Monetary Economics*, 2016, *77* (February), 48–52.

Straub, Ludwig and Iván Werning, "Positive Long Run Capital Taxation: Chamley-Judd Revisited," Technical Report, National Bureau of Economic Research, Cambridge, MA aug 2014.

Tanaka, Tomomi, Colin F Camerer, and Quang Nguyen, "Risk and Time Preferences: Linking Experimental and Household Survey Data from Vietnam," *The American Economic Review*, 2010, *100* (1), 557–571.

Tversky, A. and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science*, sep 1974, *185* (4157), 1124–1131.

Yu, Pei Cheng, "Optimal Retirement Policies," *Working Paper*, 2016.

Appendix

Outline. The Appendix is organized as follows. Appendix A presents proofs for the two-period model including our main theoretical results. Appendix B extends our main results to a multi-period life-cycle model. Appendix C describes details of the numerical solution algorithm. Finally, Appendix D contains the proof characterizing the generalized problem.

A Proofs for two-period model

A.1 Relevant IC constraints in a simple model with 2×2 types

In this section, we analyze which IC constraints bind in the simple model with 2×2 types presented in Section 2.2. We first show that the IC constraint of type (θ_H, β_L) with respect to low-ability types is binding. Assume by way of contradiction that the IC constraint of type (θ_H, β_L) with respect to low-ability agents' allocation does not bind. Then there is no reason to distort savings among low-ability types, and thus $c_1(\theta_L) = c_2(\theta_L)$ at the solution. We have already shown that $c_2(\theta_H, \beta_H) \geq c_1(\theta_H, \beta_H)$, therefore $c_2(\theta_H, \beta_H) \geq c_2(\theta_L)$ to preserve IC between the patient high-ability type and low-ability types. Together with the fact that the IC constraint of type (θ_H, β_H) with respect to low-ability types must bind in the second-best, this implies that

$$u(c_1(\theta_L)) + \beta_L \delta u(c_2(\theta_L)) \geq u(c_1(\theta_H, \beta_H)) - \frac{v(y_1(\theta_H, \beta_H))}{\theta_H} + \beta_L \delta u(c_2(\theta_H, \beta_H))$$

Consequently, all IC constraints for type (θ_H, β_L) are strictly slack—a contradiction. We conclude that the IC constraint of type (θ_H, β_L) must be binding with respect to θ_L-type agents.

Next, we show that the IC constraint of type (θ_H, β_L) is slack with respect to type (θ_H, β_H). Assume by way of contradiction that it binds. Since the IC constraint of type (θ_H, β_L) with respect to the θ_L-types' allocation is binding, we would have

$$u(c_1(\theta_H, \beta_H)) - \frac{v(y_1(\theta_H, \beta_H))}{\theta_H} + \beta_L \delta u(c_2(\theta_H, \beta_H)) = u(c_1(\theta_L)) + \beta_L \delta u(c_2(\theta_L))$$

The solution has $c_2(\theta_H, \beta_H) > c_2(\theta_L)$ or else $c_2(\theta_L) \geq c_2(\theta_H, \beta_H) > c_2(\theta_H, \beta_L)$, which cannot be

optimal with no IC constraints binding from low to high ability levels. Then we would have

$$u\left(c_1\left(\theta_H,\beta_H\right)\right)-\frac{v\left(y_1\left(\theta_H,\beta_H\right)\right)}{\theta_H}+\delta u\left(c_2\left(\theta_H,\beta_H\right)\right)>u\left(c_1\left(\theta_L\right)\right)+\delta u\left(c_2\left(\theta_L\right)\right)$$

But we know $c_2\left(\theta_H,\beta_H\right)>c_2\left(\theta_H,\beta_L\right)$ from strict separation of high-ability types, and therefore all IC constraints of type (θ_H,β_H) agents are strictly slack. This can not be optimal, as a transfer from type (θ_H,β_H) to low ability would improve welfare—a contradiction. We conclude that the IC constraint of type (θ_H,β_L) is slack with respect to type (θ_H,β_H).

Finally, the pattern of binding IC constraints for type (θ_H,β_H) depends on fundamentals. For $\beta_L\approx 1$, by continuity of the allocation in its primitives we have $c_2\left(\theta_H,\beta_L\right)>c_2\left(\theta_L\right)$, which combined with the fact that the IC constraint of type (θ_H,β_L) binds with respect to θ_L-types yields

$$u\left(c_1\left(\theta_H,\beta_L\right)\right)-\frac{v\left(y_1\left(\theta_H,\beta_L\right)\right)}{\theta_H}+\delta u\left(c_2\left(\theta_H,\beta_L\right)\right)>u\left(c_1\left(\theta_L\right)\right)+\delta u\left(c_2\left(\theta_L\right)\right)$$

In this case, the IC constraint of type (θ_H,β_H) with respect to low-ability types is slack. On the other hand, if $\beta_L\approx 0$ then for low enough $\lambda\left(\theta_H\right)$ we get $c_2\left(\theta_H,\beta_L\right)\approx 0<c_2\left(\theta_L\right)$, and the IC constraint of type (θ_H,β_H) with respect to low-ability types binds. Hence, the bindingness of (θ_H,β_H)-types' IC constraints is a function of model parameters.

It is worth noting that the indeterminacy of which IC constraints bind is precisely what renders solutions to multi-dimensional screening problems elusive (Armstrong, 1996; Rochet and Choné, 1998). Therefore, a complete theoretical characterizations of the solution to the class of problems that we study is infeasible. Conveniently, our characterization of savings rates throughout the income distribution does not depend on this particular feature of the solution.

A.2 Proofs of general two-period results

A.2.1 Problem reformulation

It is useful to restate the planner's problem in terms of utility levels by defining $u_t\left(\theta,\beta\right)=u\left(c_t\left(\theta,\beta\right)\right)$ for $t=1,2$, and $v\left(\theta,\beta\right)=v\left(y\left(\theta,\beta\right)\right)$. Let $c_t\left(\theta,\beta\right)=C\left(u_t\left(\theta,\beta\right)\right)$ where $C=u^{-1}$,

and $y(\theta,\beta) = Y(v(\theta,\beta))$ where $Y = v^{-1}$. Then the *planner's problem* can be stated as

$$\begin{aligned}
&\min_{u_1,u_2,v} -\sum_{\theta,\beta} \pi(\theta,\beta)\lambda(\theta)\left[u_1(\theta,\beta) - \frac{v(\theta,\beta)}{\theta} + \delta u_2(\theta,\beta)\right] \qquad (6)\\
&\text{s.t. } \left[u_1(\theta',\beta') - \frac{v(\theta',\beta')}{\theta} + \beta\delta u_2(\theta',\beta')\right] - \left[u_1(\theta,\beta) - \frac{v(\theta,\beta)}{\theta} + \beta\delta u_2(\theta,\beta)\right] \leq 0 \quad \forall(\theta,\beta),(\theta',\beta')\\
&\quad -\sum_{\theta,\beta}\pi(\theta,\beta)\left[Y(v(\theta,\beta)) - C(u_1(\theta,\beta)) - \frac{C(u_2(\theta,\beta))}{R}\right] \leq 0
\end{aligned}$$

Clearly, the objective and IC constraints are linear. Since $u(\cdot)$ is increasing and strictly concave and $v(\cdot)$ is increasing and strictly convex, then $C(\cdot)$ is strictly convex and $Y(\cdot)$ is strictly concave, so the feasibility constraint is strictly convex. Hence, the planner's problem (6) is a convex problem. This will be useful for the following proofs and also for our numerical solution algorithm.

A.2.2 Precursory results

We begin by proving three Lemmas that will be useful in the proofs of the main results. Define the utility of individuals as $U(\theta,\beta) = u_1(\theta,\beta) - v(\theta,\beta)/\theta + \beta\delta u_2(\theta,\beta)$ and the utility of the government as $V(\theta,\beta) = u_1(\theta,\beta) - v(\theta,\beta)/\theta + \delta u_2(\theta,\beta)$. The first Lemma shows that if the lowest labor earnings ability is sufficiently low, then agents with lowest earnings ability will have all their IC constraints with respect to higher ability types strictly slack.

Lemma 1. *Assume $\lambda'(\theta) \leq 0$. Given $\{\theta_2,\ldots,\theta_N\}$, there is $\underline{\theta} > 0$ such that if $\theta_1 < \underline{\theta}$, then at the solution to the planner's problem we have*

$$u_1(\theta_1,\beta) - \frac{v(\theta_1,\beta)}{\theta_1} + \beta\delta u_2(\theta_1,\beta) > u_1(\theta',\beta') - \frac{v(\theta',\beta')}{\theta_1} + \beta\delta u_2(\theta',\beta')$$

for $\theta' \in \{\theta_2,\ldots,\theta_N\}$.

Proof. Consider the case of $\theta_1 = 0$. Then for any $v_1(\theta_1,\beta) > 0$ we would have $V(\theta_1,\beta) = -\infty$. Since $\pi(\cdot)$ has full support and $\lambda(\cdot)$ is weakly decreasing, hence nonzero at θ_1, then $\pi(\theta_1,\beta)\lambda(\theta_1) > 0$ and this cannot be optimal. Therefore $v(\theta_1,\beta) = 0$, so θ_1-types optimally do not work at the solution to the planner's problem. Next, we show that all types $\theta' > 0$ work positive amounts. Assume by way of contradiction that $\mathrm{v}(\theta',\beta') = 0$ for some $\theta' > 0$ and $\beta' \in \{\beta_1,\ldots,\beta_M\}$. Since $Y'(0) = +\infty$ then this agent could work an infinitesimal amount and produce enough resources

to make all agents strictly better off—a contradiction.[29] Therefore $v(\theta', \beta') > 0$ and hence a deviation by θ_1-types into (θ', β')-types' allocation is not possible since $u_1(\theta', \beta') - v(\theta', \beta')/\theta_1 + \beta\delta u_2(\theta', \beta') = -\infty$. Hence, Lemma 1 holds for $\theta_1 = 0$.

By the Theorem of the Maximum, as we increase θ_1 the solution to the planner's problem is continuous in θ_1 and the above properties are preserved. Hence, there exists $\underline{\theta} > 0$ such that for all $\theta_1 < \underline{\theta}$ and for all (θ', β') with $\theta' > \theta_1$ and $\beta' \in B$ the solution to the planner's problem satisfies the desired property. □

Lemma 1 will be useful because for low enough ability types we need not worry about IC constraints binding upward in the ability dimension. The second Lemma uses a similar argument to show that if the highest ability type is sufficiently high, then IC constraints of all other agents with respect to that type are strictly slack.

Lemma 2. *Assume $\lambda'(\theta) \leq 0$. Given $\{\theta_1, \ldots, \theta_{N-1}\}$, there exists $\overline{\theta} < +\infty$ such that if $\theta_N > \overline{\theta}$ then at the solution to the planner's problem we have*

$$u_1(\theta', \beta') - \frac{v(\theta', \beta')}{\theta'} + \beta'\delta u_2(\theta', \beta') > u_1(\theta_N, \beta) - \frac{v(\theta_N, \beta)}{\theta'} + \beta'\delta u_2(\theta_N, \beta)$$

and

$$v(\theta_N, \beta) > v(\theta', \beta')$$

for all $\theta' \in \{\theta_1, \ldots, \theta_{N-1}\}$ and for all $\beta, \beta' \in \{\beta_1, \ldots, \beta_M\}$.

Proof. We use an analogous argument to that presented in the proof of Lemma 1. As $\overline{\theta} \to +\infty$, then $v(\theta_N, \beta) - v(\theta', \beta') \to +\infty$, since the relative social value from making θ_N-types work more tends to infinity, and hence upward-binding IC constraints are strictly slack for all $\theta' < \theta_N$. By a limiting argument and using continuity of the solution to the planner's problem, there exists $\overline{\theta} < +\infty$ that satisfies the desired properties. □

Lemma 2 will be useful because we can ignore IC constraints of lower ability types with respect to the highest ability types. The third Lemma shows that the most present-biased high-ability types are compensated for their higher work effort with higher consumption in either period.

[29] Formally, consider a perturbation to all agents such that $\tilde{v}(\theta, \beta) = v(\theta, \beta) + \varepsilon$ and $u_1(\theta, \beta) = u_1(\theta, \beta) + \nu$ for $\varepsilon, \nu > 0$. Since the original allocation satisfies IC and the perturbation is uniform, the perturbation also satisfies IC. The implied resource cost is $dE = \sum_{\theta,\beta} \pi(\theta, \beta) [C'(u_1(\theta, \beta))\nu - Y'(v(\theta, \beta))\varepsilon]$. Since $C'(u_1(\theta, \beta)) < \infty$, then for any $\varepsilon > 0$ we have $dE = -\infty$ as $Y'(0) = +\infty$. Welfare changes by $dW = \sum_{\theta,\beta} \pi(\theta, \beta) \lambda(\theta)(\nu - \varepsilon) = \nu - \varepsilon$. For $\nu > \varepsilon$ and small enough ν and ε, this perturbation is also feasible and increases welfare, a contradiction.

Lemma 3. *Assume $\lambda'(\theta) \leq 0$. Given $\{\theta_1, \ldots, \theta_{N-1}\}$, there exists $\overline{\theta} < +\infty$ such that if $\theta_N > \overline{\theta}$ then at the solution to the planner's problem we have either $u_1(\theta_N, \beta) > u_1(\theta', \beta')$ or $u_2(\theta_N, \beta) > u_2(\theta', \beta')$ for all $\theta' \in \{\theta_1, \ldots, \theta_{N-1}\}$ and $\beta, \beta' \in B$.*

Proof. By IC for θ_N-types we have

$$u_1(\theta_N, \beta) - \frac{1}{\theta_N}\left[v(\theta_N, \beta) - v(\theta', \beta')\right] + \beta\delta u_2(\theta_N, \beta) \geq u_1(\theta', \beta') + \beta\delta u_2(\theta', \beta')$$

From Lemma 2, there exists $\overline{\theta} < +\infty$ such that $v(\theta_N, \beta) > v(\theta', \beta')$ and hence

$$u_1(\theta_N, \beta) + \beta\delta u_2(\theta_N, \beta) > u_1(\theta', \beta') + \beta\delta u_2(\theta', \beta')$$

Therefore, the desired property holds. □

A.2.3 Proof of Theorem 1

Part 1.

Proof. From Lemma 1 we know that there exists $\underline{\theta} > 0$ such that

$$u_1(\theta_1, \beta) - \frac{v(\theta_1, \beta)}{\theta_1} + \beta\delta u_2(\theta_1, \beta) > u_1(\theta', \beta') - \frac{v(\theta', \beta')}{\theta_1} + \beta\delta u_2(\theta', \beta')$$

for all $\beta, \beta' \in \{\beta_1, \ldots, \beta_M\}$ and $\theta' \in \{\theta_2, \ldots, \theta_N\}$. Assume by way of contradiction that types $(\theta_1, \beta) \neq (\theta_1, \beta')$ receive different allocations. If $(u_1(\theta_1, \beta), u_2(\theta_1, \beta)) = (u_1(\theta_1), u_2(\theta_1))$ for all $\beta \in B$ then we must have $y(\theta_1, \beta) = y(\theta_1)$ by IC among θ_1-types. Hence the relevant case features $(u_1(\theta_1, \beta), u_2(\theta_1, \beta)) \neq (u_1(\theta_1, \beta'), u_2(\theta_1, \beta'))$ for some $\beta, \beta' \in B$. Then consider a perturbation

$$\tilde{u}_t(\theta_1, \beta) = \tilde{u}_t(\theta_1) = \sum_{\beta}\left(\frac{\pi(\theta_1, \beta)}{\sum_{\beta'}\pi(\theta_1, \beta')}\right) u_t(\theta_1, \beta)$$
$$\tilde{v}(\theta_1, \beta) = \tilde{v}(\theta_1) = \sum_{\beta}\left(\frac{\pi(\theta_1, \beta)}{\sum_{\beta'}\pi(\theta_1, \beta')}\right) v(\theta_1, \beta)$$

and keep all other allocations the same. Since all IC constraints are linear in u_t and v, a convex combination of their arguments preserves IC. Since the allocation for (θ_1, β) was initially different from that for (θ_1, β'), and since $C(\cdot)$ is strictly convex and $Y(\cdot)$ is strictly concave, the new allocation saves a strictly positive amount of resources while maintaining a constant welfare level.

But then the planner could improve welfare by distributing the extra resources uniformly across agents—a contradiction. Hence, agents of type (θ_1, β) for all $\beta \in \{\beta_1, \ldots, \beta_M\}$ are bunched. □

Part 2.

Proof. Assume by way of contradiction that all agents with types (θ_N, β) for $\beta \in B$ receive the same allocation, $(u_1(\theta_N), u_2(\theta_N), v(\theta_N))$. This implies that for some constant $\kappa > 0$:

$$\frac{R\delta C'(u_1(\theta_N))}{C'(u_2(\theta_N))} = \kappa$$

For $\kappa < 1$ (over-saving), consider the following perturbation, which keeps welfare constant:

$$\tilde{u}_1(\theta_N) = u_1(\theta_N) + \delta\varepsilon$$
$$\tilde{u}_2(\theta_N) = u_2(\theta_N) - \varepsilon$$

Agents with a present bias level β will perceive this perturbation as a decision utility change of

$$dU(\theta_N, \beta) = (1 - \beta)\delta\varepsilon$$

Whenever $\varepsilon > 0$ this perturbation is incentive compatible. Its marginal resource cost is

$$dE = \frac{\sum_\beta \pi(\theta_N, \beta) C'(u_2(\theta_N))}{R}(\kappa - 1)\varepsilon$$

If we had $\kappa < 1$ then this perturbation would generate extra resources for the government—a contradiction. Then it must be the case that $\kappa \geq 1$.

For $\kappa > 1$ (under-saving), consider a perturbation to the allocation offered to (θ_N, β_M)-types:

$$\tilde{u}_1(\theta_N, \beta_M) = u_1(\theta_N) - \delta\varepsilon$$
$$\tilde{u}_2(\theta_N, \beta_M) = u_2(\theta_N) + \varepsilon$$

This perturbation keeps welfare constant and preserves IC since agents with type $\beta < 1$ dislike the new allocation as long as $\varepsilon > 0$. The marginal resource cost is

$$dE = \frac{\pi(\theta_N, \beta_M) C'(u_2(\theta_N))}{R}(1 - \kappa)\varepsilon$$

Hence for $\varepsilon > 0$ we have $dE < 0$ since $\kappa > 1$, meaning that the planner generates extra resources while preserving IC and keeping welfare constant—a contradiction. We conclude that $\kappa = 1$.

Now consider the case when $\kappa = 1$ (saving at efficient rate). From Lemma 3 we have that either $u_1(\theta_N) > u_1(\theta, \beta)$ or $u_2(\theta_N) > u_2(\theta, \beta)$ for all $\theta \in \{\theta_1, \ldots, \theta_{N-1}\}$ and all $\beta \in B$. We prove the Theorem for the first case, as the second case is analogous. Consider the following perturbation:

$$\tilde{u}_1(\theta_N, \beta_1) = u_1(\theta_N) + \beta_1 \delta \varepsilon + \nu$$
$$\tilde{u}_2(\theta_N, \beta_1) = u_2(\theta_N) - \varepsilon$$

where $\varepsilon > 0$ and $\nu > 0$. While keeping all else unchanged, for $(\theta, \beta) \neq (\theta_N, \beta_1)$ we set

$$\tilde{u}_1(\theta, \beta) = \tilde{u}_1(\theta, \beta) + \nu$$

It is easy to check this preserves IC. The marginal resource cost of this perturbation is

$$dE = \pi(\theta_N, \beta_1)(\beta_1 - 1)\frac{C'(u_2(\theta_N))}{R}\varepsilon + \sum_{\theta,\beta} \pi(\theta, \beta) C'(u_1(\theta, \beta))\nu$$

By setting $dE = 0$ we get

$$\nu = \pi(\theta_N, \beta_1)(1 - \beta_1)\delta \frac{C'(u_1(\theta_N))}{\sum_{\theta,\beta} \pi(\theta, \beta) C'(u_1(\theta, \beta))}\varepsilon$$

The implied welfare change is then

$$dW = \pi(\theta_N, \beta_1)(1 - \beta_1)\delta \left[\frac{C'(u_1(\theta_N))}{\sum_{\theta,\beta} \pi(\theta, \beta) C'(u_1(\theta, \beta))} - \lambda(\theta_N)\right]\varepsilon$$

Since $\sum_{\theta,\beta} \pi(\theta, \beta)\lambda(\theta) = 1$ and $\lambda'(\cdot) \leq 1$ we have $\lambda(\theta_N) \leq 1$, which together with the fact that $u_1(\theta_N) > u_1(\theta, \beta)$ for all $\theta < \theta_N$ and $\beta \in B$ by Lemma 3 implies $dW > 0$. Hence this perturbation increases welfare, preserves IC, and is cost-neutral—a contradiction.

We conclude that not all agents with types (θ_N, β) for $\beta \in B$ receive the same allocation. □

A.2.4 Proof of Theorem 2

Part 1.

Proof. We begin by proving the results for θ_1-types. By Lemma 1, given $\{\theta_2, \ldots, \theta_N\}$, there exists $\underline{\theta} > 0$ such that for $\theta_1 < \underline{\theta}$ the IC constraints of type (θ_1, β) are strictly slack with respect to (θ', β')-types' allocations for $\theta' > \theta_1$. By Theorem 1, agents with type (θ_1, β) receive the same allocation for all $\beta \in B$, so for some constant $\kappa > 0$:

$$\frac{R\delta C'\left(u_1\left(\theta_1\right)\right)}{C'\left(u_2\left(\theta_1\right)\right)} = \kappa$$

Then consider the following perturbation to θ_1-types' allocation, which keeps the welfare constant:

$$\tilde{u_1}\left(\theta_1\right) = u_1\left(\theta_1\right) - \delta\varepsilon$$
$$\tilde{u}_2\left(\theta_1\right) = u_2\left(\theta_1\right) + \varepsilon$$

For $\varepsilon > 0$ sufficiently small, this perturbation is incentive compatible because IC constraints of types θ_1 were slack to begin with and all agents find the new allocation provides (weakly) lower utility than the old allocation. The marginal resource cost of this perturbation is

$$\begin{aligned} dE &= \sum_{\beta} \pi\left(\theta_1, \beta\right)\left[\frac{C'\left(u_2\left(\theta_1\right)\right)}{R} - \delta C'\left(u_1\left(\theta_1\right)\right)\right]\varepsilon \\ &= \sum_{\beta} \pi\left(\theta_1, \beta\right)\left(1-\kappa\right)\frac{C'\left(u_2\left(\theta_1\right)\right)}{R}\varepsilon \end{aligned}$$

If $\kappa > 1$ then for $\varepsilon > 0$ we have $dE < 0$, so that the perturbation generates extra resources—a contradiction. Hence $\kappa \leq 1$. Recalling the definition of $C\left(\cdot\right)$, we have $C'\left(\cdot\right) = 1/u'\left(C\left(\cdot\right)\right)$ and thus $\tau^E\left(\theta_1\right) \leq 0$. It follows that $\tau^D\left(\theta_1, \beta\right) < 0$ for $\beta < 1$ and $\tau^D\left(\theta_1, \beta_M\right) \leq 0$. □

Part 2.

Proof. We now turn to the results for θ_N-types. We first show that (θ_N, β_M)-types face a weakly negative decision wedge. Assume by way of contradiction that

$$\frac{R\delta C'\left(u_1\left(\theta_N, \beta_M\right)\right)}{C'\left(u_2\left(\theta_N, \beta_M\right)\right)} > 1$$

Consider the following perturbation to the allocation of (θ_N, β_M), while leaving all else unchanged:

$$\tilde{u}_1(\theta_N, \beta_M) = u_1(\theta_N, \beta_M) - \delta\varepsilon$$
$$\tilde{u}_2(\theta_N, \beta_M) = u_2(\theta_N, \beta_M) + \varepsilon$$

This perturbation is incentive compatible as agents with $\beta_M = 1$ are left indifferent, while agents with $\beta < 1$ prefer the original allocation. But this perturbation generates extra resources as

$$dE = \pi(\theta_N, \beta_M)\left(1 - \frac{R\delta C'(u_1(\theta_N, \beta_M))}{C'(u_2(\theta_N, \beta_M))}\right)\varepsilon$$

so for $\varepsilon > 0$ we have $dE < 0$ so the perturbation saves resources—a contradiction. Hence $\tau^D(\theta_N, \beta_M) = \tau^E(\theta_N, \beta_M) \leq 0$.

To show that $\tau^E(\theta_N, \beta_1) > 0$ we proceed analogously to the proof for Part 2 of Theorem 1. □

B Characterization of a multi-period life-cycle model

B.1 Model setup

In this section, we present a multi-period life-cycle model with stochastic earnings ability and self-control shocks. We characterize the efficient dynamic provision of insurance and commitment in this environment. Extending our results from the 2-period model, we show that a trade-off between providing insurance and providing commitment arises for agents who experience high income shocks, but not for agents with low income shocks. As a result, commitment is optimally provided only at low income levels.

In the following setup, we assume hyperbolic preferences shocks over the life cycle. While related work uses off-equilibrium path allocations to separate different degrees of time-inconsistency (Esteban and Miyagawa, 2004; Galperti, 2015; Yu, 2016), we effectively sidestep these intricacies by introducing stochastic time inconsistency levels. While studying a model with constant present bias is of great theoretical interest, our setup simplifies the analysis significantly and has two further advantages. First, our setup allows for changes in individuals' present bias over the life cycle, such as myopia that decreases with age. Second, our setup is robust to small stochastic perturbations in the hyperbolic discount factor, which the other setup abstracts from in order to generate

perfectly persistent private information.

The economy is composed of a measure one of agents whose life cycle consists of $T \geq 3$ periods, divided into T_w periods of working life and $T - T_w$ periods of retirement.[30] At each $t = 1, \ldots, T_w$, agents face an earnings ability shock $\theta_t \in \Theta = \{\theta_1, \ldots, \theta_N\}$, where $\theta_1 < \ldots < \theta_N$, with transition probabilities $\rho_{t+1}(\theta_{t+1}|\theta_t)$. We allow transition probabilities to vary over the life-cycle and and assume full support over Θ at all t and for all $\theta_t \in \Theta$. We also assume that ρ_{t+1} is stochastically ordered so that higher levels of θ_t imply a distribution that first order stochastically dominates a distribution for lower levels of θ_t. With a slight abuse of notation, we denote by $\rho_1(\theta_1)$ the probability distribution over the initial earnings ability θ_1 and assume that it also has full support.

Furthermore, At each period $t = 1, \ldots, T-1$ each agent faces a hyperbolic self-control shock $\beta_t \in B = \{\beta_1, \ldots, \beta_M\}$, where $\beta_1 < \ldots < \beta_M$, which we assume to be independently distributed both over time and from earnings ability shocks. We allow the probability distribution of self-control shocks at period t, denoted $\gamma_t(\beta_t)$, to vary over the life-cycle as long as there is full support.

We denote an agent's joint type by $h_t = (\beta_t, \theta_t) \in H_t$ and its distribution at time t by π_t. We mark by superscript t the history of types realized until period t, so that $h^t = (h_1, \ldots, h_t) \in H^t$. We let π_t denote the probability distribution over H^t.

The period payoff during working life periods $t = 1, \ldots, T_w$ over consumption and obtained earnings is given by $u_W(c_t, y_t; h_t) = u(c_t) - v(y_t)/\theta_t$, where we assume $u' > 0$, $u'' < 0$, $v' > 0$, $v'(0) = 0$, $v'' > 0$, and $v(0) = 0$. During retirement periods $t = T_w + 1, \ldots, T$, the agent is retired and consumes without working ($y_t = 0$), with period payoff given by $u_R(c_t) = u(c_t)$. The generalized period payoff function is then

$$u_t(c_t, y_t; h_t) = \begin{cases} u_W(c_t, y_t; h_t) & \text{for } t \leq T_w \\ u_R(c_t) & \text{for } t > T_w \end{cases}$$

A planner cannot directly observe agents' types but designs an incentive compatible and feasible mechanism that maximizes social welfare. As previously, we apply the Revelation Principle to characterize implementable allocations in this environment.[31] In this environment, an allocation

[30] We implicitly assume that retirement lasts for at least one period.

[31] We show in Appendix B.3 that it suffices to consider mechanisms in which at each period agent report their current type instead of their whole history of types. This result follows from our assumption that hyperbolic preference shocks are independent over time, and differs from the approach taken in Galperti (2015) and Yu (2016).

can be written as a sequence of functions $(c_t, y_t) : H^t \to \mathbb{R}^2_+$ for each t. We define an *allocation* as $\mathcal{A} = (c, y)$, where c and y denote the entire set of history-dependent consumption and labor allocations. The planner evaluates welfare according to the period 0 preferences, or *experienced utility*, of agents in the economy:

$$W_t(c, y) = \sum_{s=1}^{T} \delta^{s-1} \sum_{h^s} \pi_s(h^s)\, u_s(c_s(h^s), y_s(h^s); h_s) \tag{7}$$

Following a large strand in the behavioral public finance literature, we interpret this as the problem of an agent at period 0 seeking the optimal level of insurance for earnings ability shocks and a commitment device for self-control shocks over the life-cycle. Therefore, the efficient allocation could be implemented either by the government or by competitive private insurance companies, as long as both are able to enforce the contract.

Agents, once they reach the decision stage, have a present-biased evaluation of life-time utility, leading them to evaluate decision utility at time t as

$$U_t(c, y; h_\tau) = u_t(c_t(h^t), y_t(h^t); \theta_t) + \beta_\tau \sum_{s=t+1}^{T} \delta^{s-t} \sum_{h^s} \pi_s(h^s)\, u_s(c_s(h^s), y_s(h^s); \theta_s)$$

where we assume agents to be sophisticated in that they expect their future selves to be subject to some degree of present bias. Hence, there is dynamic disagreement between different period selves of the same (β, θ)-type as in Laibson (1997). A contract satisfies IC at time t if

$$h_t = \arg\max_{h'_t} U_t(c, y; h'_t) \tag{8}$$

We assume that there is a fixed gross rate of return R per period. A contract is *feasible* at time t if

$$\sum_{s=t}^{T} \frac{1}{R^{s-1}} \sum_{h^s} \pi(h^s) \left[y_s(h^s) - c_s(h^s) \right] \geq 0 \tag{9}$$

An allocation is *implementable* if it satisfies both IC (8) and feasibility (9) in all periods t.

Definition 3. The planner's problem is to choose a *second-best* or *constrained efficient* allocation $\mathcal{A}^{**}$ that maximizes welfare (7) subject to being implementable. We say an allocation $\mathcal{A}^*$ is *first-best* or *efficient* if it maximizes welfare (7) subject to feasibility (9) at all t.

B.2 General results

We now characterize properties of the planner's problem solution, which provides the efficient balance of insurance against earnings ability shocks and self-control shocks. By insurance of self-control shocks we mean that in any period t agents with lack of self-control ($\beta_t < 1$) and agents with self control ($\beta_t = 1$) are assigned the same allocation conditional on the whole history of earnings ability shocks they reported. Therefore, it is natural to interpret insurance of self-control shocks as provision of commitment by the planner.

Bunching and separation. Our first main result extends Theorem 1 to this dynamic economy, showing that full commitment is provided only to parts of the population.

Theorem 4. *Fix $\{\theta_2, \ldots, \theta_{N-1}\}$ and $\{\beta_2, \ldots, \beta_M\}$. Then there exist scalars $\underline{\theta} > 0$, $\overline{\theta} < +\infty$, and $\underline{\beta} > 0$ such that at the solution to the planner's problem:*

1. *If $\theta_1 < \underline{\theta}$, then for any $t = 1, \ldots, T-1$ and history h^{t-1} agents with types $\left\{\left(h^{t-1}, (\theta_1, \beta)\right) : \beta \in B\right\}$ are all assigned the same level of consumption and earnings in period t and are assigned the same continuation allocation for all future periods:*

$$\begin{aligned} c_t\left(h^{t-1}, (\theta_1, \beta)\right) &= c_t\left(h^{t-1}, (\theta_1, \beta')\right) \\ y_t\left(h^{t-1}, (\theta_1, \beta)\right) &= y_t\left(h^{t-1}, (\theta_1, \beta')\right) \\ c_{t+s}\left(h^{t-1}, (\theta_1, \beta), (h_{t+1}, \ldots, h_{t+s})\right) &= c_{t+s}\left(h^{t-1}, (\theta_1, \beta'), (h_{t+1}, \ldots, h_{t+s})\right) \\ y_{t+s}\left(h^{t-1}, (\theta_1, \beta), (h_{t+1}, \ldots, h_{t+s})\right) &= y_{t+s}\left(h^{t-1}, (\theta_1, \beta'), (h_{t+1}, \ldots, h_{t+s})\right) \end{aligned}$$

   for all $\beta, \beta' \in B$ and for all $s \geq 1$;

2. *If $\theta_N > \overline{\theta}$ and $\beta_1 \leq \underline{\beta}$, then for any $t = 1, \ldots, T-1$ and history h^{t-1} not all agents with types $\left\{\left(h^{t-1}, (\theta_N, \beta)\right) : \beta \in B\right\}$ are assigned the same current allocation and continuation allocations.*

Proof. See Appendix B.3.3. □

The planner values insurance against both earnings ability shocks and self-control shocks. Theorem (4) shows that it is efficient to provide perfect commitment at low earnings but not at high

earnings. This result is due to the interaction between the planner's two motives. Agents value flexibility after the realization of a self-control shock, demanding more immediate gratification than their prior selves' plans. Without an insurance motive, the planner would provide no such flexibility and instead provide commitment to all agents.[32] However, the planner also pursues the motive of consumption insurance, which in the presence of asymmetric information will be imperfectly provided. Therefore, the planner can charge high-ability agents for flexibility and use the proceeds to improve insurance against labor earnings shocks accrued to lower-ability agents. At low earnings, such a trade is feasible but not optimal since low-ability agents are unable to compensate the planner for the welfare loss associated with flexibility.

Optimal savings distortions. Our second result characterizes the distortions of time-inconsistent agents in this dynamic environment. A natural measure of distortions is the wedge relative to a path of time-consistent intertemporal consumption decisions. Without self-control shocks ($\beta = 1$), efficient insurance implies that intertemporal choices satisfy an *inverse Euler equation*:[33]

$$\sum_{\theta_{t+1} \in \Theta_{t+1}} \rho_{t+1}\left(\theta_{t+1} | \theta_t\right) \frac{u'\left(c_t\left(\theta^t\right)\right)}{\delta R u'\left(c_{t+1}\left(\theta^t, \theta_{t+1}\right)\right)} = 1$$

Whenever this intertemporal condition holds, the detrimental effects of time-inconsistency have been completely dealt with. We can define the *time inconsistency wedge* in our economy for agents with history h^t as

$$\tau\left(h^t\right) = \sum_{h_{t+1} \in H_{t+1}} \pi_{t+1}\left(h_{t+1} | h^t\right) \frac{u'\left(c_t\left(h^t\right)\right)}{\delta R u'\left(c_{t+1}\left(h^t, h_{t+1}\right)\right)} - 1$$

If agents face self-control problems when left on their own absent commitment devices, this would be represented as a negative time consistency wedge. Our second main result extends Theorem 2 to our dynamic economy.

Theorem 5. *Fix $\{\theta_2, \ldots, \theta_{N-1}\}$ and $\{\beta_2, \ldots, \beta_M\}$. Then there exist scalars $\underline{\theta} > 0$, $\overline{\theta} < +\infty$, and $\underline{\beta} > 0$ such that at the solution to the planner's problem:*

1. *If $\theta_1 \leq \underline{\theta}$, then for any $t = 1, \ldots, T-1$ and history h^{t-1} : $\tau\left(h^{t-1}, (\theta_1, \beta)\right) \geq 0$ for all $\beta \in B$;*

[32] For example, this is the case when $\theta_t = \theta_0$ for all agents in the economy at all histories.

[33] For applications in the context of optimal taxation see Diamond and Mirrlees (1977), Golosov et al. (2003), Golosov and Tsyvinski (2006), Farhi and Werning (2012), Farhi and Werning (2013), Stantcheva (2015), and Golosov et al. (2016).

2. *If $\theta_N > \overline{\theta}$ and $\beta_1 \leq \underline{\beta}$, then for any $t = 1, \ldots, T-1$ and history h^{t-1}:*

   - $\tau\left(h^{t-1}, (\theta_N, \beta_M)\right) \geq 0$;
   - $\tau\left(h^{t-1}, (\theta_N, \beta_1)\right) < 0$.

Proof. See Appendix B.3.4. □

Analogous to the intuition behind Theorem 2, savings distortions optimally vary throughout the income distribution. For low enough productivity types, the planner fully undoes low-ability types' self-control problem, and at times may induce savings above the first-best rate ($\tau\left(h^t\right) = 0$) as a screening device. On the other hand, high-ability types are differentially distorted, with the most patient agents ($\beta_M = 1$) weakly over-saving, but the lowest ability types strictly under-saving relative to the efficient level. Thus, not only does the planner provide imperfect commitment at higher ability levels, but it is also optimal to offer greater choice in savings for this part of the population.

B.3 Proofs of general results in multi-period life-cycle model

B.3.1 Problem reformulation

Types are unobservable and we rely on the Revelation Principle to characterize implementable allocations. To this end, we define an allocation as a pair of functions $(c_t, y_t) : H^1 \times \cdots \times H^{t-1} \times H^t \to \mathbb{R}^2_+$ for each period t that assigns a consumption level and an earnings level for any reported history $h^t \in H^t$ at period t and any past reported history $\hat{r}^{t-1} = \left(h^1, \ldots, h^{t-1}\right) \in H^1 \times \cdots \times H^{t-1}$. A strategy for an agent is a sequence of reporting strategies $\sigma_t : H^1 \times \cdots \times H^{t-1} \times H^t \to H^t$. The overall payoff after history h^t, previous reports $\hat{r}^{t-1} = \left(r^1, \ldots, r^{t-1}\right) \in H^1 \times \cdots \times H^{t-1}$ and following a strategy $(\sigma_s)_{s=t}^T$ from period t on is given by

$$U_t\left(\hat{r}^{t-1}, h^t, (\sigma_s)_{s=t}^T\right) = u\left(c_t\left(\hat{r}^{t-1}, \sigma_t\left(\hat{r}^{t-1}, h^t\right)\right)\right) - \frac{v\left(y_t\left(\hat{r}^{t-1}, \sigma_t\left(\hat{r}^{t-1}, h^t\right)\right)\right)}{\theta_t}$$
$$+ \beta_t \sum_{s=t+1}^{T} \delta^{s-t} \sum_{h^s \succ h^t} \pi_s\left(h^s | \theta_t\right) \left[u\left(c_s\left(\sigma_s\left(\hat{r}^{s-1}, h^s\right)\right)\right) - \frac{v\left(y_s\left(\sigma_s\left(\hat{r}^{s-1}, h^s\right)\right)\right)}{\theta_s}\right]$$

Note that preferences are hyperbolic with quasi-geometric discount factor β_t in period t.[34]

[34] We denote by $h^s \succ h^t$ the continuation histories at times $s > t$ that are consistent with h^t.

We assume that agents are sophisticated in that they take into account their present bias problems in the future. Define the truth-telling strategy as $\sigma_t^{Truth}\left(\hat{r}^{t-1}, h^t\right) = h^t$. An allocation satisfies IC if truth-telling is a sub-game perfect equilibrium of the game played between the selves in different periods, so that after any history of reports $\hat{r}^{t-1} \in H^1 \times \cdots \times H^{t-1}$ and any realized type h^{t-1} truth-telling is the optimal one-shot deviation:

$$\sigma_t^{Truth} \in \arg\max_{\sigma_t'} U_t\left(\hat{r}^{t-1}, h^t, \left(\sigma_t', \left(\sigma_s^{Truth}\right)_{s=t+1}^{T}\right)\right)$$

Taking into account that future selves will consider it optimal to report the truth, reporting the truth in period t after history h^t is optimal given any reports history $\hat{r}^{t-1}$. Since this is a Bayesian game with positive probabilities at all nodes of the game, the Revelation Principle guarantees that the outcome of any mechanism can be obtained using the allocations defined above.

Our assumptions of full support over types, the Markovian nature of the stochastic process over types and the planner's objective allow us to further simplify IC constraints in this environment. The Markovian structure implies that, conditional on $\hat{r}^{t-1}$, the preferences after any history $\tilde{h}^t \in H^t$ with $h_t = \tilde{h}_t$ have the same ordering as the preferences after history h^t. As we will show below, the planner's objective function is strictly concave, which implies that the optimal allocation in period t treats agents of type $\tilde{h}^t$ and h^t identically. Hence we can write

$$\begin{aligned} c_{t+s}\left(\hat{r}^{t-1}, h^t, \ldots, h^s\right) &= c_{t+s}\left(\hat{r}^{t-1}, h_t, \ldots, h^s\right) \\ y_{t+s}\left(\hat{r}^{t-1}, h^t, \ldots, h^s\right) &= y_{t+s}\left(\hat{r}^{t-1}, h_t, \ldots, h^s\right) \end{aligned}$$

Using this argument recursively for all periods $s > t$ we obtain

$$\begin{aligned} c_{t+s}\left(\hat{r}^{t-1}, h^t, \ldots, h^s\right) &= c_{t+s}\left(\hat{r}_1, \ldots, \hat{r}_{t-1}, h_t, \ldots, h_s\right) \\ y_{t+s}\left(\hat{r}^{t-1}, h^t, \ldots, h^s\right) &= y_{t+s}\left(\hat{r}_1, \ldots, \hat{r}_{t-1}, h_t, \ldots, h_s\right) \end{aligned}$$

where we used that $\hat{r}_1, \ldots, \hat{r}_t$ are optimal reports for an agent with that history of types. Therefore it is without loss of generality that the mechanism requires only reporting of the current period type and not of the full history of types.[35]

[35]This characterization implies that only equilibrium path allocations are important for IC (Fernandes and Phelan, 2000; Kapička, 2013). This argument can break down in problems with perfectly correlated types, demonstrated by

From here onward, we denote by $\left(u_t\left(h^t\right), v_t\left(h^t\right)\right)$ the intra-period allocation in utility space.

B.3.2 Precursory results

The following result is the extension of Lemma 1 to the dynamic economy.

Lemma 4. *Given $\{\theta_2, \ldots, \theta_N\}$, there is $\underline{\theta} > 0$ such that if $\theta_1 < \underline{\theta}$ then at the solution to the planner's problem we have*

$$\underbrace{U_t\left(h^{t-1}, (\beta, \theta_1)\right)}_{\text{truthful report}} \geq \underbrace{U_t\left((\beta', \theta_1)\,|h^{t-1}, (\beta, \theta_1)\right)}_{\text{deviation in } \beta} > \underbrace{U_t\left((\beta', \theta')\,|h^{t-1}, (\beta, \theta_1)\right)}_{\text{deviation in } \theta}$$

and

$$v_t\left(h^{t-1}, (\beta', \theta')\right) > v_t\left(h^{t-1}, (\beta, \theta_1)\right)$$

for all $\theta' > \theta_1$, for all $\beta' \neq \beta$, for all h^{t-1}, and for all $t = 1, \ldots, T$.

Proof. If $\theta_1 = 0$, then $y_t\left(h^{t-1}, (\beta, \theta_1)\right) = 0$ for all $\beta \in B$ and all h^{t-1}. For any $\theta' > 0$, then $y_t\left(h^{t-1}, (\beta', \theta')\right) > 0$, therefore

$$U_t\left((\beta', \theta')\,|h^{t-1}, (\beta, \theta_1)\right) = -\infty$$

which proves the second, strict inequality. The first, weak inequality is required by IC. Continuity of the solution to the planner's problem then implies that for fixed $\{\theta_2, \theta_3, \ldots, \theta_N\}$ there is $\underline{\theta} > 0$ such that for all $\theta_1 \leq \underline{\theta}$ the desired sequence of inequalities holds. □

Lemma 5. *Given $\{\theta_1, \ldots, \theta_{N-1}\}$, there exists $\overline{\theta} < +\infty$ such that if $\theta_N > \overline{\theta}$ then at the solution to the planner's problem we have*

$$U_t\left(h^{t-1}, (\beta', \theta')\right) > U_t\left((\beta, \theta_N)\,|h^{t-1}, (\beta', \theta')\right)$$

and

$$v_t\left(h^{t-1}, (\beta, \theta_N)\right) > v_t\left(h^{t-1}, (\beta', \theta')\right)$$

for all h^{t-1} , for all $\theta' \in \{\theta_1, \ldots, \theta_{N-1}\}$, and for all $\beta, \beta' \in B$.

an example in Battaglini and Lamba (2015). The assumption of full support of β_t for all t and histories is crucial for this characterization to be valid. If there is no full support in β_t, then it is possible to design a mechanism in which off-equilibrium path allocations relax incentive constraints on the equilibrium path, as in Galperti (2015) and Yu (2016).

Proof. The proof is analogous to that of Lemma 2, extended to the dynamic setting. □

Lemma 6. *Given* $\{\theta_1, \ldots, \theta_{N-1}\}$ *and* $\{\beta_2, \ldots, \beta_M\}$*, there exists* $\overline{\theta} < +\infty$ *and* $\underline{\beta} > 0$ *such that if* $\theta_N > \overline{\theta}$ *and* $\beta_1 < \underline{\beta}$ *then at the solution to the planner's problem we have* $u_t\left(h^{t-1}, (\beta_1, \theta_N)\right) > u_t\left(h^{t-1}, (\beta, \theta)\right)$ *for all* h^{t-1}*, for all* $\theta \in \{\theta_1, \ldots, \theta_{N-1}\}$*, for all* $\beta \in B$ *and for all* $t = 1, \ldots, T-1$*.*

Proof. From Lemma 5, we know that $v_t\left(h^{t-1}, (\beta_1, \theta_N)\right) > v_t\left(h^{t-1}, (\beta, \theta)\right)$. Note that if $\beta_1 = 0$, IC requires $u_t\left(h^{t-1}, (\beta_1, \theta_N)\right) > u_t\left(h^{t-1}, (\beta, \theta)\right)$. By continuity of the solution to the planner's problem, there exists $\underline{\beta}_t > 0$ such that this inequality continues to be strict for all $\beta_1 < \underline{\beta}_t$. Since $T < \infty$ we can pick a uniform level of $\underline{\beta} = \min_t \left\{\underline{\beta}_t\right\} > 0$ that satisfies the desired property. □

B.3.3 Proof of Theorem 4

Part 1.

Proof. Consider the problem in terms of utility levels from consumption and disutility levels from working as in Appendix A.2.1. Assume by way of contradiction that for a fixed $t < T$ and fixed history $h^{t-1} \in H^{t-1}$ and for $\beta, \beta' \in B_t$ the solution to the planner's problem features

$$u_t\left(h^{t-1}, (\beta, \theta_1)\right) > u_t\left(h^{t-1}, (\beta', \theta_1)\right)$$

Consider a new allocation that is a convex combination between between $\left(h^{t-1}, (\beta^*, \theta_1)\right)$-types' allocations for all $\beta^* \in B$ and that is offered after history h^{t-1}:

$$\tilde{u}_t\left(h^{t-1}, (\beta^*, \theta_1), h_{t+1}, \ldots, h_T\right) = \sum_{b \in B} \frac{\pi_t\left((b, \theta_1) | h^{t-1}\right)}{\sum_{b' \in B} \pi_t\left((b', \theta_1) | h^{t-1}\right)} u_t\left(h^{t-1}, (b, \theta_1), h_{t+1}, \ldots, h_{t+s}\right)$$
$$\tilde{v}_t\left(h^{t-1}, (\beta^*, \theta_1), h_{t+1}, \ldots, h_T\right) = \sum_{b \in B} \frac{\pi_t\left((b, \theta_1) | h^{t-1}\right)}{\sum_{b' \in B} \pi_t\left((b', \theta_1) | h^{t-1}\right)} v_t\left(h^{t-1}, (b, \theta_1), h_{t+1}, \ldots, h_{t+s}\right)$$

By Lemma (4), there exists $\underline{\theta} > 0$ such that for $\theta_1 < \underline{\theta}$ we have

$$U_t\left(h^{t-1}, (\beta, \theta_1)\right) \geq U_t\left((b, \theta_1) | h^{t-1}, (\beta, \theta_1)\right) > U_t\left((\beta'', \theta') | h^{t-1}, (\beta, \theta_1)\right)$$

for all $\theta' > \theta_1$ and for all $\beta'' \in B$. Therefore, IC constraints at nodes $\left(h^{t-1}, (b, \theta_1)\right)$ are satisfied for all $b \in B$. From linearity of the objective function, IC constraints of $\left(h^{t-1}, (b, \theta)\right)$-types are also satisfied for all $\theta > \theta_1$ and all $b \in B$. Therefore this perturbation preserves IC at period

t. Furthermore, from the planner's point of view the perturbation continuation utility at h^{t-1} remains unchanged. Therefore welfare is unchanged, while for agents with hyperbolic preferences all IC constraints for period $s \leq t-1$ are satisfied. For histories $h^s \succ h^{t-1}$ for $s > t$, taking a convex combination leaves incentives unchanged because the objective is linear. But $C(u) = u^{-1}(u)$ is strictly convex, so for $u_t\left(h^{t-1}, (\beta, \theta_1)\right) > u_t\left(h^{t-1}, (\beta', \theta_1)\right)$ and $\pi_t(\cdot)$ having full support this perturbation saves a strictly positive amount of resources—a contradiction. Hence, agents with types $\left\{\left(h^{t-1}, (\theta_1, \beta)\right) : \beta \in B\right\}$ are bunched. □

Part 2.

Proof. Assume by way of contradiction that for some h^{t-1} we have that $\left(h^{t-1}, (\beta_t, \theta_N)\right)$-types for all $\beta_t \in B_t$ share the same allocation. Then

$$\mathbb{E}_t\left[\frac{C'\left(u_{t+1}\left(h^{t-1}, (\beta_t, \theta_N), (\beta_{t+1}, \theta_{t+1})\right)\right)}{\delta R_t C'\left(u_t\left(h^{t-1}, (\beta_t, \theta_N)\right)\right)} | \theta_N\right] = \kappa$$

for some constant $\kappa > 0$. Recalling that $\beta_M = 1$, consider the following perturbation for the allocation of type (β_M, θ_N):

$$\begin{aligned}
\tilde{u}_t\left(h^{t-1}, (\beta_M, \theta_N)\right) &= u_t\left(h^{t-1}, (\beta_M, \theta_N)\right) - \varepsilon \\
\tilde{u}_{t+1}\left(h^{t-1}, (\beta_m, \theta_n), (\beta_{t+1}, \theta_{t+1})\right) &= u_{t+1}\left(h^{t-1}, (\beta_M, \theta_N), (\beta_{t+1}, \theta_{t+1})\right) + \frac{1}{\delta}\varepsilon
\end{aligned}$$

for $\varepsilon > 0$. Welfare of type $\left(h^{t-1}, (\beta_M, \theta_N)\right)$ is kept constant by such a change. Types $\left(h^{t-1}, (\beta_j, \theta_N)\right)$ for $\beta_j < 1$ dislike this perturbation, so it preserves IC. The marginal resource cost is

$$\begin{aligned}
dE &= -C'\left(u_t\left(h^{t-1}, (\beta_M, \theta_N)\right)\right)\varepsilon + \frac{1}{R_t}\mathbb{E}_t\left[C'\left(u_{t+1}\left(h^{t-1}, (\beta_M, \theta_N), (\beta_{t+1}, \theta_{t+1})\right)\right)|\theta_N\right]\frac{1}{\delta}\varepsilon \\
&= \left[\frac{1}{\delta R_t}\mathbb{E}_t\left[C'\left(u_{t+1}\left(h^{t-1}, (\beta_M, \theta_N), (\beta_{t+1}, \theta_{t+1})\right)\right)|\theta_N\right] - C'\left(u_t\left(h^{t-1}, (\beta_M, \theta_N)\right)\right)\right]\varepsilon \\
&= (\kappa - 1)\, C'\left(u_t\left(h^{t-1}, (\beta_M, \theta_N)\right)\right)\varepsilon
\end{aligned}$$

For the original allocation to be optimal, we require $dE \geq 0$ and thus $\kappa \geq 1$.

Suppose further that $\kappa > 1$. Recall that $\beta_t \leq 1$ for all $\beta_t \in B_t$ and consider the following

perturbation to all types $\left(h^{t-1},(\beta_t,\theta_N)\right)$:

$$\tilde{u}_t\left(h^{t-1},(\beta_t,\theta_N)\right)=u_t\left(h^{t-1},(\beta_t,\theta_N)\right)+\varepsilon$$
$$\tilde{u}_{t+1}\left(h^{t-1},(\beta_t,\theta_N),(\beta_{t+1},\theta_{t+1})\right)=u_{t+1}\left(h^{t-1},(\beta_t,\theta_N),(\beta_{t+1},\theta_{t+1})\right)-\frac{1}{\delta}\varepsilon$$

Type $\left(h^{t-1},(\beta_M,\theta_N)\right)$ is indifferent between the original allocation and the new one, while types $\left(h^{t-1},(\beta_t,\theta_N)\right)$ with $\beta_t<1$ strictly prefer the new allocation for $\varepsilon>0$. Since no IC constraint for types $\{\theta_1,\theta_2,\ldots,\theta_{N-1}\}$ are binding with respect to θ_N-types, then the perturbation preserves IC for $\varepsilon>0$ small enough. The associated resource cost is

$$dE=(1-\kappa)\,C'\left(u_t\left(h^{t-1},(\beta_t,\theta_N)\right)\right)\varepsilon$$

But $\kappa>1$, leading to a resource gain—a contradiction. This leaves us with the case when $\kappa=1$.

Suppose that $\kappa=1$ so that for agents bunched at θ_N the inverse Euler equation holds. By Lemma 3 and θ_N-type agents are bunched, we have $u_t\left(h^{t-1},(\beta_t,\theta_N)\right)>u_t\left(h^{t-1},(\beta_t,\theta_j)\right)$ for all $j<N$. Then consider the following perturbation:

$$\tilde{u}_t\left(h^{t-1},(\beta_1,\theta_N)\right)=u_t\left(h^{t-1},(\beta_1,\theta_N)\right)+\varepsilon-\nu$$
$$\tilde{u}_{t+1}\left(h^{t-1},(\beta_1,\theta_N),(\beta_{t+1},\theta_{t+1})\right)=u_{t+1}\left(h^{t-1},(\beta_1,\theta_N),(\beta_{t+1},\theta_{t+1})\right)-\frac{1}{\delta\beta_2}\varepsilon$$

From the point of view of β_1-types the payoff change is

$$dU(\beta_1)=\left(1-\frac{\beta_1}{\beta_2}\right)\varepsilon-\nu$$

Since $\beta_1<\beta_2$, we can choose $\varepsilon>0$ and $\nu>0$ such that $(1-\beta_1/\beta_2)\,\varepsilon=\nu$. All $\left(h^{t-1},(\beta_1,\theta_N)\right)$-types are left indifferent by this perturbation. Furthermore, agents with type $\beta>\beta_1$ dislike this perturbation. Therefore, since other IC constraints with respect to type $\left(h^{t-1},(\beta_1,\theta_N)\right)$ are slack, the perturbation preserves IC. In terms of resources, however, we have

$$\begin{aligned}dE&=C'\left(u_t\left(h^{t-1},(\beta_1,\theta_N)\right)\right)(\varepsilon-\nu)-\frac{1}{R_t}\mathbb{E}_t\left[C'\left(u_{t+1}\left(h^{t-1},(\beta_M,\theta_N),(\beta_{t+1},\theta_{t+1})\right)\right)|\theta_N\right]\frac{1}{\delta\beta_2}\varepsilon\\&=C'\left(u_t\left(h^{t-1},(\beta_1,\theta_N)\right)\right)\left(\frac{\beta_1}{\beta_2}-\frac{\kappa}{\beta_2}\right)\varepsilon\end{aligned}$$

Since $\beta_2 \leq 1$, then for $\varepsilon > 0$ and $\nu > 0$ we get $dE < 0$, so that the planner saves resources. Since $u_t\left(h^{t-1}, (\beta_t, \theta_N)\right) > u_t\left(h^{t-1}, (\beta_t, \theta_j)\right)$ for all $j < N$, there exists $\varepsilon > 0$ small enough such that redistributing these extra resources improves welfare—a contradiction.. □

B.3.4 Proof of Theorem 5

Part 1.

Proof. Fix a period t and a history h^{t-1}. From Theorem 4, we know that there exists $\underline{\theta} > 0$ such that all agents with a history in $\left\{\left(h^{t-1}, (\beta, \theta_1)\right) : \beta \in B\right\}$ for $\theta_1 < \underline{\theta}$ are bunched at the same continuation allocation. In particular, those agents face the same inverse Euler equation distortion

$$\sum_{(\beta_{t+1}, \theta_{t+1}) \in B \times \Theta} \gamma_{t+1}(\beta_{t+1}) \rho_{t+1}(\theta_{t+1}|\theta_1) \left[\frac{C'\left(u_{t+1}\left(h^{t-1}, (\beta, \theta_1), (\beta_{t+1}, \theta_{t+1})\right)\right)}{\delta R_t C'\left(u_t\left(h^{t-1}, (\beta, \theta_1)\right)\right)} | \theta_N\right] = \kappa$$

for all $\beta \in B$ and for some constant $\kappa > 0$. The desired result holds if and only if $\kappa \geq 1$. Assume by way of contradiction that $\kappa < 1$. Then consider the following perturbation:

$$\begin{aligned}
\tilde{u}_t\left(h^{t-1}, (\beta, \theta_1)\right) &= u_t\left(h^{t-1}, (\beta, \theta_1)\right) - \delta\varepsilon \\
\tilde{u}_{t+1}\left(h^{t-1}, (\beta, \theta_1), (\beta_{t+1}, \theta_{t+1})\right) &= u_{t+1}\left(h^{t-1}, (\beta, \theta_1), (\beta_{t+1}, \theta_{t+1})\right) + \varepsilon
\end{aligned}$$

for all $(\beta_{t+1}, \theta_{t+1}) \in B \times \Theta$. This perturbation keeps welfare constant, hence does not affect IC at period $s < t$, when due to quasi-geometric discounting agents and the planner agree about the intertemporal trade-off between periods t and $t+1$. From Lemma 4, there exists $\underline{\theta} > 0$ such that for $\theta_1 < \underline{\theta}$ agents with histories in $\left\{\left(h^{t-1}, (\beta, \theta_1)\right) : \beta \in B\right\}$ have strictly slack IC constraints with respect to any other agent not in this group. For $\varepsilon > 0$, agents with $\beta \leq 1$ find themselves weakly worse off under this perturbation, so the perturbation preserves incentive compatible. The marginal resource cost is

$$dE = (\kappa - 1)\delta C'\left(u_t\left(h^{t-1}, (\beta, \theta_1)\right)\right)\varepsilon$$

For $\varepsilon > 0$ and $\kappa < 1$ we get $dE < 0$, so the perturbation saves resources—a contradiction. □

Part 2.

Proof. For the first part, let

$$\sum_{(\beta_{t+1},\theta_{t+1})\in B\times\Theta}\gamma_{t+1}\left(\beta_{t+1}\right)\rho_{t+1}\left(\theta_{t+1}|\theta_N\right)\left[\frac{C'\left(u_{t+1}\left(h^{t-1},\left(\beta_M,\theta_N\right),\left(\beta_{t+1},\theta_{t+1}\right)\right)\right)}{\delta R_t C'\left(u_t\left(h^{t-1},\left(\beta_M,\theta_N\right)\right)\right)}|\theta_N\right]=\kappa_H$$

for some $\kappa_H>0$. Assume by way of contradiction that $\kappa_H<1$. Then consider the perturbation

$$\begin{aligned}\tilde{u}_t\left(h^{t-1},\left(\beta_M,\theta_N\right)\right)&=u_t\left(h^{t-1},\left(\beta_M,\theta_N\right)\right)-\delta\varepsilon\\ \tilde{u}_{t+1}\left(h^{t-1},\left(\beta_M,\theta_N\right),\left(\beta_{t+1},\theta_{t+1}\right)\right)&=u_{t+1}\left(h^{t-1},\left(\beta_M,\theta_N\right),\left(\beta_{t+1},\theta_{t+1}\right)\right)+\varepsilon\end{aligned}$$

for all $(\beta_{t+1},\theta_{t+1})\in B\times\Theta$. Since $\beta_M=1$, this perturbation preserves IC for $\varepsilon>0$. The marginal resource cost is

$$dE=\left(\kappa_H-1\right)\delta C'\left(u_t\left(h^{t-1},\left(\beta_M,\theta_N\right)\right)\right)\varepsilon$$

so for $\kappa_H<1$ we have that $dE<0$ whenever $\varepsilon>0$, which saves a strict amount of resources—a contradiction.

For the second part, note that from Lemma 6 and from Theorem 4 we have that there exists $\overline{\theta}<+\infty$ and $\underline{\beta}>0$ such that for $\theta_N>\overline{\theta}$ and $\beta_1<\underline{\beta}$ we have that agents with histories in $\left\{\left(h^{t-1},(\beta,\theta)\right):\beta\in B,\theta<\theta_N\right\}$ strictly prefer their own allocation to the allocation of any agent with history $\left\{\left(h^{t-1},(\beta,\theta_N)\right):\beta\in B\right\}$. Furthermore, we have $u_t\left(h^{t-1},(\beta_1,\theta_N)\right)>u_t\left(h^{t-1},(\beta,\theta)\right)$ for all $\theta<\theta_N$ and all $\beta\in B$ by Lemma 6. Suppose now that

$$\sum_{(\beta_{t+1},\theta_{t+1})\in B\times\Theta}\gamma_{t+1}\left(\beta_{t+1}\right)\rho_{t+1}\left(\theta_{t+1}|\theta_N\right)\left[\frac{C'\left(u_{t+1}\left(h^{t-1},\left(\beta_1,\theta_N\right),\left(\beta_{t+1},\theta_{t+1}\right)\right)\right)}{\delta R_t C'\left(u_t\left(h^{t-1},\left(\beta_1,\theta_N\right)\right)\right)}|\theta_N\right]=\tilde{\kappa}_H$$

for some $\tilde{\kappa}_H>0$. Assume by way of contradiction that $\tilde{\kappa}_H\geq 1$. Then consider the perturbation:

$$\begin{aligned}\tilde{u}_t\left(h^{t-1},\left(\beta_1,\theta_N\right)\right)&=u_t\left(h^{t-1},\left(\beta_1,\theta_N\right)\right)+\beta_1\delta\varepsilon+\nu\\ \tilde{u}_{t+1}\left(h^{t-1},\left(\beta_1,\theta_N\right),\left(\beta_{t+1},\theta_{t+1}\right)\right)&=u_{t+1}\left(h^{t-1},\left(\beta_1,\theta_N\right),\left(\beta_{t+1},\theta_{t+1}\right)\right)-\varepsilon\\ \tilde{u}_t\left(h^{t-1},\left(\beta,\theta\right)\right)&=\tilde{u}_t\left(h^{t-1},\left(\beta,\theta\right)\right)+\nu\end{aligned}$$

for all $(\beta_{t+1},\theta_{t+1})\in B\times\Theta$ and all $(\beta,\theta)\neq(\beta_1,\theta_N)$. For $\varepsilon>0$ and $\nu>0$, this perturbation is incentive compatible since agents with $\beta\geq\beta_1$ find it (weakly) less attractive. To keep welfare

unchanged, we require

$$dW = \pi\left(\beta_1, \theta_N | h^{t-1}\right)(\beta_1 - 1)\,\delta\varepsilon + \nu = 0$$

so that $\nu = \pi\left(\beta_1, \theta_N | h^{t-1}\right)(1-\beta_1)\,\delta\varepsilon$. The marginal resource cost is

$$\begin{aligned} dE =& \pi\left(\beta_1, \theta_N | h^{t-1}\right) C'\left(u_t\left(h^{t-1}, (\beta_1, \theta_N)\right)\right)\delta\left(\beta_1 - \tilde{\kappa}_H\right)\varepsilon + \nu \sum_{(\beta,\theta)} \pi\left(\beta, \theta | h^{t-1}\right) C'\left(u_t\left(h^{t-1}, (\beta, \theta)\right)\right) \\ =& \pi\left(\beta_1, \theta_N | h^{t-1}\right)\delta C'\left(u_t\left(h^{t-1}, (\beta_1, \theta_N)\right)\right)(1-\beta_1)\left[\left(\frac{\beta_1 - \tilde{\kappa}_H}{1-\beta_1}\right) + \sum_{(\beta,\theta)} \pi\left(\beta, \theta | h^{t-1}\right)\frac{C'\left(u_t\left(h^{t-1}, (\beta, \theta)\right)\right)}{C'\left(u_t\left(h^{t-1}, (\beta_1, \theta_N)\right)\right)}\right]\varepsilon \end{aligned}$$

Since $u_t\left(h^{t-1}, (\beta_1, \theta_N)\right) > u_t\left(h^{t-1}, (\beta, \theta)\right)$ for $\theta < \theta_N$ by Lemma 6, and clearly $u_t\left(h^{t-1}, (\beta_1, \theta_N)\right) \geq u_t\left(h^{t-1}, (\beta, \theta_N)\right)$ for all $\beta \in B$, then

$$\sum_{(\beta,\theta)} \pi\left(\beta, \theta | h^{t-1}\right)\frac{C'\left(u_t\left(h^{t-1}, (\beta, \theta)\right)\right)}{C'\left(u_t\left(h^{t-1}, (\beta_1, \theta_N)\right)\right)} < 1$$

Since $\beta_1 < 1 \leq \tilde{\kappa}_H$ we conclude that $dE < 0$ for $\varepsilon > 0$, meaning that this perturbation saves a strictly positive amount of resources—a contradiction. □

C Details of numerical solution algorithm

We start by defining a general global non-linear maximization problem

$$\max_{x \in A} f(x)$$

for $A \subset \mathbb{R}^K$ and $f : \mathbb{R}^K \to \mathbb{R}$ where $K \in \mathbb{N}$. We impose the following four regularity conditions, which are satisfied for a large class of problems including the one we study:

Assumption 1. *$A = \cap_{i \in I} D_i$ where D_i are convex sets, for some finite set I.*

Assumption 2. *f is concave and has a unique maximum at any set $\cap_{j \in J} D_j$ for $J \subset I$.*

Assumption 3. *There exists $J^* \subset I$ such that $\#J^* < K$ and*

$$\arg \max_{x \in \cap_{j \in J^*} D_j} f(x) = \arg \max_{x \in A} f(x)$$

Assumption 4. *For all $J \subset I$, there exists $S_J \subset J$ such that $\#S_J < K-1$ and*

$$\arg \max_{x \in \cap_{j \in S_J} D_j} f(x) = \arg \max_{x \in \cap_{j \in J} D_j} f(x)$$

Assumptions 1 is satisfied since all constraints in the re-stated planner's problem (6) are either linear or strictly convex. Assumption 2 holds because we seek the solution to a convex program. Assumptions 3 and 4 are sufficient conditions for the LICQ to hold at the solution to the planner's problem. We check them numerically at each step in our routine. We then propose the following numerical algorithm, which finds the smallest set of binding constraints at the optimum by looping through subsets of the set of global IC constraints and solving a sequence of relaxed problems:

1. Start with a set $J_0 \subset I$ such that $\#J_0 \leq K-1$

   1.1 If $\#J_0 = K-1$, then find $J_0' \subset J_0$ that makes assumption 4 hold

   1.2 Let $J_0 = J_0'$.

2. Solve for $x(J_0) = \arg\max_{x \in \cap_{j \in J_0}} f(x)$

   2.1 Let $J^D(J_0) = \left\{ j \in J_0 : x(J_0 \backslash \{j\}) \in \cap_{i \in J_0} D_i \right\}$ be the set of slack constraints

   2.2 If $\#J^D(J_0) = 0$, let $J_0' = J_0$

   2.3 If $\#J^D(J_0) > 0$:

      i. Randomly select $J_0^D \subset J^D(J_0)$

      ii. Let $J_0' = J_0 \backslash J_0^D$

   2.4 Let $J^V(J_0') = \{ i \in I : x(J_0') \notin D_i \}$

   2.5 Stop if $\#J^V(J_0') = 0$

3. Randomly select a subset $J_0^V \subset J^V(J_0)$ with $K - \#J_0' - 1$ elements

4. Let $J_1 = \left(J_0 \backslash J_0^D\right) \cup J_0^V$, then $\#J_1 = K-1$

This algorithm iteratively finds the smallest set of binding constraints by adding a subset of violated but excluded constraints, and dropping a subset of redundant constraints in each iteration. Adding a stochastic component to the constraint selection criterion avoids cycles and guarantees that the algorithm finds the unique global optimum of the above program in finite time.

Theorem 6. *The algorithm converges with probability one to the global solution of the problem*

$$plim\ x\left(J_n\right) = x^*$$

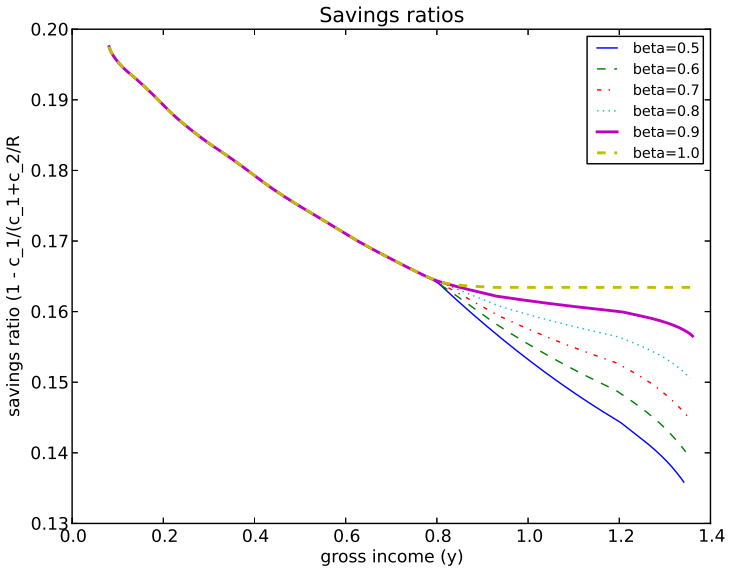
where $x^* = \arg\max_{x \in A} f\left(x\right)$.

Proof. The proof relies on the convexity of the problem as defined in Section A.2.1, and proceeds as follows: First, note that $f\left(x\left(J_n\right)\right)$ is monotonically decreasing, so it converges. Suppose the limit is above $f\left(x^*\right)$. Then since I is finite, from some sufficiently high n onward we have $f\left(J_n\right) = f\left(J_{n-1}\right)$ for all n. But $\#J^V\left(J_n\right) > 0$, so there is a positive probability that $f\left(J_{n+1}\right) < f\left(J_n\right)$. Therefore, there is zero probability of a limit above $f\left(x^*\right)$. Thus plim $f\left(J_n\right) = f\left(x^*\right)$. Since x^* is unique and f is continuous, we have plim$x\left(J_n\right) = x^*$. □

Large-scale implementation. We implement our algorithm using the Interior Point Optimizer (IPOPT) large-scale nonlinear optimization library in Python. To ensure stability of our solution, we try different algorithms and starting points, and ensure that solutions coincide. We illustrate the capabilities of our algorithm by solving a problem with $\left(\left|\Theta\right|, \left|B\right|\right) = \left(1000, 6\right)$, that is $6,000$ types, $18,000$ choice variables, and $35,994,001$ constraints. Our algorithm solves this problem in approximately two minutes on a 2013 MacBook Pro. We find that well below one percent of constraints bind at the optimum, and that most binding constraints—though not all—are "local."

Figure 7 plots optimal savings ratios, defined as $s = \left(c_2/R\right) / \left(c_1 + c_2/R\right)$, as a function of individuals' gross income (x-axis) and present bias level (colored lines). The numerical solution extends in interesting ways our main theoretical results, which stated that the lowest ability types were bunched and the highest types were separated (Theorem 1), and that optimal savings were above the first-best rate at the bottom but below first-best at the top (Theorem 2). In the intermediate range, we observe bunching at higher than first-best savings rates up to some threshold, and separation into savings rates ordered by β above this threshold. Notably, high-ability agents that agree with the planner $\left(\beta = 1\right)$ converge to the first-best savings rate, while those who disagree $\left(\beta < 1\right)$ save above their preferred rate but below the first-best.

Figure 7. Optimal savings rates in a numerical illustration of the solution algorithm



Note: Illustration of numerical algorithm solving a problem with $(|\Theta|, |B|) = (1000, 6)$ types, $18,000$ choice variables, and $35,994,001$ constraints. Savings rate is defined as $s = (c_2/R) / (c_1 + c_2/R)$.

D Proof of Theorem 3 for the general problem

Part 1.

Proof. The proof is analogous to that of Theorem 1 and relies only on convexity of the set of IC constraints when we rewrite the problem in terms of utility levels and preference wedges. □

Part 2.

Proof. Note that since $\theta_1 < \underline{\theta} < \overline{\theta} < \theta_N$ we have $v(\theta_N, \alpha) > v(\theta_1)$. Therefore, IC implies

$$u(\theta_N, \alpha) + \alpha\varepsilon(\theta_N, \alpha) > u(\theta_1) + \alpha\varepsilon(\theta_1)$$

In particular, for $\alpha = 0$ we have $u(\theta_N, \alpha) > u(\theta, \alpha')$ for all (θ, α'). Assume by way of contradiction that $u(\theta_N, \alpha) = u(\theta_N)$, $v(\theta_N, \alpha) = v(\theta_N)$ and $a(\theta_N, \alpha) = a(\theta_N)$ at the solution to the planner's problem. Then we have $u(\theta_N) > u(\theta_1)$. Let $\alpha_{max} = \max\{A\} \geq 0$. If $b'(a(\theta_N)) > p$, then consider

the following change to the allocation:

$$\begin{aligned}\tilde{a}\left(\theta_N, \alpha_{max}\right) &= \alpha\left(\theta_N\right) + \nu \\ \tilde{u}\left(\theta_N, \alpha_{max}\right) &= u\left(\theta_N\right) - \alpha_{max}\nu + \eta \\ \tilde{u}\left(\theta, \alpha\right) &= \tilde{u}\left(\theta, \alpha\right) + \eta\end{aligned}$$

for all $(\theta, \alpha) \neq (\theta_N, \alpha_{max})$. For $\nu > 0$ this perturbation is incentive compatible since for $\alpha < \alpha_{max}$ the change in the deviation payoff into the allocation of (θ_N, α_{max}) is $(\alpha - \alpha_{max})\nu < 0$ and the original allocation is incentive compatible. The marginal change in resources used is

$$\begin{aligned}dE =& \pi\left(\theta_N, \alpha_{max}\right)\left[-\alpha_{max}\nu C'\left(u\left(\theta_N\right)\right) - \left(\frac{b'\left(\varepsilon^{-1}\left(\varepsilon\left(\theta_N\right)\right)\right)}{\varepsilon'\left(\varepsilon^{-1}\left(\varepsilon\left(\theta_N\right)\right)\right)} - \frac{p}{\varepsilon'\left(\varepsilon^{-1}\left(\varepsilon\left(\theta_N\right)\right)\right)}\right)\nu\right] \\ &+ \eta\sum_{\theta,\alpha}\pi\left(\theta, \alpha\right)C'\left(u\left(\theta, \alpha\right)\right)\end{aligned}$$

If $\alpha_{max} = 0$, then this yields a contradiction since for $\nu > 0$ we can set $\eta > 0$ small enough to increase welfare. For $\alpha_{max} > 0$, the marginal change in the government's objective is given by

$$dW = -\pi\left(\theta_N, \alpha_{max}\right)\lambda\left(\theta_N\right)\alpha_{max}\nu + \eta$$

If $\eta = \pi\left(\theta_N, \alpha_{max}\right)\lambda\left(\theta_N\right)\alpha_{max}\nu$, then $dW = 0$ and the marginal change in resource cost is

$$\begin{aligned}\frac{dE}{\pi\left(\theta_N, \alpha_{max}\right)\alpha_{max}C'\left(u\left(\theta_N\right)\right)} =& \nu\left\{\lambda\left(\theta_N\right)\sum_{\theta,\alpha}\pi\left(\theta, \alpha\right)\frac{C'\left(u\left(\theta, \alpha\right)\right)}{C'\left(u\left(\theta_N\right)\right)} - 1\right\} \\ &- \frac{\nu}{\alpha_{max}C'\left(u\left(\theta_N\right)\right)}\left(\frac{b'\left(\varepsilon^{-1}\left(\varepsilon\left(\theta_N\right)\right)\right)}{\varepsilon'\left(\varepsilon^{-1}\left(\varepsilon\left(\theta_N\right)\right)\right)} - \frac{p}{\varepsilon'\left(\varepsilon^{-1}\left(\varepsilon\left(\theta_N\right)\right)\right)}\right)\end{aligned}$$

Since $\lambda\left(\theta_N\right) \leq 1$ and since $C\left(\cdot\right)$ is convex and $u\left(\theta_N\right) \geq u\left(\theta, \alpha\right)$, then $dE < 0$ for $\nu > 0$, a contradiction. Therefore we must have $b'\left(a\left(\theta_N\right)\right) \leq p$. In this case, we can construct an analogous perturbation for $\alpha_{min} = \min\left\{A\right\} \leq 0$. Hence, we conclude that $b'\left(a\left(\theta_N\right)\right) = p$. If $\alpha_{max} > 0$, note that the perturbation above also implies that $dE < 0$ since $u\left(\theta_N\right) > u\left(\theta, \alpha\right)$ for $\theta < \theta_N$, a contradiction. The case for $\alpha_{min} < 0$ is analogous. This exhausts all possible cases, thus contradicting the initial assumption that all agents in $\left\{\left(\theta_N, \alpha\right) : \alpha \in A\right\}$ receive the same allocation. We conclude that θ_N-types are separated in the α-dimension. □