Optimal Dynamic Taxes

Mikhail Golosov
Princeton

Maxim Troshkin
Yale and Minnesota

Aleh Tsyvinski*
Yale

November 2011

Abstract

We study optimal labor and savings distortions in a lifecycle model with idiosyncratic shocks. We show a tight connection between its recursive formulation and a static Mirrlees model with two goods, which allows us to derive elasticity-based expressions for the dynamic optimal distortions. We derive a generalization of a savings distortion for non-separable preferences and show that, under certain conditions, the labor wedge tends to zero for sufficiently high skills. We estimate skill distributions using individual data on the U.S. taxes and labor incomes. Computed optimal distortions decrease for sufficiently high incomes and increase with age.

*Email: golosov@princeton.edu, maxim.troshkin@yale.edu, a.tsyvinski@yale.edu. We thank Stefania Albanesi, Fernando Alvarez, V.V. Chari, Dirk Krueger, Larry Jones, Igor Livshits, Stephen Morris, James Poterba, Emmanuel Saez, Ali Shourideh, Nancy Qian, Hongda Xiao, Pierre Yared, and audiences at ASU, Bank of Japan, Boston University, Chicago, Chicago Booth, Chicago Fed, Cornell, Cowles, EIEF, Gerzensee, Minnesota Macro, NBER PF, Northwestern, NY Fed, NYU, Princeton, Rochester, SED, Texas, UCSD. Marianne Bruins, James Duffy and Nicolas Werquin provided outstanding research assistance. We gratefully acknowledge the use of software licences provided at the Institute on Computational Economics. Golosov and Tsyvinski thank EIEF for hospitality and NSF for support. Troshkin thanks Minneapolis Fed for hospitality and support. Tsyvinski thanks IMES of the Bank of Japan and John Simon Guggenheim Foundation.

A sizeable New Dynamic Public Finance (NDPF) literature studies optimal taxation in dynamic settings[1]. The models in this literature extend the classic Mirrlees equity-efficiency trade-offs to dynamic settings in which agents' skills change stochastically over time. A key theoretical insight of our paper is to show a connection of the dynamic model with a static optimal taxation model with two goods. A recursive formulation of the optimal problem allows us to think of the dynamic problem as one in which an agent each period derives utility from two goods, consumption today and a suitably defined future promises, as well as labor. This allows us to derive formulas that facilitate interpretation of the forces behind the optimal income taxation results in dynamic settings and to generalize the analysis of the savings distortion to the non-separable preferences as well as to show the conditions under which the labor wedges for the high skilled agents tend to zero.

In the static model Diamond (1998) derived the expressions for the optimal labor distortions and showed that they are determined by three key parameters: the shape of the income distribution, the redistributionary objectives of the government, and labor elasticity. The dynamic model introduces three significant differences: (i) the use of dynamic incentives adds a force lowering labor wedges; (ii) conditional rather than unconditional distributions of skills are key determinants of wedges; (iii) persistence of shocks acts as a more redistributionary motive for the planner. We then show that, under certain conditions, the labor wedge tends to zero for sufficiently high skills. Importantly, this result relies on understanding the forces and their interactions behind the savings distortion and the labor distortions and highlights the usefulness of the formulas that we derive. Our results on the wedge tending to

[1]See, for example, Golosov, Kocherlakota, and Tsyvinski (2003) or reviews in Golosov, Tsyvinski, and Werning (2006) and Kocherlakota (2010).

zero for the high skills is in sharp contrast to the static case with the Pareto tail of the skills distribution of Diamond (1998) and Saez (2001), who show that the taxes on the high skill agents are increasing and tend to high levels (50-70%) depending on the chosen elasticity of labor supply.

The theoretical analysis points to empirical skill distributions as a crucial input for a quantitative analysis. We construct a dataset of individual skills and their evolution over lifetime implied by the observed micro level data for the U.S. The main difficulty in estimating skills from the data is that skills are unobservable. One can use wages as a proxy for skills but it does not necessarily correspond to skills which measure the return to effort. We use the data on the actual U.S. tax code and labor income choices to infer the unobservable skill level. Since the details of the actual U.S. system of taxes and transfers are observable, we compute the implied individual skills from the necessary conditions for the individual optimum. The methodology is a dynamic extension of that of Saez (2001) who used a similar approach to infer cross-sectional distribution of skills in the population.

We then numerically simulate the optimal labor and savings wedges in a realistically calibrated economy based on the empirical income distributions. The dynamic wedges are significantly different from the static taxes, emphasizing the importance of the theoretical forces we study. We find that the labor distortion for the early periods are smaller than for the later periods. Importantly, the labor wedges for the high skilled agents tend to zero in our dynamic model while in calibrated static models of Diamond (1998) and Saez (2001) they typically reach 50-70%. We provide simulations of the savings wedge and find it numerically significant and increasing with the labor income. The consideration of conditional rather than the unconditional empirical distributions of income and skills significantly alters the pattern of wedges compared

to the static (or the i.i.d.) case. Agents face very different labor distortions conditional on the previous shocks. This is due to the differences among the conditional distributions and also due to the increase of the planner's redistributionary objectives to deter earlier deviations.

We then compute the welfare gains of using the optimal policy. First, we follow an important insight of Farhi and Werning (2010) to compare the constrained efficient optimum to that with the optimal linear taxes, and confirm their findings in our setup. The optimal age-dependent linear labor wedges yield a welfare loss of 0.9% of consumption compared to the constrained optimum. The optimal age-*in*dependent labor distortion yields a welfare loss of 1.6%. While these magnitudes are non-trivial, linear taxes can still yield reasonably good policies as in Farhi and Werning (2010). Then, we consider a case of a more redistributive social planner. The analysis of the static Mirrlees problems (e.g., Mirrlees (1971), Atkinson and Stiglitz (1976), Tuomala (1990)) also points out that if the planner is more redistributive than utilitarian planner, the tax policy is substantially different from linear, and nonlinear taxes may yield large welfare gains. We calculate welfare gains of using optimal policies when the social planner is more redistributive, in particular Rawlsian. The optimal age-dependent linear labor wedges yield a welfare loss of 4.6% compared to the constrained optimum. The optimal age-*in*dependent labor distortion yields a welfare loss of 5.1%. We conclude that the welfare gains of using optimal nonlinear policies are significant.

There are several papers related to our work. The first-order approach for persistent shocks is developed in Kapička (2010) and Pavan, Segal, and Toikka (2010). In our numerical simulations we verify its sufficiency.

An important contribution of Farhi and Werning (2010) derives a formula describing a dynamic behavior of the labor income tax rate in both continu-

ous and discrete case, provides a simulation of a lifecycle economy, and derives additional insights using a continuous time approach. Our work focuses on a study of cross-sectional properties of optimal wedges, on deriving elasticity based formulas, and on numerical simulations based on calibrated skill distribution that we estimate from the U.S. data extending the analysis of Diamond (1998) and Saez (2001) to dynamic settings.

Numerical simulations in our paper are also related to Weinzierl (2011). He derives theoretically and analyzes numerically an elasticity-based formula with which he studies optimal age-dependent taxation, in a dynamic Mirrlees setting. Albanesi and Sleet (2006) is a comprehensive numerical and theoretical study of optimal capital and labor taxes in a dynamic economy with i.i.d. shocks. Kocherlakota and Pistaferri (2007) and Kocherlakota and Pistaferri (2009) use micro level data to evaluate predictions of dynamic optimal policy models. Golosov and Tsyvinski (2006) study a disability insurance model with fully persistent shocks. Golosov, Tsyvinski, and Werning (2006) is a two-period numerical study of the determinants of dynamic optimal taxation in the spirit of Tuomala (1990). Ales and Maziero (2007) numerically solve a version of a life cycle economy with i.i.d. shocks drawn from a discrete, two-type distribution, and find that the labor distortions are lower earlier in life. Fukushima (2010) simulates a policy reform which replaces an optimal flat tax with an optimal non-linear tax that is age and history dependent and finds sizeable welfare gains in a model where the welfare function places zero Pareto weight on any finite number of cohorts. Battaglini and Coate (2008) provide a complete characterization of the optimal program with Markovian agents. While incorporating persistence in abilities, most of their analysis for tractability assumes only two ability types and risk neutral individuals.

1 Environment

We consider an economy that lasts T periods, denoted by $t = 1, ..., T$ ($T < \infty$)[2]. Each agent's preferences are described by a time separable utility function over consumption good $c_t \geq 0$ and labor $l_t \geq 0$,

$$\mathbb{E}_1 \sum_{t=1}^{T} \beta^{t-1} U(c_t, l_t), \tag{1}$$

where $\beta \in (0, 1)$ is a discount factor, $\mathbb{E}_1$ is a period 1 expectations operator, and $U : \mathbb{R}^2_+ \to \mathbb{R}$.

In period $t = 1$, agents draw their initial type (skill), θ_1, from a distribution $F_1(\theta)$. For $t \geq 2$, skills follow a Markov process $F_t(\theta|\theta_{t-1})$, where θ_{t-1} is agent's skill realization in period $t-1$. We denote the probability density function by $f_t(\theta|\theta_{t-1})$ and assume that f_t is differentiable in both arguments. We assume that, in each period t, skills are non-negative: $\theta_t \in \Theta = \mathbb{R}_+$. The set of possible histories up to period t is denoted by Θ^t.

An agent of type θ_t who supplies l_t units of labor produces $y_t = \theta_t l_t$ units of output. The skill shocks and the history of shocks are privately observed by the agent. Output $y_t = \theta_t l_t$ and consumption c_t are observed by the planner. In period t, the agent knows his skill realization only for the first t periods $\theta^t = (\theta_1, ..., \theta_t)$. Denote by $c_t\left(\theta^t\right) : \Theta^t \to \mathbb{R}_+$ agent's allocation of consumption and by $y_t\left(\theta^t\right) : \Theta^t \to \mathbb{R}_+$ agent's allocation of output in period t. Denote by $\sigma_t\left(\theta^t\right) : \Theta^t \to \Theta^t$ agent's report in period t. We denote the set of all such reporting strategies in period t, $\left(\sigma_1\left(\theta^1\right), ..., \sigma_t\left(\theta^t\right)\right)$ by Σ^t. Resources can be transferred between periods with a rate on savings $\delta > 0$. The observability of consumption implies that all savings are publicly observable. Hence, without

[2]The recursive formulation of the problem that follows makes it easy to extend the analysis to the case of infinitely lived agents. In fact, the calibration and numerical analysis is greatly simplified in the case of infinitely lived agents.

loss of generality, we can assume that the social planner controls all the savings. We also assume that the social planner has a social welfare function defined over lifetime utilities of the agents, $G : \mathbb{R} \to \mathbb{R}$, where G is increasing and concave. Since the lifetime utility of the agent is given by (1), the social welfare is given by $\int G\left(\mathbb{E}_1 \sum_{t=1}^{T} \beta^{t-1} U\left(c_t, l_t\right)\right) dF_1(\theta)$.

We denote partial derivatives of U with respect to c and l as U_c and U_l and define all the second derivatives and cross-partials accordingly. Since $U(c,l) = U(c, y/\theta)$, we also use notation $U_y = U_l \frac{1}{\theta}$ and $U_\theta = U_y\left(-\frac{y}{\theta}\right)$ to denote derivatives with respect to y and θ. We make the following assumptions on U.

Assumption 1. *U is twice continuously differentiable in both arguments, satisfies $U_c > 0, U_l < 0, U_{cc} < 0, U_{ll} < 0, U_{cl} \geq 0$, and*

$$\frac{\partial}{\partial \theta} \frac{U_y\left(c, y; \theta\right)}{U_c\left(c, y; \theta\right)} \geq 0.$$

These assumptions are standard. The last restriction is the single crossing property. The assumption that $U_{cl} \geq 0$ ensures that consumption and leisure are substitutes, which is generally considered to be the empirically relevant case.

In parts of our analysis we will need to use the notion of elasticity of labor supply, holding current period consumption fixed, which is defined as[3]

$$\frac{1}{\varepsilon} \equiv \frac{U_{yy} U_c - U_{cy} U_y}{\left(U_c\right)^2} y \frac{U_c}{U_y}. \tag{2}$$

By definition, when U is separable in c and y, ε is a Frisch elasticity of labor supply. When U has no income effects then ε is the uncompensated elasticity of labor supply. We make the following assumption on ε.

Assumption 2. *The elasticity ε is positive and bounded away from 0.*

[3] It is easy to see that this is elasticity of labor supply by differentiating the intratemporal first order conditions for the household.

For most of the analysis we assume that F_t is differentiable with the p.d.f. f_t, for all t. We denote the partial derivative of f_t with respect to the $t-1$ period shock, θ_-, by $f_{2,t}$. For some parts of the analysis we will need further restrictions on f, in particular, the two assumptions that follow.

Assumption 3. *For all t and for all θ_-, $(1-F_t(\theta|\theta_-))/\theta f_t(\theta|\theta_-)$ is bounded from above.*

Assumption 4. *For all t, $F_t(\cdot|\theta_-)$ first order stochastically dominates $F_t\left(\cdot|\hat{\theta}_-\right)$ if $\theta > \hat{\theta}$ and $f_{2,t}(\theta|\theta_-)/f_t(\theta|\theta_-)$ is increasing in θ .*

The first assumption is very weak, and most empirically relevant distributions satisfy it. The second assumption ensures that high θ types get higher expected utility and introduces a certain form of persistence. The second part of the assumption is satisfied if, for example, the distribution functions have a property that if $\theta_H > \theta_L$ then $f(\theta|\theta_L)/f(\theta|\theta_H)$ is decreasing in θ. This assumption can be further relaxed for many results, but in its current form it significantly simplifies the exposition.

The optimal allocations solve the dynamic mechanism design problem (see, e.g., Golosov, Kocherlakota, and Tsyvinski (2003)):

$$\max_{\left\{c_t\left(\theta^t\right),y_t\left(\theta^t\right)\right\}_{\theta^t\in\Theta^t;t=1,..,T}} \int G\left(\mathbb{E}_1\sum_{t=1}^{T}\beta^{t-1}U\left(c_t\left(\theta^t\right),y_t\left(\theta^t\right)/\theta_t\right)\right)dF_1(\theta) \quad (3)$$

subject to the incentive compatibility constraint:

$$\mathbb{E}_0\left\{\sum_{t=1}^{T}\beta^{t-1}U\left(c_t\left(\theta^t\right),y_t\left(\theta^t\right)/\theta_t\right)\right\}$$
$$\geq \mathbb{E}_0\left\{\sum_{t=1}^{T}\beta^{t-1}U\left(c_t\left(\sigma_t\left(\theta^t\right)\right),y_t\left(\sigma_t\left(\theta^t\right)\right)/\theta_t\right)\right\}, \forall\sigma^T\in\Sigma^T, \quad (4)$$

and the feasibility constraint:

$$\mathbb{E}_0\left\{\sum_{t=1}^{T}\delta^{t-1}c_t\left(\theta^t\right)\right\} \leq \mathbb{E}_0\left\{\sum_{t=1}^{T}\delta^{t-1}y_t\left(\theta^t\right)\right\}. \quad (5)$$

The expectation $\mathbb{E}_0$ above is taken over all possible realizations of histories. Note that the expectation in the objective function is taken after the first period shocks are realized.

We follow Fernandes and Phelan (2000) and Kapička (2010) to briefly describe the recursive formulation and refer to these two papers for the technical details. Let $\omega\left(\tilde{\theta}|\theta\right) : \Theta \times \Theta \to \mathbb{R}$ denote *promised utility* to an agent of skill θ who reports skill $\tilde{\theta}$. We use notation $\boldsymbol{\omega}(\theta)$ and $\boldsymbol{\omega}$ to denote functions $\omega(\theta|\cdot)$ and $\omega(\cdot|\cdot)$, respectively. Let $c : \Theta \to \mathbb{R}_+$ and $y : \Theta \to \mathbb{R}_+$.

The optimal allocations solve the cost minimization problem for period $t = 1$:

$$V_1(\omega_0) = \min_{c,y,\boldsymbol{\omega}} \int \left(c(\theta) - y(\theta) + \delta V_2\left(\boldsymbol{\omega}(\theta), \theta\right)\right) f_1(\theta)\, d\theta$$

subject to the incentive compatibility constraint:

$$U\left(c(\theta), y(\theta)/\theta\right) + \beta\omega(\theta|\theta) \geq U\left(c\left(\tilde{\theta}\right), y\left(\tilde{\theta}\right)/\theta\right) + \beta\omega\left(\tilde{\theta}|\theta\right), \ \forall\tilde{\theta} \in \Theta, \theta \in \Theta, \quad (6)$$

and to the promise keeping constraint:

$$\omega_0 \leq \int G\left(U\left(c(\theta), y(\theta)/\theta\right) + \beta\omega(\theta|\theta)\right) f_1(\theta)\, d\theta.$$

The initial promised utility ω_0 is a solution to $V_1(\omega_0) = 0$.

For $t > 1$, the social planner takes the period $t-1$ realization of the shock and the chosen promised utility function $\hat{\boldsymbol{\omega}}(\theta_-)$ as given and solves:

$$V_t\left(\hat{\boldsymbol{\omega}}(\theta_-), \theta_-\right) = \min_{c,y,\boldsymbol{\omega}} \int \left(c(\theta) - y(\theta) + \delta V_{t+1}\left(\boldsymbol{\omega}(\theta), \theta\right)\right) f_t(\theta|\theta_-)\, d\theta \quad (7)$$

subject to the incentive compatibility constraint (6) and

$$\hat{\omega}(\theta_-|\tilde{\theta}) = \int \left(U\left(c(\theta), y(\theta)/\theta\right) + \beta\omega(\theta|\theta)\right) f_t\left(\theta|\tilde{\theta}\right) d\theta \text{ for all } \tilde{\theta} \in \Theta. \quad (8)$$

The function $V_{T+1}(\boldsymbol{\omega}(\theta), \theta) = 0$, if $\boldsymbol{\omega}(\theta) = \mathbf{0}$, and $V_{T+1}(\boldsymbol{\omega}(\theta), \theta) = \infty$, otherwise. All other functions V_t are defined by backward induction. The function V_t is the resource cost of delivering promised utilities $\boldsymbol{\omega}(\theta)$.

The incentive compatibility constraint states that an agent prefers to reveal his true type θ, receive utility $U(c(\theta), y(\theta)/\theta)$ and a continuation utility $\omega(\theta|\theta)$ rather than claim a different type $\tilde{\theta}$, receive utility $U\left(c\left(\tilde{\theta}\right), y\left(\tilde{\theta}\right)/\theta\right)$ and continuation utility $\omega\left(\tilde{\theta}|\theta\right)$. The promise keeping constraints (8) ensure that next period allocations indeed deliver the expected utility $\omega\left(\tilde{\theta}|\theta\right)$ to any type $\tilde{\theta}$ who sends a report θ.

We proceed in this section by using the first order approach developed by Kapička (2010) and Pavan, Segal, and Toikka (2010) to obtain a more manageable recursive formulation. One needs to keep track only of the "on the path" promised utility $\omega(\theta|\theta)$ and the utility from a local deviation $\omega_2(\theta|\theta)$, where $\omega_2(\theta|\theta)$ is the derivative of ω with respect to its second argument evaluated at $(\theta|\theta)$. Then defining functions $w : \Theta \to \mathbb{R}$ and $w_2 : \Theta \to \mathbb{R}$, the maximization problem (7) can be re-written as

$$V_t(\hat{w}, \hat{w}_2, \theta_-) = \min_{c,y,u,w,w_2} \int \left(c(\theta) - y(\theta) + \delta V_{t+1}(w(\theta), w_2(\theta), \theta)\right) f_t(\theta|\theta_-)\, d\theta \tag{9}$$

$$u'(\theta) = U_\theta(c(\theta), y(\theta)/\theta) + \beta w_2(\theta), \tag{10}$$

$$\hat{w} = \int u(\theta)\, f_t(\theta|\theta_-)\, d\theta, \tag{11}$$

$$\hat{w}_2 = \int u(\theta)\, f_{2,t}(\theta|\theta_-) d\theta, \tag{12}$$

$$u(\theta) = U(c(\theta), y(\theta)/\theta) + \beta w(\theta). \tag{13}$$

There are three state variables in this recursive formulation: $\hat{w}$ is the promised utility associated with the promise-keeping constraint (11); $\hat{w}_2$ is the state

variable associated with the threat-keeping constraint (12); θ_- is the reported type in period $t-1$. In what follows we assume that solution to (9) is differentiable.[4]

The first-order approach is valid only if at the optimum the local constraints (10) are sufficient to guarantee that global incentive constraints (6) are satisfied. It is well known, that there are no general conditions either in the static mechanism design problem with multiple goods (see, e.g., Mirrlees (1976)) or in dynamic models (see, e.g., Kapička (2010)) that guarantee that only local incentive constraints bind. In the next lemma we show sufficient conditions that the optimal allocations must satisfy to guarantee that local constraints (10) imply (6).

Assumption 5. *The optimal allocation satisfies*

$$c'(\theta) \geq 0,\ \omega_1\left(\theta|\hat{\theta}\right) \geq 0,\ \omega_{12}\left(\theta|\hat{\theta}\right) \geq 0,\ \textit{for all } \theta, \hat{\theta}. \tag{14}$$

Lemma 1. *Suppose that Assumption 1 and Assumption 5 are satisfied. Then (10) implies (6)*

In the numerical part of the paper we verify that Assumption 5 is satisfied for the calibrated model.

As stated, problem (9) does not need to be convex. However, if it is not, welfare can be improved by allowing randomizations over w and w_2. Then the differentiability of V_t can be established following the methods similar to, e.g., Acemoglu, Golosov, and Tsyvinski (2008). To avoid cumbersome notation, for the rest of the paper we make the assumption that follows.

[4]It is well known that there are circumstances when solutions to this problems are not differentiable, for example, when it is optimal to bunch different types. There are standard methods to characterize this problem in such situations at an expense of introducing additional notational complexity.

Assumption 6. *V_t is convex and differentiable in $\hat{w}, \hat{w}_2$.*

Before characterizing the problem, we re-write the optimal problem to highlight the effects of persistence. To minimize on notation, we drop explicit conditioning of functions on θ, and use, for example, notation c instead of $c(\theta)$ whenever this does not cause confusion.

Lemma 2. *Let $(c^*, y^*, u^*, w^*, w_2^*)$ be a solution to (9) for $t > 1$. Then $(c^*, y^*, u^*, w^*, w_2^*)$ is a solution to*

$$\min_{(c,y,u,w,w_2)} \int \left(c - y + \delta V_{t+1}\left(w, w_2, \theta\right)\right) f_t\left(\theta|\theta_-\right) d\theta \tag{15}$$

subject to (10), (13), and

$$\hat{w} = \int \left(1 - \zeta \frac{f_2(\theta|\theta_-)}{f(\theta|\theta_-)}\right) u f\left(\theta|\theta_-\right) d\theta, \tag{16}$$

for some constant ζ.

In the constraint (16) utility $u\left(\theta\right)$ is multiplied by the term $\left(1 - \zeta \frac{f_2(\theta|\theta_-)}{f(\theta|\theta_-)}\right)$. This pseudo-objective is equivalent to the objective function of a social planner that has (non-normalized) weights $\left(1 - \zeta \frac{f_2(\theta|\theta_-)}{f(\theta|\theta_-)}\right)$ instead of the utilitarian weights equal to 1 for all types θ in period t. As we will argue later in our analysis, the relevant case in most circumstances is $\zeta > 0$. The term $\left(1 - \zeta \frac{f_2(\theta|\theta_-)}{f(\theta|\theta_-)}\right)$ assigns the highest weight to the lowest type and monotonically decreases for the higher types. In other words, the planner's objective is more redistributionary towards the lower types in period t. The intuition for this change in weights is as follows. Consider a marginal deviation in period $t-1$. Suppose type $\theta_- + \epsilon$ claims to be type θ_- for some small ϵ. Under the above assumption on $f\left(\theta|\theta_-\right)$, this type is relatively more likely to receive high shocks θ and relatively less likely to receive low shocks θ in period t. The social planner who is more redistributive in period t and puts higher (pseudo) weights on the low

types allocates relatively low utility to this agent. The type θ_- is not significantly affected, since his probability of having high shocks θ is relatively low. This agent benefits from more redistribution as for him the high shocks θ in period t are less likely. The same intuition generalizes for other stochastic processes. The main insight is that the social planner allocates relatively higher pseudo weights on those realizations of shocks θ for which there is a relatively large difference in the probability of occurrence between types θ_- and types close to θ_-.

Now we define and proceed to characterize optimal distortions. For an agent with the history of shocks θ^t at time t, we define a labor distortion:

$$1-\tau_t^y\left(\theta^t\right)\equiv\frac{-U_l\left(c_t\left(\theta^t\right),y_t\left(\theta^t\right)/\theta_t\right)}{\theta_t U_c\left(c_t\left(\theta^t\right),y_t\left(\theta^t\right)/\theta_t\right)} \tag{17}$$

and a savings distortion

$$1-\tau_t^s\left(\theta^t\right)=\left(\frac{\delta}{\beta}\right)\frac{U_c\left(c_t\left(\theta^t\right),y_t\left(\theta^t\right)/\theta_t\right)}{\mathbb{E}_t\left\{U_c\left(c_{t+1}\left(\theta^{t+1}\right),y_{t+1}\left(\theta^{t+1}\right)/\theta_{t+1}\right)\right\}}. \tag{18}$$

For some results it will also be useful to define a life-time savings distortion, $\bar{\tau}_t^s$, as

$$1-\bar{\tau}_t^s\left(\theta^t\right)=\left(\frac{\delta}{\beta}\right)^{T-t}\frac{U_c\left(c_t\left(\theta^t\right),y_t\left(\theta^t\right)/\theta_t\right)}{\mathbb{E}_t\left\{U_c\left(c_T\left(\theta^T\right),y_T\left(\theta^T\right)/\theta_T\right)\right\}}. \tag{19}$$

2 Characterization of distortions

In the appendix, we proceed by setting up Hamiltonian to (15) and characterizing the solution under the assumptions made in the previous section. In this section, we instead provide a heuristic analysis of the problem in which we emphasize a close connection of the dynamic mechanism design problem (9) and static tax problem with two goods, as in Mirrlees (1976) and Mirrlees (1986). We show how the insights from the static model can be used to provide characterization of the distortions in the dynamic model.

We now re-write the problem in a more intuitive form. Define function $H_t(r, \hat{w}_2, \theta_-)$ implicitly from $V_t(H_t, \hat{w}_t, \theta_-) = r$ and denote by $H_{r,t} = \left(\frac{\partial V_t}{\partial w}\right)^{-1}$ the partial derivative of H with respect to r. $H_t(r, \hat{w}_2, \theta_-)$ is the maximal utility that the social planner can provide to an agent in period t if the agent has r amount of resources. Parameter $\hat{w}_2$ captures how much additional redistribution the planner promised in the previous period, as discussed in Lemma 2. That is, $H_t(r, \hat{w}_2, \theta_-)$ is an indirect utility function for an agent who was type θ_- in the previous period, has r units of savings, and faces the optimal schedule of distortions over his lifetime. With this notation, we can write a dual problem to (15) as

$$H_t(\hat{r}, \hat{w}_2, \theta_-) = \max_{u,c,y,r,w_2} \int \left(1 - \zeta \frac{f_{2,t}(\theta|\theta_-)}{f_t(\theta|\theta_-)}\right) u f_t(\theta|\theta_-)\, d\theta$$

subject to the incentive compatibility constraint (10), and

$$\int (c - y + \delta r)\, f_t(\theta|\theta_-)\, d\theta = \hat{r}, \tag{20}$$

$$u = U(c, y/\theta) + \beta H_{t+1}(r, w_2, \theta). \tag{21}$$

If we take w_2 as being set optimally, the optimization with respect to (u, c, r, y) in this problem is analogous to the maximization in the static model with labor and two goods, c and r, which have relative prices 1 and δ with respect to labor y. To simplify notation, we drop explicit notation with respect to t and θ_-, whenever this does not cause confusion. If μ is a multiplier on (10), λ is the multiplier on (20), and η is a multiplier on (21), the necessary conditions can be written, using notation $p_c = 1$, $p_r = \delta$, and $p_y = -1$ as

$$(-f p_j - \mu U_{\theta j}) = \eta U_j \text{ for } j \in \{c, y\}, \tag{22}$$

$$-f p_r \lambda = \eta \beta H_r, \tag{23}$$

$$\left(1 - \frac{f_2}{f}\zeta\right) f + \eta = \mu'. \tag{24}$$

These first order conditions are equivalent to those in a static model with two goods (see Mirrlees (1976); equations (33) and (34)). As Mirrlees (1976) discusses, generally we expect that the multiplier on the incentive compatibility constraint μ is nonnegative (which corresponds to downward binding incentive constraints), although it is difficult to rule out that μ can take negative values for some θ. We proceed by assuming that μ is non-negative everywhere, and show in the online appendix that this is indeed the case when $U_{cl} = 0$ or shocks are independent across periods.

2.1 Savings distortions

We first characterize savings distortions. Direct comparison of (22) and (23) shows that when $\mu \geq 0$,

$$\frac{1}{p_c} U_{c,t} \leq \frac{1}{p_r} \beta H_{r,t+1}, \tag{25}$$

with strict inequality when $U_{cl} = 0$. This result has a direct analogue in the static multi-good model. Mirrlees (1976) shows that when μ is non-negative, it is optimal to distort the good which is more complementary with leisure. The intuition for this result is as follows. The role of the distortions is to provide incentives for the agents with the high skill not to pretend to be of low skill. When such a deviation occurs, a deviating agent enjoys more leisure than the truth telling agent of the low type[5]. Therefore, taxing goods which are complementary with leisure helps to relax the incentive constraints. In the context of our model, consumption today, c, is a substitute with leisure, because $U_{cl} \geq 0$, while $H_{rl} = 0$. This implies that good r should be more distorted than good c.

[5] See Kaplow (2008) for a detailed discussion of nonseparable preferences in a static context.

To explore the implication of (25) for savings distortions, we now discuss the relationship between $H_{r,t+1}$ and $\mathbb{E}_t U_{c,t+1}$. By the envelope theorem, $H_{r,t+1}$ is equal to an increase in the expected utility from any incentive compatible allocation of an additional unit of resources tomorrow. An allocation of a unit of resources equally across for all the realizations of shocks in period $t+1$ is generally not incentive compatible, which implies that $H_{r,t+1} \neq \mathbb{E}_t U_{c,t+1}$. When preferences are separable between c and l it can be shown (see, e.g., Golosov, Kocherlakota, and Tsyvinski (2003) or Farhi and Werning (2009) that an incentive compatible perturbation increases utility of consumption, $U(c(\theta))$, equally for all the realization of θ, and that $H_{r,t+1} = \left(\mathbb{E}_t \frac{1}{U_{c,t+1}}\right)^{-1}$. When there is uncertainty about consumption in period $t+1$, Jensen's inequality implies that $H_{r,t+1} < \mathbb{E}_t U_{c,t+1}$, which, together with (25), implies that $\tau_t^s > 0$. When preferences are non-separable, $H_{r,t+1}$ can be greater or less than $\mathbb{E}_t U_{c,t+1}$, and in general it is not possible to sign τ_t^s. We show, however, that under some conditions it is possible to sign the life-time savings distortion, $\bar{\tau}_t^s$.

Integrate (24), using the boundary conditions $\mu(0) = \mu(\infty) = 0$ and the fact that $\int_0^\infty f_2 d\theta = 0$, to get $-\int \eta d\theta = 1$. If we substitute the expression for η from (23) and the equality $H_{r,t} = \lambda$, we get

$$\left(\frac{\delta}{\beta}\right) \mathbb{E}_t \frac{H_{r,t}}{H_{r,t+1}} = 1. \tag{26}$$

This expression is a generalization of the "inverse Euler equation" which is obtained with separable preferences.[6] Similarly to the inverse Euler equation, it implies that it is optimal to have a positive distortion between marginal utility of resources in period t, $H_{r,t}$, and period $t+1$, $\mathbb{E}_t H_{r,t+1}$. Iteration of (26) implies that

$$\left(\frac{\delta}{\beta}\right)^{T-t} \mathbb{E}_t \frac{H_{r,t}}{H_{r,T}} = 1.$$

[6] To see this, substitute (25), which holds with equality in separable case.

This expressions has important implications for the lifetime savings distortion, $\bar{\tau}_t^s$, in environments in which agents eventually retire. If all agents retire by period T, $H_{r,T} = U_c(c_T, 0)$, which, combining with inequality (26) implies that $\bar{\tau}_t^s > 0$. The simplest restriction on fundamentals that ensure retirement is the assumption that $\theta_T = 0$ with probability 1, for any history of shocks. We summarize this discussion in the following proposition

Proposition 1. *Suppose that assumptions 1, 5, and 6 hold. Suppose that* $F_T(0|\theta_-) = 1$ *for all* θ_- *and* $\mu \geq 0$. *Then* $\bar{\tau}_t^s \geq 0$, *with the strict inequality in all states in which the incentive constraint (10) is binding and* $U_{cl} > 0$.

Although it is difficult in general to characterize the cross-sectional implications for the optimal savings wedge τ^s, one can find an expression for the wedge τ^w defined as $\tau^w = 1 + \left(\frac{\beta}{\delta}\right) U_y / H_r$. By analogy with τ^y, which measures distortion between labor supply and consumption today, τ^w measures a distortion between labor supply and consumption tomorrow. In the appendix we show that

$$\tau^w = \left[1 + \frac{\varepsilon}{\varepsilon + 1} \frac{U_{cl}}{U_c} \left(\frac{y}{\theta}\right)\right] \tau^y. \tag{27}$$

In the next section we derive an expression for the labor wedge $\tau^y(\theta)$ and use expression (27) to get insights about dynamics of τ^w. We also will use it to show an important insights about asymptotic behavior of the wedges. As long as the degree of substitutability of consumption and leisure, U_{cl}/U_c does not go to zero, (27) shows that the gap between the two distortions, τ^w/τ^y widens if more productive types supply more effort. In the next section, we show that for some widely used classes of preferences this implies that τ^y must go to zero for high θ types.

2.2 Labor distortions

Next, we turn to characterization of labor distortions. Again, we can use the standard arguments of Mirrlees (1971) and Mirrlees (1976) to re-write the first order condition (22) as

$$\frac{\tau^y}{1-\tau^y} = \left(1+\frac{1}{\varepsilon}\right)\frac{\mu}{\theta f}U_c. \tag{28}$$

One can then express η from (22), substitute into (24), and integrate to obtain the expression for μ. In the appendix, we describe in details how to proceed and derive the expression that follows for the labor wedge:

$$\frac{\tau^y}{1-\tau^y} = \left(1+\frac{1}{\varepsilon}\right)\frac{1-F(\theta)}{\theta f(\theta)}\int_\theta^\infty \left\{\exp\left[\int_\theta^{\hat\theta}\left(\left(1-\frac{\zeta^u}{\zeta^c}\right)\frac{\dot y}{y}\right)d\theta'\right]\right.$$
$$\left.\times\left(1-\lambda\left(1-\frac{f_2(\hat\theta)}{f(\hat\theta)}\zeta\right)U_c(\hat\theta)\right)\exp\left[\int_\theta^{\hat\theta}\beta\frac{U_{cc}w_1}{(U_c)^2}d\theta'\right]\frac{f(\hat\theta)}{1-F(\theta)}\right\}d\hat\theta, \tag{29}$$

where ζ^c and ζ^u are compensated and uncompensated elasticities of labor supply *holding savings fixed*. We wrote the expression in terms of the elasticities holding savings fixed to facilitate comparison with the optimal taxes in the static models, e.g. Diamond (1998) and Saez (2001).

There are three key differences with the formula for the labor wedges in the static economy such as Saez (2001), Proposition 1. The first difference the term $\exp\left[\int_\theta^{\hat\theta}\beta\frac{U_{cc}w_1}{(U_c)^2}d\theta'\right]$, which depends on the future promised utility and on the current realization of the shock. This term is less than 1, which points out that there is a force that pushes the wedges lower in the dynamic setting. The second difference is that the social weight that the planner applies to agents of different types changes when the shocks are persistent differs from the true social weight $G'\left(\bar U_c(\theta)\right)$. We already discussed the intuition behind this changing the social welfare in Lemma 2. Finally, the expression $\frac{1-F_t(\theta|\theta_-)}{\theta f_t(\theta|\theta_-)}$ in

(29) depends on the distribution of shocks conditional on last period realization rather than the cross-sectional distribution of shocks in static models.

Diamond (1998) used the static analogue of (29) to show that if utility is quasi-linear then the first integral in (29) asymptotically converges to 1 and $\frac{\tau^y}{1-\tau^y}$ converges to $\left(1+\frac{1}{\varepsilon}\right)\frac{1-F^{cs}}{\theta f^{cs}}$ from below, where f^{cs} and F^{cs} denote the cross-sectional distribution of the types. His analysis can easily be extended to any quasi-linear preferences of the form

$$U(c,l)=\tilde{U}\left(c-h(l)\right). \tag{30}$$

As Diamond (1998) and Saez (2001) discuss, the empirical distribution of income $\frac{1-F^{cs}}{\theta f^{cs}}$ is such that the implied asymptotic labor distortions may be quite high, with $\tau^y \to 0.8$ for some specifications of ε. We show next that the conclusion about the size of the optimal labor distortions may change significantly in the dynamic models.

Let the preferences have the form (30). In this case, $U_{cl}/U_c=\frac{-\tilde{U}''}{\tilde{U}'}h'(l)$ and $h'(l)=(1-\tau^y)\,\theta$. Substitute these expressions into (27) to obtain

$$1+\frac{\delta}{\beta}\frac{U_y}{H_r}=\left[1+\frac{\varepsilon}{1+\varepsilon}\frac{-\tilde{U}''}{\tilde{U}'}\left(1-\tau^y\right)y\right]\tau^y,$$

where we used (28) and (30). The left hand side of this expression is less than 1. Consider the right hand side of this expression. Suppose that τ^y did not converge to 0. As long as τ^y is bounded away from 1, this implies that the term $(1-\tau^y)\,y$ goes to infinity. If the coefficient of the absolute risk aversion $\frac{-\tilde{U}''}{\tilde{U}'}$ is bounded away from zero, this implies that the right hand side of this expression is unbounded and eventually becomes greater than 1 which leads to a contradiction. Therefore, we have the following result which we prove formally in the appendix.

Proposition 2. *Suppose that assumptions 1, 2, 3, 5, 4, and 6 hold. Suppose that preferences are of the form (30), and $-U_{cc}/U_c$ is bounded away from zero. Suppose that $\left(1 - \frac{f_2(\hat{\theta})}{f(\hat{\theta})}\zeta\right) \geq 0$, or that $U_c(\hat{\theta})$ and $\frac{f_2(\hat{\theta})}{f(\hat{\theta})}$ are bounded from above. Then $\tau_t^y \to 0$ for $t < T$.*

3 Quantitative analysis

We now turn to the quantitative study of a calibrated model. The theoretical analysis above unambiguously points to empirical skill distributions as a crucial input for a quantitative analysis. Before we proceed to estimate skill distributions, we start by constructing a dataset of individual skills, θ, implied by the observed micro level data for the U.S.

The main difficulty in estimating skills from the data is that skills are unobservable. One can use wages as a proxy for skills but it is not clear that this measure corresponds to θ, which measures the return to *effort*. Because of these conceptual problems with using wage data we chose a different approach. We use the data on the actual U.S. tax code and labor income choices to infer the unobservable skill level. Since the details of the actual U.S. system of taxes and transfers are observable, we compute the implied individual skills from the necessary conditions for individual optimum. The methodology is a dynamic extension of that of Saez (2001) who used a similar approach to infer cross-sectional distribution of skills in the population.

Both to simplify the analysis and to be directly comparable to previous work of Diamond (1998) and Saez (2001), we choose quasi-linear preferences (30) with a constant elasticity of labor supply ε. For these preferences, the implied skill $\theta_{i,year}$ for an individual i in a given year is computed from the

individual first-order conditions as follows:

$$\theta_{i,year} = \frac{y_{i,year}}{\left\{y_{i,year}\left[1 - T'_{year}\left(y_{i,year}\right)\right]\right\}^{\varepsilon/(1+\varepsilon)}},$$

where $y_{i,year}$ is the labor income of individual i observed in a given $year$, and $T'_{year}\left(y_{i,year}\right)$ is the effective marginal tax rate that the individual faces when she earned her labor income. Since there are no income effects with quasi-linear preferences (30), an individual labor supply decision is unaffected by individual savings choice, and thus a static consumption-labor margin determines the implied skill.

We briefly outline our empirical and computational strategy. The online appendix contains a complete description of our approach with further details and summary statistics. Our main data source is the U.S. Panel Study of Income Dynamics (PSID). We use the data collection waves from 1990 onward with the latest currently available data wave of 2007, which contains data from 2006. Recent waves (from 1996) come in two year intervals, hence, we consider a total of nine waves with two years in between. To be consistent with the data, a period in our calibrated model correspond to two years, i.e., $T = 20$, and we model 40 years of working life. The labor income, $y_{i,year}$, is obtained directly from the PSID waves, converted to constant 1990 dollars. We consider total labor income, which is a sum of a list of variables in the PSID that contain data on salaries and wages, separate bonuses, the labor portion of business income, overtime pay, tips, commissions, professional practice or trade payments, market gardening, additional job income, and other miscellaneous labor income.[7]

Effective marginal tax rates, $T'_{year}\left(y_{i,year}\right)$, are estimated for each individual using TAXSIM - a National Bureau of Economic Research's program for cal-

[7] The online appendix contains specific details and variable names.

culating individual effective liabilities under the U.S. Federal and State income tax laws from individual data. For each individual with labor income we also have in the PSID a collection of personal data that are in most cases sufficient to estimate individual and year specific effective marginal tax rates. Finally, we use the constructed total individual labor incomes and the estimated effective marginal tax rates to compute implied skills from the individual optimality conditions as described above.

Estimation approach. We first estimate the initial unconditional distribution of implied skills among the initial young workers, $F_1(\theta)$. We consider the 25 year old from all of the PSID waves to obtain the sample of 8,231 observations. We estimate $F_1(\theta)$ non-parametrically using a kernel density estimation. The resulting distribution is shown as "initial young, unconditional" distribution depicted by a dotted line in Figure 1. One concern with a cross-sectional distribution is that high income individuals may be undersampled in the PSID or that the PSID is "top coded", i.e., there is an income cutoff level above which no observations are collected. To address this concern we fit a Pareto distribution to the right tail of our skill distribution. Specifically, we let skills to be Pareto distributed above the income level of $\$150,000$.

To estimate conditional distributions $F_t(\theta|\theta\_)$, i.e. transition probabilities, we exploit the panel feature of the PSID. We start by considering all individual skill transitions between adjacent data waves. Furthermore, we break all these wave-to-wave skill transitions into two age groups – when the individual is younger than 45 at the beginning of the transition and when the individual is 45 or older at the beginning of the transition. We therefore estimate two separate conditional distributions $F_{young}(\theta|\theta\_)$ and $F_{old}(\theta|\theta\_)$. Hence we assume age dependence between the age groups and age-independent transitions within

each age group.[8] In other words, we allow younger individuals to experience different transition probabilities than older individuals; within each age group, we assume age-independent transition probabilities.[9]

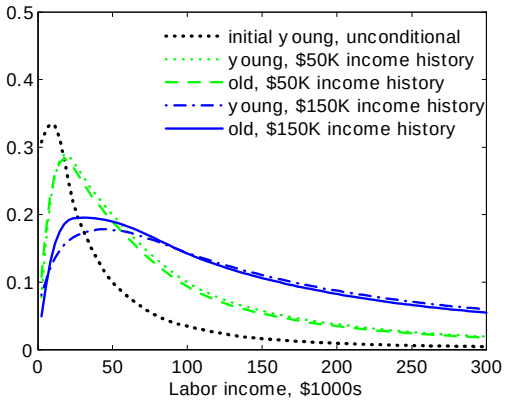

Figure 1: Initial unconditional vs. some of the conditional distributions

To provide an easy comparison, Figure 1 displays the initial unconditional distribution (the dotted line) together with just four examples of the estimated conditional distributions, two from each age group. An old skill distribution conditional on $\$50,000$ income, with the conditional expected income of $\$54,937$, appears as the closest to the initial unconditional distribution. The difference between the conditional expected income and the unconditional expected income of $\$23,963$ is primarily due to a significantly thicker right tail. A young skill distribution conditional on $\$50,000$ income is the next closest to the unconditional one with the conditional expected income of $\$58,599$. The

[8] We also check our results for robustness by removing this age dependence in the estimated conditional distributions.

[9] We stop at just two age groups to have sufficient number of data points to estimate all conditional distributions. There is nothing in our computational solution method that would stop us from having a different transition matrix for each period, provided that we had enough data to obtain those transition matrices.

other two examples of conditional distributions differ more significantly from the initial unconditional distribution. An old skill distribution conditional on $\$150,000$ income has the conditional expected income of $\$109,893$. A young distribution conditional on $\$150,000$ income has the conditional expected income of $\$115,508$. If one were to compare the initial unconditional distribution with a cross-sectional unconditional distribution, it would reveal that the initial young unconditional distribution appears less unequal with a somewhat thinner right tail, which is perhaps not surprising for a sample of young 25 year old workers.

The estimated conditional distributions imply a persistent skill shock process. Depending on the specification of the stochastic process, estimates of the persistence of skills simulated with our conditional distributions range from 0.73 to 0.81. These persistence estimates are not as high as the estimates of 0.95 and higher in Storesletten, Telmer, and Yaron (2004) and are closer to the estimates around 0.8 in Guvenen (2009). All these estimates are significantly higher than the estimate of 0.5 in Heaton and Lucas (1996), who do not condition on age.

Calibration. We model a life cycle of 40 years of working life, i.e., the individuals between the ages of 25 and 65, with one period representing two years as dictated by our main data source, the PSID. We choose preferences of the form

$$-\frac{1}{\psi}\exp\left(-\psi\left(c-\frac{l^{\gamma}}{\gamma}\right)\right).$$

As we explain in the online appendix, exponential preferences allow to reduce the dimension of the state space in the recursive formulation. We set the coefficient of *absolute* risk aversion, ψ, equal to 10. Numerical simulations are re-scaled so that consumption ranges from 0.1 to 1 implying that relative risk aversion ranges from 1 to 10 when $\psi = 10$. The elasticity parameter, γ, is set

to 3 with the Frisch elasticity of labor supply then $\varepsilon = 1/(\gamma - 1) = 0.5$. The annual discount factor is $\beta = 0.9804$ and the marginal rate of transformation across years is $\delta = 1.02$ so that the social planner at the solution of the optimal program chooses not to transfer resources between periods.

Finally, in the benchmark analysis we assume that the social welfare criterion is utilitarian. We discuss the implications of relaxing this assumption later in the section.

Results. Given the parameters of the calibrated model and the empirically estimated skill distributions, we proceed to solve the model numerically by exploiting the recursive formulation of the dual problem. Figure 2 presents the results of numerical simulations. Consider first panels A and B. Panel A displays labor distortions at the initial age of 25 and at ages 27, 35, 45, 55, and 65 for the agent with a history of shocks up to that period such that in each previous period he had income of $\$50,000$. Panel B displays labor distortions at the same ages for the agent with a history of shocks up to that period such that in each previous period she had income of $\$150,000$. Both for agents with $\$50,000$ and $\$150,000$ income histories, the lowest, dotted line is the unconditional labor wedge at the initial age of 25, which is identical in both panels, and generally higher lines represent distortions at older ages. These two examples correspond to the two examples (for each age group) of the estimated conditional distributions of skills in Figure 1.

There are several key features of interest with the labor wedge results. First, both for the agent with the history of $\$50,000$ incomes and for the agent with the history of $\$150,000$ incomes, the average conditional labor wedges are increasing with age. This is consistent with our theoretical findings where the provision of incentives dynamically allows to lower labor wedges early in life. The planner then wants to distort the provision of the incentives in the future,

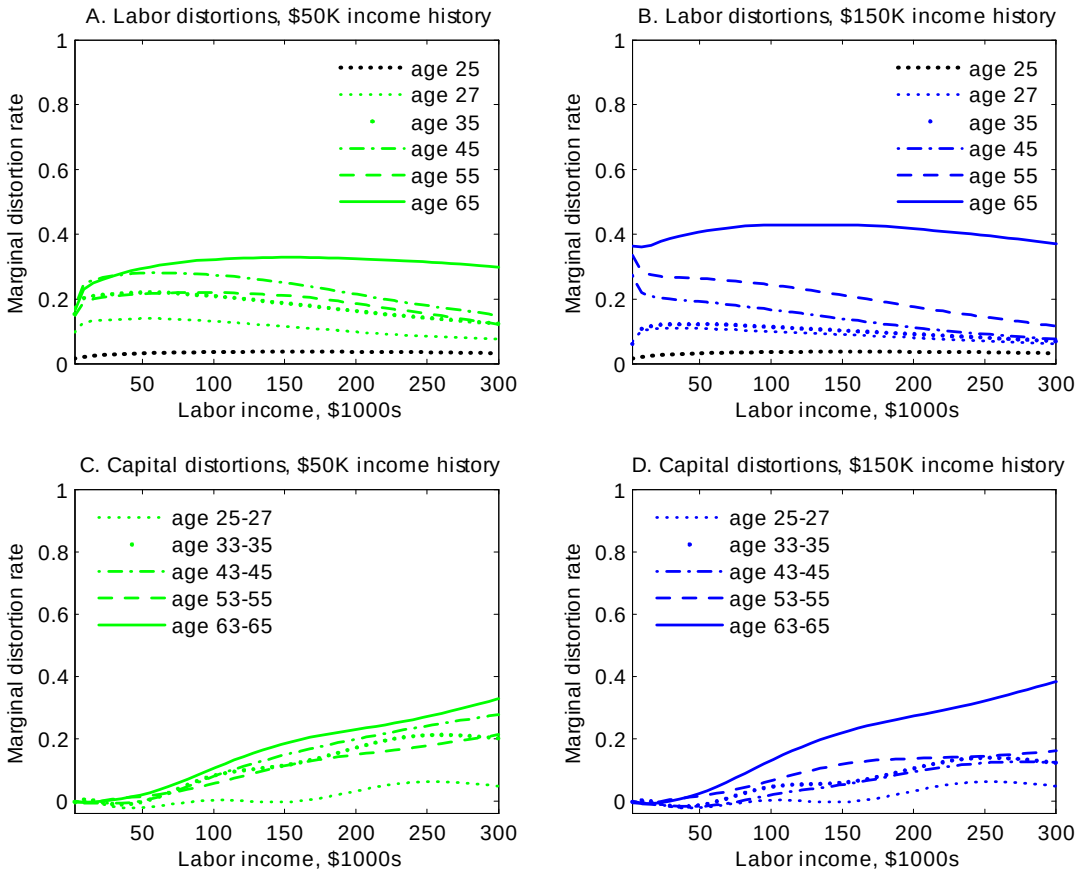

Figure 2: Labor and capital distortions with persistent shocks

and therefore to distort the intertemporal margin. The further away the agents from the end of the working life, the bigger are the planner's opportunities to distort the intertemporal margin which also relates to our analysis of the lifetime savings distortion. This allows the planner to substitute the labor wedge for the intertemporal wedge. As agents near the end of the working life, the power of the intertemporal distortions decreases, and the labor wedges are higher.

Second, the conditional labor wedges for an agent with a history of $\$50,000$ incomes are generally lower, especially at older ages, than those for an agent

of the same age with a history of incomes of $150,000. In particular at age 65, the distortions for an agent with a $50,000 income history start from 0.16 and increase to 0.33 at $150,000 current income before decreasing to 0.27 at $300,000 current income. The distortions for an agent with a $150,000 income history start from 0.36 and increase to 0.43 at $150,000 current income before decreasing to 0.34 at $300,000 current income. There are two forces driving these differences that follow from the discussion in the theoretical analysis: (*i*) the additional redistribution over time implied by the persistent shocks and (*ii*) the differences between conditional and unconditional distributions of skills as well as the differences in conditional distributions among agents, specifically between those with a history of relatively low incomes and those with a history of relatively high incomes as is evident from the discussion of the examples in Figure 1 above.

Third, consistent with Proposition 2, the labor wedge decreases for the high incomes at every age for any history. Figure 2 shows this for two particular histories of $50,000 incomes and $150,000 incomes with the decrease in wedges at all ages for incomes above $150,000.

Next, consider capital wedges in panels C and D of Figure 2. Panel C presents capital distortions between the initial age of 25 and age 27, and between ages 33-35, 43-45, 53-55, and 63-65, with generally lower lines representing younger ages, for the agent with a history of shocks up to that period such that in each previous period he had income of $50,000. Panel D displays savings distortions at the same ages, once again with generally lower lines representing younger ages, for the agent with a history of shocks up to that period such that in each previous period she had income of $150,000. In both examples, the conditional savings distortions are generally increasing in current period realization of income as well as with age. The distortions are

close to zero for current incomes below $\$50,000$. For the agents with a history of $\$50,000$ incomes, savings distortions at current income of $\$300,000$ reach as high as 0.06 at ages 25-27 and 0.34 at ages 63-65. For the agents with a history of $\$150,000$ incomes, savings distortions reach at current income of $\$300,000$ as high as 0.07 at ages 25-27 and 0.39 at ages 63-65.

Welfare losses from simple tax policies. Farhi and Werning (2010) solve a dynamic model with idiosyncratic shocks and argue that in their parametrization the fully optimal non-linear tax system can be well approximated by simpler linear taxes when the social planner is utilitarian. We use a different stochastic process for skills and assume different preferences. Still, as figure 2, the optimal history-dependent non-linear labor distortions appear to be mostly flat. Next, we explore the magnitude of welfare losses in our model from using simpler tax instruments.

A natural benchmark for comparison is optimal linear taxes. We keep optimal savings distortions and transfer all revenues lump sum back to the agents. First, consider the case of the utilitarian social planner. Using optimal age-dependent linear labor wedges instead of the constrained optimal wedges results in a welfare loss of 0.9% of consumption equivalent. Using optimal age-*in*dependent labor distortions increases the welfare loss to 1.6%. While these magnitudes are non-trivial, linear taxes can still yield reasonably good policies.

The result that welfare gains from non-linear taxes are small for utilitarian social planner has parallels in the static Mirrlees models. Mirrlees (1971), Atkinson and Stiglitz (1976), Tuomala (1990) also found in numerical simulations that the optimal labor distortions appear to be mostly flat in this case. They also argued that as the importance of nonlinearities increases, the social planner becomes more redistributive. To investigate this in dynamic settings,

we compute the welfare gains of using optimal policies when the social planner is Rawlsian. The optimal age-dependent linear labor wedges yield a welfare loss of 4.6% of consumption compared to the constrained optimum. The optimal age-*in*dependent labor wedges yield a welfare loss of 5.1%. We conclude that the welfare gains of using optimal non-linear policies are significant.

4 Conclusion

In this paper, we take a step toward reconciling two literatures: the dynamic optimal taxation literature and the classic static optimal taxation literature. We show that a dynamic optimal taxation model shares many similarities with the static model with two goods. A suitably written recursive formulation of the dynamic model separates the analysis as one involving utility of consumption (and labor) today and the future utility over transferred resources. This allows one to use the insights of the static literature to study the optimal dynamic labor and savings wedges and extend the analysis of the forces behind the optimal wedges in Diamond (1998) and Saez (2001) to the dynamic settings. We show that while there are many similarities, the dynamics importantly alters the prescription of the static optimal taxation literature. Perhaps, the most important difference is that while the static literature prescribes labor taxes to be as high as 50-70% for the high skilled individuals in the calibrated models, in the dynamic settings the labor wedge tends to zero for the high skilled. Other important differences include the use of conditional rather than unconditional distribution of skills, an ability to use intertemporal distortions to lower labor wedges, especially, early in the life of the agents, and the behavior and implications for the savings distortions not present in the static models.

Importantly, we calibrate our model by estimating the skill distribution and its evolution over the lifetime as this is one of the key determinants of the dynamic labor and savings distortions. We compute optimal labor and savings distortions and find them significantly different from the static ones. As such, we conclude that while conceptually one can think of the dynamic optimal taxation model in its recursive form as of a static model with two goods, the dynamics adds important insights that significantly change the policy prescriptions.

Conesa, Kitao, and Krueger (2009) take a different approach to the analysis of the dynamic optimal taxation models. They study tax reforms and the optimal taxes within a set of the parametrically restricted tax functions. One advantage of that approach over solving for the full informationally constrained optimum is that it is computationally more feasible and allows one to study optimal taxes which are most commonly used in practice. Our paper points out to the elements that may be important in choosing the parameters of such functions.

5 Appendix

To keep our analysis here succinct, let $x = (c, y)$, $p = (1, -1)'$, $p_c = 1$, $p_y = -1$, and use shorthand notation for utility $U(x, \theta)$. Then a Hamiltonian to (9)

$$\mathcal{H} = \left(\lambda u \left(1 - \frac{f_2}{f}\zeta\right) - (px + \delta V)\right) f + \eta \left[u - U - \beta w\right] - \mu U_\theta - \mu\beta w_2,$$

where μ, λ, $-\zeta/\lambda$, and η are the respective multipliers on (10), (11), (12), and (13).

The first order conditions are as follows: with respect to good $j \in \{c, y\}$:

$$-fp_j - \mu U_{\theta j} = \eta U_j; \tag{31}$$

with respect to u

$$\lambda\left(1-\frac{f_2}{f}\zeta\right)f+\eta=\mu'; \tag{32}$$

with respect to w and w_2 (where we use V_1 and V_2 to denote relevant cross partial derivatives)

$$-\delta V_{1,t+1}f=\eta\beta, \tag{33}$$

and

$$-\delta V_{2,t+1}f=\mu\beta. \tag{34}$$

Also, note that the envelope theorem implies that

$$V_{1,t+1}=\lambda_t, V_{2,t+1}=-\zeta/\lambda_t. \tag{35}$$

Use (31) for c to find η and substitute it into (31) for y

$$\begin{aligned}\left(p_c\frac{U_y}{U_c}-p_y\right)f &= \mu\left(U_{y\theta}-\frac{U_y}{U_c}U_{c\theta}\right)\\ &= -\frac{\mu U_c}{\theta}\frac{U_y}{U_c}\left(1+\frac{1}{\varepsilon}\right).\end{aligned}$$

Since $p_c\frac{U_y}{U_c}-p_y=\tau^y$, it implies that

$$\frac{\tau^y}{1-\tau^y}=\left(1+\frac{1}{\varepsilon}\right)U_c\frac{\mu}{\theta f}. \tag{36}$$

This expression shows that if U satisfies Assumption 1 then the sign of τ^y is equal to the sign of μ. The expression for the multiplier μ can be obtained by integrating (32) with a boundary condition $\mu(0)=0$. By substituting η from the FOCs for w or c, we show in the online appendix that

$$\begin{aligned}\mu &= \int_0^\theta\left[\lambda\left(1-\frac{f_2(\hat\theta)}{f(\hat\theta)}\zeta\right)-\frac{\delta}{\beta}V_1(\hat\theta)\right]f(\hat\theta)d\hat\theta \qquad (37)\\ &= \int_0^\theta\left(\lambda\left(1-\frac{f_2(\hat\theta)}{f(\hat\theta)}\zeta\right)-\frac{1}{U_c(\hat\theta)}\right)\exp\left[-\int_{\hat\theta}^\theta\frac{U_{c\theta}(\theta')}{U_c(\theta')}d\theta'\right]f(\hat\theta)d\hat\theta\end{aligned}$$

As we already discussed, it is difficult to determine the sign of μ and λ. The analogous problem arises in the static model with multiple goods, as explained by Mirrlees (1976). Similar to that literature, we focus on the case of $\mu \geq 0$ and $\lambda \geq 0$. In the online appendix we show that these assumptions are indeed satisfied if preferences are separable or shocks are i.i.d.

We characterize the savings distortions first. It is useful to characterize the distortion τ^w, defined as

$$1 - \tau^w \equiv -\frac{\delta}{\beta} V_1 U_y. \tag{38}$$

This is a wedge between the marginal cost of labor today and the marginal cost of providing one util to this agent tomorrow. Use the first order conditions (31) and (33) to obtain

$$\left(1 - \frac{\delta}{\beta} V_1 U_c\right) f = -\mu U_{c\theta}. \tag{39}$$

Since $U_{c\theta} = U_{cl}\left(-\frac{y}{\theta^2}\right) \leq 0$,

$$\frac{\delta}{\beta} V_{1,t+1} U_{c,t} \leq 1. \tag{40}$$

We can also rewrite (39) as

$$\left(1 - \frac{\delta}{\beta} V_1 \left(-U_y\right) \frac{U_c}{-U_y}\right) = \frac{\mu U_c}{\theta f} \frac{U_{cl}}{U_c} \left(\frac{y}{\theta}\right).$$

Substitute (36) and the definitions of τ^y and τ^w and re-arrange

$$\tau^w = \left[1 + \frac{\varepsilon}{\varepsilon + 1} \frac{U_{cl}}{U_c} \left(\frac{y}{\theta}\right)\right] \tau^y. \tag{41}$$

This is the expression (27) in the body of the paper. Since the expression in square brackets is greater than 1, this implies that $\tau^w \geq \tau^y$ (see the online appendix for details).

When $U_{cl} = 0$ this expression implies the Inverse Euler Equation. To see this consider the the second line of (37). The boundary condition $\mu(\infty)$ implies

that $\lambda = \int_0^\infty \frac{1}{U_c(\hat{\theta})} f(\hat{\theta}) d\hat{\theta}$. The envelope condition (35) gives the expression for λ, so that $V_{1,t+1} = \mathbb{E}_t \left(U_{c,t+1}\right)^{-1}$. Combine it with (40) (which holds with equality in separable case) to get $\frac{\delta}{\beta} \mathbb{E}_t \frac{U_{c,t}}{U_{c,t+1}} = 1$.

We now characterize the lifetime saving distortion. Use the first line of (37) with the boundary condition $\mu(\infty) = 0$ to get

$$\frac{\delta}{\beta} \mathbb{E}_t V_{1,t+1} = \lambda_t = V_{1,t} \tag{42}$$

where the second equality follows from (35). Since $V_1 = 1/H_r$, this equation implies (26).

We now ready to prove Proposition 1. When there is no labor supply in the last period, then $U(c_T, 0) = w$ and $V_T(w) = U^{-1}(w, 0)$. Therefore, $V_{1,T} = \frac{1}{U_c(c_T,0)}$. Use (42) to show that

$$V_{1,t} = \frac{\delta}{\beta} \mathbb{E}_t \left\{V_{1,t+1}\right\} = \left(\frac{\delta}{\beta}\right)^2 \mathbb{E}_t \left\{\mathbb{E}_{t+1} \left\{V_{1,t+2}\right\}\right\} = \cdots = \left(\frac{\delta}{\beta}\right)^{T-t} \mathbb{E}_t \left\{\frac{1}{U_{c,T}}\right\}.$$

Then (40) implies that

$$1 \geq \left(\frac{\delta}{\beta}\right)^{T-t} U_{c,t} \mathbb{E}_t \left\{\frac{1}{U_{c,T}}\right\}, \text{ for all } t,$$

which by Jensen's inequality implies that

$$1 \geq \left(\frac{\delta}{\beta}\right)^{T-t} \frac{U_{c,t}}{\mathbb{E}_t \left\{U_{c,T}\right\}},$$

and proves Proposition 1.

We now explore the determinants of the labor wedges. A key term in (36) is $U_c\left(\theta\right) \mu(\theta)$. Use the second line of (37) and the boundary conditions on μ to determine $U_c \mu$:

$$\begin{aligned} & U_c\left(\theta\right) \mu(\theta) \\ & \quad = U_c\left(\theta\right) \int_\theta^\infty \left\{ \frac{1}{\tilde{U}_c(\hat{\theta})} \left(1 - \lambda \left(1 - \frac{f_2(\hat{\theta})}{f(\hat{\theta})} \zeta\right) U_c(\hat{\theta})\right) \right. \\ & \qquad \left. \times \exp\left[\int_\theta^{\hat{\theta}} \frac{U_{c\theta}(\theta')}{U_c(\theta')} d\theta'\right] f(\hat{\theta}) \right\} d\hat{\theta} \end{aligned}$$

$$
\begin{aligned}
&= \int_{\theta}^{\infty} \left\{ \exp\left[-\int_{\theta}^{\hat{\theta}} \frac{\partial \ln U_c(\theta')}{d\theta'} d\theta' \right] \left(1 - \lambda \left(1 - \frac{f_2(\hat{\theta})}{f(\hat{\theta})} \zeta \right) U_c(\hat{\theta}) \right) \right. \\
&\quad \left. \times \exp\left[\int_{\theta}^{\hat{\theta}} \frac{U_{c\theta}(\theta')}{U_c(\theta')} d\theta' \right] f(\hat{\theta}) \right\} d\hat{\theta} \\
&= \int_{\theta}^{\infty} \left\{ \exp\left[\int_{\theta}^{\hat{\theta}} \frac{U_{c\theta} - \left(U_{cc}(\theta')\dot{c} + U_{cl}\dot{l} \right)}{U_c(\theta')} d\theta' \right] \right. \\
&\quad \left. \times \left(1 - \lambda \left(1 - \frac{f_2(\hat{\theta})}{f(\hat{\theta})} \zeta \right) U_c(\hat{\theta}) \right) f(\hat{\theta}) \right\} d\hat{\theta}.
\end{aligned}
$$

Use $u'(\theta) = U_c\dot{c} + U_l\dot{l} + \beta w_1 + \beta w_2$ and (10) to derive $\dot{c} = \left[U_\theta - U_l\dot{l} - \beta w_1 \right] / U_c$. This implies that

$$
\begin{aligned}
&\int_{\theta}^{\hat{\theta}} \frac{U_{c\theta} - \left(U_{cc}(\theta')\dot{c} + U_{cl}\dot{l} \right)}{U_c(\theta')} d\theta' \\
&= \int_{\theta}^{\hat{\theta}} \left(\left(\frac{U_{c\theta}}{U_c} - \frac{U_{cc}U_\theta}{(U_c)^2} \right) - \left(\frac{U_{cl}}{U_c} - \frac{U_{cc}U_l}{(U_c)^2} \right) \dot{l} + \beta \frac{U_{cc}w_1}{(U_c)^2} \right) d\theta' \\
&= \int_{\theta}^{\hat{\theta}} \left(\left(l\frac{U_{cc}U_l}{(U_c)^2} - l\frac{U_{cl}}{U_c} \right) \frac{\dot{y}}{y} + \beta \frac{U_{cc}w_1}{(U_c)^2} \right) d\theta'.
\end{aligned}
$$

Following Saez (2003) we can show that $\left(l\frac{U_{cc}U_l}{(U_c)^2} - l\frac{U_{cl}}{U_c} \right) = \frac{\zeta^c - \zeta^u}{\zeta^c}$ and obtain the expression for the labor wedge (29).

We now proceed to the proof of Proposition 2. If $\left(1 - \frac{f_2(\hat{\theta})}{f(\hat{\theta})}\zeta \right) \geq 0$, for all $\hat{\theta}$, or if $\left(1 - \lambda \left(1 - \frac{f_2(\hat{\theta})}{f(\hat{\theta})}\zeta \right) \tilde{U}_c(\hat{\theta}) \right)$ is bounded from above, then expression (29) implies that τ^y is bounded away from 1. With quasi-linear preferences

$$
h'(l) = (1 - \tau^y)\,\theta. \tag{43}
$$

Therefore, as τ^y is bounded away from 1, $l(\theta) \to \infty$. Suppose τ^y does not converge to zero. Then, since τ^y is bounded from above, condition (43) implies that $l \to \infty$. When U is of the form (30), $U_{cl}/U_c = \frac{-U_{cc}}{U_c} h' = \frac{-U_{cc}}{U_c} (1 - \tau^y)\,\theta$. Since $\frac{-U_{cc}}{U_c}$ is bounded away from zero, this expression becomes arbitrarily large

for some θ when τ^y does not converge to zero. From (41) this implies that from some θ, $\tau^w(\theta) > 1$. However, since $U_y \leq 0$ and $V_1 \geq 0$, expression (38) implies that $\tau^w(\theta) \leq 1$ for all θ, which leads to a contradiction.

References

Acemoglu, Daron, Mikhail Golosov, and Aleh Tsyvinski. 2008. "Markets versus governments." *Journal of Monetary Economics* 55 (1):159–189.

Albanesi, Stefania and Christopher Sleet. 2006. "Dynamic optimal taxation with private information." *Review of Economic Studies* 73 (1):1–30.

Ales, Laurence and Pricila Maziero. 2007. "Accounting for private information." *working paper* .

Atkinson, Anthony and Joseph E. Stiglitz. 1976. "The design of tax structure: direct versus indirect taxation." *Journal of Public Economics* 6 (1-2):55–75.

Battaglini, Marco and Stephen Coate. 2008. "Pareto efficient income taxation with stochastic abilities." *Journal of Public Economics* 92 (3-4):844–868.

Conesa, Juan Carlos, Sagiri Kitao, and Dirk Krueger. 2009. "Taxing capital? Not a bad idea after all!" *American Economic Review* 99 (1):25–48.

Diamond, Peter. 1998. "Optimal Income Taxation: An Example with a U-Shaped Pattern of Optimal Marginal Tax Rates." *American Economic Review* 88 (1):83–95.

Farhi, Emmanuel and Iván Werning. 2009. "Capital Taxation: Quantitative Explorations of the Inverse Euler Equation." *working paper* .

———. 2010. "Insurance and Taxation over the Life Cycle." *working paper* .

Fernandes, Ana and Christopher Phelan. 2000. "A recursive formulation for repeated agency with history dependence." *Journal of Economic Theory* 91 (2):223–247.

Fukushima, Kenichi. 2010. "Quantifying the welfare gains from flexible dynamic income tax systems." *mimeo* .

Golosov, Mikhail, Narayana Kocherlakota, and Aleh Tsyvinski. 2003. "Optimal Indirect and Capital Taxation." *Review of Economic Studies* 70 (3):569–587.

Golosov, Mikhail and Aleh Tsyvinski. 2006. "Designing optimal disability insurance: A case for asset testing." *Journal of Political Economy* 114 (2):257–279.

Golosov, Mikhail, Aleh Tsyvinski, and Iván Werning. 2006. "New dynamic public finance: A user's guide." *NBER Macroeconomics Annual* 21:317–363.

Guvenen, Fatih. 2009. "An empirical investigation of labor income processes." *Review of Economic Dynamics* 12:58–79.

Heaton, John and Deborah J. Lucas. 1996. "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing." *Journal of Political Economy* 104:443–487.

Judd, Kenneth L. 1996. "Approximation, perturbation, and projection methods in economic analysis." *Handbook of Computational Economics* 1:509–585.

———. 1998. *Numerical methods in economics*. The MIT Press.

Kapička, Marek. 2010. "Efficient allocations in dynamic private information economies with persistent shocks: A first order approach." *working paper* .

Kaplow, Louis. 2008. *The Theory of Taxation and Public Economics*. Princeton University Press.

Kocherlakota, Narayana. 2010. *The New Dynamic Public Finance*. Princeton University Press, USA. Forthcomming.

Kocherlakota, Narayana and Luigi Pistaferri. 2007. "Household Heterogeneity and Real Exchange Rates." *The Economic Journal* 117 (512):C1–C25.

———. 2009. "Asset pricing implications of Pareto optimality with private information." *Journal of Political Economy* 117 (3):555–590.

Mirrlees, James. 1971. "An Exploration in the Theory of Optimum Income Taxation." *Review of Economic Studies* 38 (2):175–208.

———. 1976. "Optimal tax theory: A synthesis." *Journal of Public Economics* 6 (4):327–358.

———. 1986. "The theory of optimal taxation." *Handbook of mathematical economics* 3:1197–1249.

Pavan, Alessandro, Ilya Segal, and Juuso Toikka. 2010. "Dynamic mechanism design: Incentive compatibility, profit maximization and information disclosure." *working paper* .

Saez, Emmanuel. 2001. "Using Elasticities to Derive Optimal Income Tax Rates." *Review of Economic Studies* 68 (1):205–229.

Storesletten, Kjetil, Christopher I. Telmer, and Amir Yaron. 2004. "Consumption and risksharing over the life cycle." *Journal of Monetary Economics* 51:609–633.

Tuomala, Matti. 1990. *Optimal income tax and redistribution.* Oxford University Press, USA.

Weinzierl, Matthew. 2011. "The Surprising Power of Age-Dependent Taxes." *Review of Economic Studies* 78 (4):1490–1518.

A Appendix For Online Publication

A.1 Proof of Lemma 1

Note that given any solution $u^*(\theta)$ following a sequence of reports $\left(\theta^{t-1}, \theta\right)$, we can construct

$$\omega\left(\theta|\hat{\theta}\right) = \int_0^{\infty} u^*\left(\theta^{t-1}, \theta, s\right) f_{t+1}\left(s|\hat{\theta}\right) ds.$$

We can re-write (6) as

$$\max_{\hat{\theta}} \mathcal{V}\left(\hat{\theta}; \theta\right) \equiv \max_{\hat{\theta}} U\left(c(\hat{\theta}), y(\hat{\theta}); \theta\right) + \beta\omega(\hat{\theta}|\theta).$$

If $\hat{\theta} = \theta$ is a local maximum for all θ, then

$$U_c\left(c(\hat{\theta}), y(\hat{\theta}); \hat{\theta}\right) c'\left(\hat{\theta}\right) + U_y\left(c(\hat{\theta}), y(\hat{\theta}); \hat{\theta}\right) y'\left(\hat{\theta}\right) + \beta\omega_1(\hat{\theta}|\hat{\theta}) = 0, \text{ for all } \hat{\theta}, \tag{44}$$

or equivalently

$$\begin{aligned} &U_c\left(c(\hat{\theta}), y(\hat{\theta}); \hat{\theta}\right) y'\left(\hat{\theta}\right) \\ &\times \left[\frac{c'\left(\hat{\theta}\right)}{y'\left(\hat{\theta}\right)} + \frac{U_y\left(c(\hat{\theta}), y(\hat{\theta}); \hat{\theta}\right)}{U_c\left(c(\hat{\theta}), y(\hat{\theta}); \hat{\theta}\right)} + \beta\frac{\omega_1(\hat{\theta}|\hat{\theta})}{y'\left(\hat{\theta}\right)} \frac{1}{U_c\left(c(\hat{\theta}), y(\hat{\theta}); \hat{\theta}\right)}\right] = 0, \text{for all } \hat{\theta}. \end{aligned}$$

Note that from (44) and assumptions 1 and 5 $y' \geq 0$.

We argue next that for any θ^*, $\mathcal{V}_1\left(\hat{\theta}; \theta^*\right) \geq 0$ for all $\hat{\theta} < \theta^*$, and $\mathcal{V}_1\left(\hat{\theta}; \theta^*\right) \leq 0$ for all $\hat{\theta} > \theta^*$, which establishes the lemma. Differentiating $\mathcal{V}$, one obtains

$$\begin{aligned} \mathcal{V}_1\left(\hat{\theta}; \theta^*\right) &= U_c\left(c(\hat{\theta}), y(\hat{\theta}); \theta^*\right) c'\left(\hat{\theta}\right) + U_y\left(c(\hat{\theta}), y(\hat{\theta}); \theta^*\right) y'\left(\hat{\theta}\right) + \beta\omega_1\left(\hat{\theta}|\theta^*\right) \\ &= U_c\left(c(\hat{\theta}), y(\hat{\theta}); \theta^*\right) y'\left(\hat{\theta}\right) \\ &\quad \times \left[\frac{c'\left(\hat{\theta}\right)}{y'\left(\hat{\theta}\right)} + \frac{U_y\left(c(\hat{\theta}), y(\hat{\theta}); \theta^*\right)}{U_c\left(c(\hat{\theta}), y(\hat{\theta}); \theta^*\right)} + \beta\frac{\omega_1\left(\hat{\theta}|\theta^*\right)}{y'\left(\hat{\theta}\right)} \frac{1}{U_c\left(c(\hat{\theta}), y(\hat{\theta}); \theta^*\right)}\right] \end{aligned}$$

$$
\begin{aligned}
= \; & U_c\left(c(\hat{\theta}), y(\hat{\theta}); \theta^*\right) y'\left(\hat{\theta}\right) \\
& \times \left[\left\{ \frac{U_y\left(c(\hat{\theta}), y(\hat{\theta}); \theta^*\right)}{U_c\left(c(\hat{\theta}), y(\hat{\theta}); \theta^*\right)} - \frac{U_y\left(c(\hat{\theta}), y(\hat{\theta}); \hat{\theta}\right)}{U_c\left(c(\hat{\theta}), y(\hat{\theta}); \hat{\theta}\right)} \right\} \right. \\
& \left. + \frac{\beta}{y'\left(\hat{\theta}\right)} \left(\frac{\omega_1\left(\hat{\theta}|\theta^*\right)}{U_c\left(c(\hat{\theta}), y(\hat{\theta}); \theta^*\right)} - \frac{\omega_1\left(\hat{\theta}|\hat{\theta}\right)}{U_c\left(c(\hat{\theta}), y(\hat{\theta}); \hat{\theta}\right)} \right) \right]
\end{aligned}
$$

The term in curly brackets takes the sign of $\theta^* - \hat{\theta}$ for any utility function that satisfies assumption 1. If $\theta^* \geq (\leq)\hat{\theta}$, then $U_c\left(c(\hat{\theta}), y(\hat{\theta}); \theta^*\right) \leq (\geq) U_c\left(c(\hat{\theta}), y(\hat{\theta}); \hat{\theta}\right)$, which, together with the assumptions that $y' \geq 0$ and w_1 is increasing in the second argument, implies that the second term in the square brackets also takes the same sign as $\theta^* - \hat{\theta}$. Since $u_c > 0$ and $y' \geq 0$, this establishes that $\mathcal{V}_1\left(\hat{\theta}; \theta^*\right)$ has the sign of $\theta^* - \hat{\theta}$.

A.2 Proof of Lemma 2

Consider a Hamiltonian to (9) and use (13) to substitute for $w(\theta)$

$$
\begin{aligned}
H \; = \; & -\left(c\left(\theta\right) - y\left(\theta\right) + \delta V_{t+1}\left(\beta^{-1}\left(u(\theta) - U(c(\theta), y(\theta)/\theta)\right), w_2(\theta), \theta\right)\right) f_t\left(\theta|\theta_-\right) \\
& + \mu(\theta)\left[U_l(c(\theta), y(\theta)/\theta)\left(-\frac{y\left(\theta\right)}{\theta^2}\right) + \beta w_2\left(\theta\right)\right] \\
& - pu\left(\theta\right) f\left(\theta|\theta_-\right) - p_2 u\left(\theta\right) f_2(\theta|\theta_-) \\
= \; & -\left(c\left(\theta\right) - y\left(\theta\right) + \delta V_{t+1}\left(\beta^{-1}\left(u(\theta) - U(c(\theta), y(\theta)/\theta)\right), w_2(\theta), \theta\right)\right) f_t\left(\theta|\theta_-\right) \\
& + \mu(\theta)\left[U_l(c(\theta), y(\theta)/\theta)\left(-\frac{y\left(\theta\right)}{\theta^2}\right) + \beta w_2\left(\theta\right)\right] \\
& - \left(1 + \frac{p_2}{p}\frac{f_2(\theta|\theta_-)}{f\left(\theta|\theta_-\right)}\right) pu\left(\theta\right) f\left(\theta|\theta_-\right)
\end{aligned}
$$

and let $(c^*, y^*, w_2^*, \mu^*, p^*, p_2^*)$ be a solution. Let $\zeta = -p_2^*/p^*$. Using direct substitution it is straightforward to verify that $(c^*, y^*, w_2^*, \mu^*, p_2^*)$ is a solution to a Hamiltonian for (15).

A.3 Analysis

Here we proved some additional results we referred to in Section 2 and in the Appendix.

Lemma 3. *Suppose that Assumptions 1, 5, and 6 hold.*

(i) If either shocks θ are independent over time; or $U_{cl} = 0$; or $V_1 \geq 0$, then $\lambda \geq 0$;

(ii) If shocks to θ are independent over time; or assumption 4 is satisfied and $U_{cl} = 0$; or assumption 4 and $V_1 \geq 0$ and is increasing in θ, then $\mu \geq 0$.

Proof. We prove part (i) first. Substitute (33) into (32) (with independent shocks $f_2 = 0$).

$$\lambda \left(1 - \frac{f_2}{f}\zeta\right) f - \frac{\delta}{\beta} V_1 f = \mu'.$$

Integrate using the boundary condition $\mu(0) = 0$

$$\mu(\theta) = \int_0^\theta \left[\lambda \left(1 - \frac{f_2(\hat{\theta})}{f(\hat{\theta})}\zeta\right) - \frac{\delta}{\beta} V_1(\hat{\theta})\right] f(\hat{\theta}) d\hat{\theta}. \tag{45}$$

Use the other boundary condition $\mu(\infty) = 0$ to get

$$\lambda = \int_0^\infty \frac{\delta}{\beta} V_1(\hat{\theta}) f(\hat{\theta}) d\hat{\theta}, \tag{46}$$

where we used the fact that $\int_0^\infty f_2(\hat{\theta}) d\hat{\theta} = 0$. $V_1 \geq 0$ implies that $\lambda \geq 0$.

Alternatively use the first order condition for c, (31), to substitute into (32) and integrate

$$\mu(\theta) = \int_0^\theta \left(\lambda \left(1 - \frac{f_2(\hat{\theta})}{f(\hat{\theta})}\zeta\right) - \frac{1}{U_c(\hat{\theta})}\right) \exp\left[-\int_{\hat{\theta}}^\theta \frac{U_{c\theta}(\theta')}{U_c(\theta')} d\theta'\right] f(\hat{\theta}) d\hat{\theta}. \tag{47}$$

The boundary condition $\mu(\infty) = 0$ implies that

$$\lambda = \frac{\int_0^\infty \frac{1}{\tilde{U}_c(\hat{\theta})} \exp\left[-\int_{\hat{\theta}}^\infty \frac{U_{c\theta}(\theta')}{U_c(\theta')} d\theta'\right] f(\hat{\theta}) d\hat{\theta}}{\int_0^\infty \left(1 - \frac{f_2(\hat{\theta})}{f(\hat{\theta})}\zeta\right) \exp\left[-\int_{\hat{\theta}}^\infty \frac{U_{c\theta}(\theta')}{U_c(\theta')} d\theta'\right] f(\hat{\theta}) d\hat{\theta}}. \tag{48}$$

When shocks are independent, $f_2 = 0$ and $\lambda > 0$. When $U_{cl} = 0$, then $U_{c\theta} = 0$ and therefore this expression becomes

$$\lambda = \frac{\int_0^\infty \frac{1}{\tilde{U}_c(\hat{\theta})} f(\hat{\theta}) d\hat{\theta}}{\int_0^\infty \left(1 - \frac{f_2(\hat{\theta})}{f(\hat{\theta})}\zeta\right) f(\hat{\theta}) d\hat{\theta}} = \int_0^\infty \frac{1}{U_c(\hat{\theta})} f(\hat{\theta}) d\hat{\theta} > 0.$$

Next we turn to part (ii). First, suppose that shocks are independent. By Assumption 5 $w(\theta)$ is increasing and therefore by assumption 6 $V_1(\theta)$ is increasing in θ. Choose $\bar{\theta}$ s.t. $\lambda = \frac{\delta}{\beta} V_1(\bar{\theta})$. Therefore $\lambda - \frac{\delta}{\beta} V_1(\theta) \geq 0$ for all $\theta \leq \bar{\theta}$ and then (45) together with $f_2 = 0$ implies that $\mu(\theta) \geq 0$ for all $\theta \leq \bar{\theta}$. Now consider $\theta \geq \bar{\theta}$. Since $\mu(\infty) = 0$, we can write $\mu(\theta)$ as $\mu(\theta) = \int_\theta^\infty \left(\frac{\delta}{\beta} V_1(\hat{\theta}) - \lambda\right) f(\hat{\theta}) d\hat{\theta}$. The expression in the brackets is positive for all $\hat{\theta} \geq \bar{\theta}$, and therefore $\mu(\theta) \geq 0$ for $\theta \geq \bar{\theta}$.

Next, suppose that shocks are persistent, assumption 4 is satisfied and V_1 is increasing in θ. Consider period 1 first. In period 1 $\zeta_{t=1} = 0$ by the definition of the recursive problem, and equation (45) takes the form

$$\mu(\theta) = \int_0^\theta \left[\lambda G'(u(\theta)) - \frac{\delta}{\beta} V_1(\hat{\theta})\right] f(\hat{\theta}) d\hat{\theta}.$$

Since F_t exhibits first order stochastic dominance, $u(\theta)$ must be increasing in θ and hence $G'(u(\theta))$ decreases in θ. If V_1 is increasing in θ we can apply the same arguments as in i.i.d case to show that $\mu_{t=1} \geq 0$. Since $\mu_{t=1} \geq 0$, the first order condition for w_2, (34) implies that $V_2 \leq 0$ which implies from the equation (35) that $\zeta_{t=2} \geq 0$. Since f satisfies assumption 4, $\frac{f_{2,t=2}(\theta)}{f_{t=2}(\theta)} \zeta_{t=2} + \frac{\delta}{\beta} V_{1,t=2}(\theta)$ is increasing in θ. Then we choose $\bar{\theta}$ such that $\lambda = \frac{f_{2,t=2}(\bar{\theta})}{f_{t=2}(\bar{\theta})} \zeta_{t=2} + \frac{\delta}{\beta} V_{1,t=2}(\bar{\theta})$ and apply the arguments of the previous paragraph to show that $\mu_{t=2} \geq 0$. By iteration we then establish this argument for all t.

Finally, suppose that shocks are persistent, assumption 4 is satisfied, and $U_{cl} = 0$. In this case, by assumption 5, $1/U_c(\theta)$ is increasing in θ. Since $U_{c\theta} = 0$

we can apply the same arguments as in the previous paragraph to equation (47). □

We state the following lemma to conclude this part of the discussion.

Lemma 4. *Suppose assumptions 1, 5, 6 are satisfied. Then* $\tau^y \geq 0$ *if and only if* $\mu \geq 0$.

Lemma 5. *Suppose assumptions 1, 2, 5, and 6 are satisfied. Suppose that* $\mu \geq 0$. *Then* $\tau^w \geq \tau^y$.

Proof. Since $\mu \geq 0$ by Lemma 4, $\tau^y \geq 0$. Assumption 2 implies that $\varepsilon > 0$ and Assumption 1 that $U_{cl}/U_c \geq 0$, therefore $\tau^w \geq \tau^y$. □

A.4 Calibration, estimation, and computation

This section follows the general outline of the quantitative analysis section but fills in the details to provides a more complete description of the estimation and computation strategies we used.

Calibrated model. We model a life cycle of 40 years of working life, i.e., the individuals between the ages of 25 and 65. The parameters of the calibrated model are summarized in Table 1. Recall that we set the coefficient of absolute risk aversion, ψ, equal to 10. Numerical simulations are re-scaled so that consumption ranges from 0.1 to 1 implying that relative risk aversion ranges from 1 to 10 when $\psi = 10$. The elasticity parameter, γ, is set to 3 with the Frisch elasticity of labor supply then $\varepsilon = 1/(\gamma - 1) = 0.5$. The annual discount factor is $\beta = 0.9804$ and the marginal rate of transformation is $\delta = 1.02$ so that the social planner at the solution of the optimal program chooses not to transfer resources between periods.

Empirical strategy. Our main data source is the PSID. We use the data

Table 1: Parameters Of The Calibrated Model

Parameter	(explanation)	Value	Notes
ψ	(absolute risk aversion)	10	consumption ranges from 0.1 to 1, implying relative risk aversion ranges from 1 to 10
γ	(elasticity)	3	Frisch elasticity of labor supply 0.5
β	(discount factor)	0.9804	
δ	(marginal rate of transformation)	1.02	$1/\beta$, no transfer of resources between periods
a_1	(age at $t = 1$)	25	
a_T	(age at $t = T$)	65	model 40 years of work life

waves from 1990 onward to the latest currently available data wave, 2006. Table 2 presents summary statistics for the PSID waves we use. Recent waves (from 1996) come in two year intervals, hence we consider a total of nine waves with two years in between. To be consistent with the data, we let one period in the model correspond to two years, i.e., $T = 20$ since we model 40 year of working life.[10]

Labor income, $y_{i,year}$, is obtained directly from the PSID waves, converted to constant 1990 dollars, and is summarized in Table 2. We consider total labor income, which is a sum of a list of variables in the PSID that contain data on salaries and wages, separate bonuses, the labor portion of business income, overtime pay, tips, commissions, professional practice or trade payments, mar-

[10] We check that our results are robust when the number of periods is doubled to $T = 40$. When we take one period in the model to be 2 years, the discount factor is β^2 and the marginal rate of tranformation between periods is δ^2.

Table 2: Summary Statistics Of The PSID Waves

		Age		Labor Income		
Year	Number of Individuals	Mean	Standard Deviation	Mean	Standard Deviation	Maximum
1990	4718	38.0	11.4	26,668	26,380	550,000
1992	4936	38.4	11.5	27,411	28,869	759,259
1994	5312	38.7	11.2	27,922	31,537	789,503
1996	5437	39.0	11.3	27,820	29,843	581,612
1998	5785	39.5	11.6	29,405	35,284	1,140,000
2000	6162	39.6	11.9	30,828	38,883	869,699
2002	6362	40.1	12.1	30,832	51,943	2,536,232
2004	6346	40.4	12.5	31,332	54,235	2,054,795
2006	6490	40.6	12.8	31,081	45,965	2,051,282

Notes: The year entries correspond to the year of the data origin of PSID waves. Individuals are heads of households and their spouses or long-term cohabitants separately. Labor income is total labor income of an individual, e.g.,.in 2006 it is the sum of PSID variables ER40921 (which is in turn a sum of several variables) and ER40900.

ket gardening, additional job income, and other miscellaneous labor income. As Table 2 illustrates, mean real total labor income grows about 8.5% over the sixteen years considered or at about 0.5% per year. The variance of labor income increases by 0.3% over the same period.

Effective marginal tax rates, $T'_{year}\left(y_{i,year}\right)$, are estimated for each individual with TAXSIM. TAXSIM is a FORTRAN program of the National Bureau of Economic Research for estimating individual effective liabilities under U.S.

Table 3: Estimated Marginal Tax Rates and Implied Skills

	Estimated Effective Marginal Tax Rate		Implied Skill, θ	
Year	Mean	Standard Deviation	Mean	Standard Deviation
1990	21.7	8.7	2.0	1.2
1992	20.4	9.4	2.1	1.3
1994	22.7	11.1	2.2	1.4
1996	21.5	10.9	2.3	1.4
1998	21.5	11.0	2.4	1.6
2000	22.3	10.7	2.6	1.8
2002	20.9	10.9	2.7	2.0
2004	19.1	10.8	2.8	2.1
2006	19.2	10.9	2.9	2.1

Notes: Effective marginal tax rates are estimated using TAXSIM.

Implied skills are computed from individual optimality conditions as described in the text.

Federal and State income tax laws from individual data.[11] For each individual with labor income we also have in the PSID a collection of personal data that are in most cases sufficient to estimate individual and year specific effective marginal tax rates. Specifically, we input into TAXSIM for each individual their wage and salary income, wage and salary income of their spouse, dividend income, other property income (e.g., interest), taxable pensions and social security benefits, other transfers (e.g., welfare), unemployment compensation, whether the individual is older than 65, state of residence, marital status, number of dependents, and tax year. The estimated effective marginal tax rates for the individuals in the PSID waves are summarized in Table 3. Notably, mean of the effective marginal tax rate remains (except in 1992) close to 22% until 2000 and then falls to 20.9 in 2002, 19.1 in 2004, and 19.2 in 2006.

Finally, we use the constructed total individual labor incomes and the estimated effective marginal tax rates to compute implied skills as described above. Table 3 provides summaries for each of the PSID waves we consider. Thus we obtain a data set of implied skills based on empirical U.S. micro data with the details of our data set sample size provided in Table 4. When using PSID waves, we treat heads of households and their spouses or long-term cohabitants as separate observations. We first restrict the sample to include only individuals with the total labor income of at least $\$1,000$ in 1990 dollars and with at least 250 total hours worked in a year. Excluding individuals that do not have enough data in the PSID to estimate effective marginal tax rates with TAXSIM results in a sample of 50,624 individuals total from all waves. We also check our results for robustness with an alternative sample where individuals older than 65 are excluded. Considering only those with enough data for TAXSIM results in an alternative sample of just above

[11] For more details and to use the program freely see http://www.nber.org/~taxsim/.

Table 4: Sample Sizes

Sample Restriction	Sample Size
At least 250 hours worked, $1,000 income	51,548
- enough personal information for TAXSIM	50,624
Younger than 65	50,277
- enough personal information for TAXSIM	49,396
In at least two adjacent waves	
(i.e. at least one skill "transition")	27,664
- younger than 45	20,410
- 45 or older	7,254
Initial young	8,387
- enough personal information for TAXSIM	8,231

Table 5: Labor Income, Tax Rates, And Skills Of The Initial Young

Statistic	Mean	Standard Deviation
Labor Income	23,963	19,929
Estimated Effective Marginal Tax Rate	19.3	11.0
Implied Skill, θ	2.2	1.2

Notes: Labor income is total labor income of an individual from the PSID. Effective marginal tax rates are estimated using TAXSIM. Implied skills are computed from individual optimality conditions as described in the text.

forty nine thousand. In contrast, to estimate conditional distributions used in this section, we use the panel characteristic of the PSID – we consider all individuals that appear in at least two adjacent waves, that is the individuals for whom we observe at least one skill "transition".[12] We have a total of 27,664 skill transitions. To estimate the initial unconditional distribution of skills we consider a cross section of 25 year old from all waves to obtain a usable sample of 8,231. Summary statistics for this last subsample are provided in Table 5.

Estimation approach. As mentioned in the body of the paper, we start by estimating the initial unconditional distribution of implied skills among the initial young workers, $F_1(\theta)$. We consider the 25 year old from all of the PSID waves to obtain the sample of 8,231 observations described above and summarized in Table 5. We estimate $F_1(\theta)$ non-parametrically using a

[12] The PSID is not a balanced panel - an individual may appear in one wave, stay for one or more waves, and then disappear. Our data points are all of the separate individual wave-to-wave skill transitions.

kernel density estimation. We use the normal kernel function and $\mathbb{R}_+$ as the support. The resulting distribution is shown as "initial young, unconditional" distribution depicted by a dotted line in Figure 1.

To estimate conditional distributions $F_t(\theta|\theta\_)$, i.e., transition probabilities, we exploit the panel feature of the PSID. We start by considering all individual skill transitions between adjacent data waves to obtain 27,664 transitions as shown in Table 4. Furthermore, we break all these skill transitions into two age groups – when the individual is younger than 45 at the beginning of the transition and when the individual is 45 or older at the beginning of the transition. This gives samples of 20,410 "young" skill transitions and 7,254 "old" skill transitions. We therefore estimate two separate conditional distributions $F_{young}(\theta|\theta\_)$ and $F_{old}(\theta|\theta\_)$. Hence we assume age dependence between the age groups and age-independent transitions within each age group. In other words, we allow younger individuals to experience different transition probabilities than older individuals; within each age group, we assume age-independent transition probabilities.[13] We estimate each conditional distribution non-parametrically using a kernel density estimation with the normal kernel function and $\mathbb{R}_+$ as the support.

Computational strategy. To be able to numerically solve the problem of this size and complexity (i.e., with multitude of periods and correlated shocks) we exploit the recursive structure of the dual formulation of the planner's problem analyzed in Section 1. Hence we need to solve a finite horizon discrete time dynamic programming problem with a three-dimensional state vector

[13] We stop at just two age groups to have sufficient number of data points to estimate all conditional distributions. There is nothing in our computational solution method that would stop us from having a different transition matrix for each period, provided that we had enough data to obtain those transition matrices.

which is continuous in each dimension. We proceed in three stages.

The first stage is a value function iteration. We start from period T and proceed by backward induction. First, we solve period $t = T$ problem for a fixed set of values of the state vector and compute V_T for each of then. Then we can approximate V_T and proceed to period $t = T - 1$ where we use the approximation as the basis for the interpolation of V_T to any value of the state vector to solve for V_{T-1}. We continue until we compute V_1. Specifically, with the exponential preferences we can show that

$$V_t\left(\hat{w}, \hat{w}_2, \theta_-\right) = a_t(-\frac{\hat{w}_2}{\hat{w}}|\theta_-) - \frac{1 + \delta + ... + \delta^{T-t}}{\psi} \ln\left(-\hat{w}\right)$$

and in particular

$$V_T(\hat{w}, \hat{w}_2, \theta_-) = a_T\left(-\frac{\hat{w}_2}{\hat{w}}|\theta_-\right) - \frac{1}{\psi}\ln(-\hat{w}).$$

This means two things for our computations. First, if we discretize the type space Θ, we only need to consider $\hat{w}$ and $\frac{\hat{w}_2}{\hat{w}}$ as the state variables for each type. That is, our state space is discretized in skill dimension and is continuous in the other two dimensions. Second, we do not need to approximate V_t as a whole, rather we only need to approximate a_t, which significantly improves the quality of the approximation of V_t. We approximate a_t's using a shape-preserving least absolute deviation (LAD) method with Chebyshev polynomials. The evaluation nodes are chosen as the roots of Chebyshev polynomials.[14] The policy functions are similarly approximated at this stage. With our preferences

[14] For more on this, see e.g. Judd (1996) and Judd (1998).

it can be shown that

$$\begin{aligned}
c_t(\hat{w}, \hat{w}_2, \theta_-) &= a_t^c\left(-\frac{\hat{w}_2}{\hat{w}}|\theta_-\right) - \frac{1}{\psi}\ln(-\hat{w}) \\
y_t(\hat{w}, \hat{w}_2, \theta_-) &= a_t^y\left(-\frac{\hat{w}_2}{\hat{w}}|\theta_-\right) \\
w_t(\hat{w}, \hat{w}_2, \theta_-) &= a_t^w\left(-\frac{\hat{w}_2}{\hat{w}}|\theta_-\right)\ \hat{w} \\
w_{2t}(\hat{w}, \hat{w}_2, \theta_-) &= a_t^{w_2}\left(-\frac{\hat{w}_2}{\hat{w}}|\theta_-\right)\ \hat{w}.
\end{aligned}$$

Once again, we approximate a_t^c's, a_t^y's, a_t^w's, and $a_t^{w_2}$'s using a shape-preserving LAD method with Chebyshev polynomials and the evaluation nodes at Chebyshev roots.

To compute the full constrained optimal allocation, we need to find w_0 such that $V_1(w_0) = 0$. This is the second stage. Given V_1 computed in the first stage, we search for an interval containing zero using binary jumps. Then we converge to w_0 with bisection (binary search).[15]

The third stage is to compute optimal labor and savings distortions. Since policy functions were approximated during the first stage, given V_t's and w_0 from the first two stages, we can now generate the optimal allocations by forward induction. We start with w_0 computed in the second stage and roll out the solution from period $t = 1$ all the way to period $t = T$. Optimal labor and savings distortions are computed from their definitions in equations (17) and (18) respectively.

Finally, we verify ex post that the first-order approach is valid. We verify within small numerical error bounds the sufficient conditions discussed above in the context of optimization problem (9) and with formal arguments in the Appendix. In particular, in addition to preferences described by the utility

[15] The rate of convergence for bisection is, admittedly, only linear, however, what is important here is guaranteed convergence.

function U that satisfies single crossing property, we numerically verify that output satisfies $y'(\theta) \geq 0$ and promised utility satisfies $\omega_1\left(\hat{\theta}|\theta\right) \geq 0$ and $\omega_{12}\left(\hat{\theta}|\theta\right) \geq 0$. These conditions are straightforward to verify numerically within the value function iteration.[16]

For this three-stage computational procedure to be feasible it is essential to have an efficient and robust optimization algorithm to solve all of the separate period t mechanism design problems of each stage at each node.[17] We solve each problem using an implementation of the interior-point optimization algorithm with conjugate gradient iteration to compute the optimization step.[18] Conjugate gradient iteration offers a way of dealing with possible Jacobian and Hessian singularities. The interior-point approach is one of the most efficient and stable methods that are currently available for solving large nonlinear optimization problems. The interior-point algorithm uses a trust-region Newton method to solve the barrier problem and an l_1 penalty barrier function. We find that the interior-point algorithm provides a good approximate estimate of the solution and the optimal set of active constraints. To compute accur-

[16] As an additional check, we also verified sufficiency using the "brute force" approach: once the solution is found, we check that there are no global deviations.

[17] The main reason is that a mechanism design problem is a bi-level maximization problem (alternatively, a mathematical programming problem with equilibrium constraints). The outer-level maximization of the planner has to take into account the best response of the agents, which is the outcome of the inner-level maximization of each agent type with respect to the type reported. In other words, incentive constraints are individual agent type maximization problems with type report as a choice variable. We follow the usual convention of computationally approaching these types of problems (e.g. Judd (1998)) by writing the incentive constraints as inequalities (without relying on simplifying the incentive compatibility constraints with the envelope theorem) as in problem (3).

[18] The implementation we use is KNITRO. To streamline the application of KNITRO we use a modelling language AMPL.

ate estimates of the solution, including Lagrange multipliers, we proceed to switch to an active-set iteration that uses the output of the interior-point algorithm as its input. The implementation of the active-set algorithm is based on the sequential linear quadratic programming. Once the problem is correctly scaled, we observe quadratic convergence to a local maximum. Our globalization strategy is to explore multiple feasible starting points.