Quantifying the Inefficiency of the US Social Security System

Mark Huggett and Juan Carlos Parra*

January 18, 2005

Abstract

We quantify the inefficiency of the retirement component of the US social security system within a model where agents receive idiosyncratic, labor-productivity shocks that are privately observed.

JEL Classification: D80, D90, E21

Keywords: Social Security, Idiosyncratic Shocks, Efficient Allocations, Private Information

*
Affiliation: Georgetown University
Address: Economics Department; Georgetown University; Washington DC 20057- 1036
E-mail: mh5@georgetown.edu and jcp29@georgetown.edu
Homepage: http://www.georgetown.edu/faculty/mh5 and www.georgetown.edu/users/jcp29
Phone: (202) 687- 6683
Fax: (202) 687- 6102

1 Introduction

One rationale for a social security system is the provision of social insurance for risks that are not easily insured in private markets. In fact, the Economic Report of the President (2004, Ch. 6, p.130) claims that the provision of social insurance for labor income risk over the life cycle is one of the main problems that justifies a government role in old-age entitlement programs. They claim that labor income is risky but difficult to insure. One reason given for why insurance is dfficult is that labor income is partly under an individual's control by the choice of (unobserved) effort or labor hours. They then claim that social security provides partial insurance through a progressive retirement benefit based on lifetime earnings. Given this argument, we find it natural to ask how well or how poorly does the retirement component of the US social security system serve this insurance role? Thus, the goal of this paper is to quantify how far a stylized version of the US social security system is from an efficient system.

Quantifying the inefficiency of social security systems is difficult. One reason is that there are many sources of risk to consider and social security systems have distinct benefits tailored to these risks. A second reason is that there are other mechanisms, such as income taxation, that potentially are important in the provision of insurance. Thus, analyzing the inefficiency of social security quickly becomes an analysis of the inefficiency of the tax-transfer system as a whole.

This paper provides a simple benchmark analysis. This simplification is gained by (i) analyzing one component of the US social security system, the retirement component, in isolation, (ii) considering social security together with income taxation to be the entire tax-transfer system and (iii) considering a single but very important source of risk. The risk that is examined here is idiosyncratic, labor-productivity risk. We focus on this risk for two reasons. First, individual workers experience substantial variation in wage rates which are not related to systematic life-cycle variation or to aggregate fluctuations.[1] Second, this risk is a natural way to model labor income as risky but difficult to insure.

The degree of inefficiency of the US social security system together with the income tax system is determined by comparing an agent's ex-ante, expected utility in the model of the US economy to the ex-ante, expected utility that a planner could achieve for the agent. In the model of the US economy it is assumed that there is a risk-free asset for transfering resources over time and that social security together with income taxation are the only means for transfering resources across states (i.e across an agent's

[1]Heathcote, Storresletten and Violante (2004) examine annual wage rate data for US males. They divide (log) wages into components capturing life-cycle, business-cycle and idiosyncratic wage variation. They further divide the idiosyncratic component into subcomponents and find substantial variation in each subcomponent.

labor-productivity histories). The planner faces two constraints. Allocations must use no more resources than are used in the US system and must be incentive compatible. The incentive problem arises from the fact that the planner only observes an agent's earnings. Earnings equal the product of labor productivity and labor hours. Thus, the planner does not know whether earnings of an agent are low because labor productivity is low or because labor hours are low. Since labor productivity is privately observed by the agent, the Revelation Principle implies that the allocations between an agent and a planner that can be acheived are precisely those that are incentive compatible.

It is useful to briefly describe some features of the retirement component of US social security system. This will be helpful for understanding sources of potential inefficiencies. Consider a one-person household. In the US system the earnings of this person are taxed at a fixed tax rate up to a yearly maximum earnings level. The marginal tax rate is zero beyond this maximum. Benefits are in the form of a retirement annuity payment received after a retirement age. The size of the retirement payment is determined by a benefit formula which is an increasing and concave function of a measure of an individual's average earnings over the lifetime.

We first focus on inefficiency in the absence of idiosyncratic, labor-productivity risk. The inefficiency of the US tax-transfer system is computed within the model by proportionally adjusting consumption in the allocation produced by the US system so that, labor held fixed, ex-ante expected utility is equal to the expected utility that a planner could deliver. Thus, this calculation states in consumption terms the utility gain of moving to the utility possibility frontier, holding fixed the expected resources paid to the planner (i.e. the planners utility). Absent risk and absent income taxation, the inefficiency of the US system in the benchmark model is xx percent of consumption per year. The inefficiency increases to yy percent when income taxation and social security are combined together.

This result fits well with standard intuition (e.g. Feldstein (1996, p. 4)). The intuition is that inefficiency is related to the magnitude of distortions. Much of the emphasis has been focused on the distortion to the consumption-labor margin. When social security is analyzed "on top of" the income tax system, there is already a substantial distortion due to positive marginal income tax rates. In the benchmark model we calculate that the present value of marginal social security benefits incurred for extra work is below the value of marginal taxes paid at all ages. Thus, the marginal rate of substitution between consumption and labor for an optimizing agent is depressed below the agent's marginal rate of transformation (i.e. the agent's labor productivity) both by income taxation and by social security. In contrast, in an efficient allocation, marginal rates of substitution equal marginal rates of transformation.

How do the results change when permanent, labor-productivity risk is added? Here the abstraction is that wage rates across agents differ in each period over the life cycle

as some agents are born with higher productivity than others. When these permanent differences are set to the values estimated in US data, we find that the inefficiency of the model economy is xx percent in the absence of income taxation and is yy percent when social security and income taxation are combined. These results raise two questions: (1) Why is inefficiency so much larger when labor-productivity risk is present? and (2) Why does inefficiency now fall when social security and income taxation are jointly considered compared to when social security is analyzed in the absence of income taxation?

The paper is organized as follows. Section 2 briefly discusses related literature. Section 3 presents the social security decision problem and the optimal planning problem. Section 4 sets model parameters. Section 5 presents results.

2 Related literature

This paper builds upon the social security and optimal contract theory literatures. We highlight the papers in these literatures which are most closely related to our work.

To address the role of social security in the provision of social insurance, one needs a model with some risk that is not easily insured in private markets. Imrohoroglu et al (1995), De Nardi et al (1999), Huggett and Ventura (1999) and Storesletten et al (1999) were among the earliest papers to provide a quantitative analysis of social security systems in the presence of idiosyncratic earnings risk. This paper shares much in common with these papers in that it uses computational methods to calculate allocations and adopts the modeling of the US social security system used in Huggett and Ventura (1999).[2]

Our work is also related to the efficiency gains literature. This literature determines whether or not specific policy changes produce Pareto improvements and calculates the magnitude of efficiency gains. For example, the classic work by Auerbach and Kotlikoff (1987, Ch. 10) computes efficiency gains from more closely linking marginal social security benefits to marginal social security taxes in a model which abstracts from aggregate and idiosyncratic risk. Our work computes efficiency gains in a model with idiosyncratic, labor-productivity risk that is privately observed. Infact, we compute the maximum efficiency gain, which we label the "inefficiency" of the social insurance system. Relatively few papers in the efficiency gains literature calculate how far social insurance systems are from efficient allocations.[3]

[2]Imrohoroglu et al (2000) survey the social security literature that emphasizes idiosyncratic earnings risk.

[3]Lindbeck and Persson (2003) review the literature on efficiency gains and social security reform. We mention three papers from this literature which differ in the risk analyzed. Hubbard and Judd (1987) determine whether social security improves upon no social security system when there is mortality risk and

This paper also builds upon the optimal contract theory literature that emphasizes privately-observed, labor-productivity risk. This literature began with Mirrlees (1971). Diamond and Mirrlees (1978, 1986) extended this framework to consider the optimal disability insurance problem. In this problem all agents are identical and able to work until hit with a privately-observed, disability shock rendering an agent permanently unable to work. Golosov and Tsyvinski (2004) reconsider the optimal disability problem. They quantify the efficiency gains to adopting an optimal disability insurance system instead of a stylized version of the US system. Our paper differs since we focus on the retirement benefit.[4]

This paper is also related to work in dynamic contract theory, such as Green (1987), Spear and Srivastava (1987), Thomas and Worrall (1990), Atkeson and Lucas (1992) and Fernandes and Phelan (2000). In this work recursive methods are used to characterize and to compute solutions to dynamic contracting problems. An important issue is the nature of tax-transfer systems that implement solutions to dynamic contracting problems with labor-productivity risk. Golosov and Tsyvinski (2004), Albanesi and Sleet (2003), Kocherlakota (2003) and Battaglini and Coate (2004) present some results on this problem.

3 Framework

3.1 Preferences

An agent's preferences over consumption and labor allocations over the life cycle are given by a calculation of ex-ante, expected utility.

$$E[\sum_{j=1}^{J} \beta^{j-1} u(c_j, l_j)] = \sum_{j=1}^{J} \sum_{s^j \in S^j} \beta^{j-1} u(c_j(s^j), l_j(s^j)) P(s^j)$$

Consumption and labor allocations are denoted $(c, l) = (c_1, ..., c_J, l_1, ..., l_J)$. Consumption and labor at age j are functions c_j and l_j mapping j-period shock histories $s^j \equiv (s_1, ..., s_j)$ into consumption and labor decisions in period j. An agent's labor productivity in period j, or equivalently at age j, is given by a function $\omega(s_j, j)$ mapping the period shock s_j and the agent's age j into labor productivity. Consumption is non-negative and labor lies in the interval $[0, 1]$. The set of possible j-period histories

private markets do not provide annuities. Krueger and Kubler (2003) determine whether there are efficiency gains to adopting a pay-as-you-go social security system in place of private pensions when there is aggregate productivity risk. Nishiyama and Smetters (2004) ask whether there are efficiency gains in moving from the US system to an individual accounts system when agents face idiosyncratic wage risk.

[4]Diamond (2003) relates work in optimal contract theory to the design of social security systems.

is denoted $S^j = \{s^j = (s_1, ..., s_j) : s_i \in S, i = 1, ..., j\}$, where S is a finite set of shocks. $P(s^j)$ is the probability of history s^j.

3.2 Incentive Compatibility

It is assumed that labor productivity is observed only by the agent. The principal observes the output of the agent which equals the product of labor productivity and work time. In this context, the Revelation Principle (see Mas-Colell et al (1995, Prop. 23.C.1)) implies that the allocations (c, l) that can be achieved between a principal and an agent are precisely those that are incentive compatible.

We now define what it means for an allocation to be incentive compatible. For this purpose, we define the report function $\sigma \equiv (\sigma_1, ..., \sigma_J)$, which is composed of period report functions σ_j that map shock histories $s^j \in S^j$ into S. The truthful report function is denoted σ^* and has the property that $\sigma_j^*(s^j) = s_j$ in any period for any j-period history. An allocation (c, l) is *incentive compatible* (IC) provided that the truthful report function always gives at least as much expected utility to the agent as any other feasible report function. The expected utility of an allocation (c, l) under a report function σ is denoted $W(c, l; \sigma, s_1)$. This is defined below, where $\hat{s}^j \equiv (\sigma_1(s^1), ..., \sigma_j(s^j))$ denotes the j-period reported history when the true history is s^j. Using this notation, (c, l) is IC provided $W(c, l; \sigma^*, s_1) \geq W(c, l; \sigma, s_1), \forall s_1, \forall \sigma$. A report function σ is feasible for an allocation (c, l) provided that in any period in any history an agent's true labor productivity $\omega(s_j, j)$ is always large enough to produce the output required by a report (i.e. $0 \leq l_j(\hat{s}^j)\omega(\sigma_j(s^j), j) \leq \omega(s_j, j), \forall j, \forall s^j$).

$$W(c, l; \sigma, s_1) \equiv \sum_j \sum_{s^j \in S^j} \beta^{j-1} u(c_j(\hat{s}^j), \frac{l_j(\hat{s}^j)\omega(\sigma_j(s^j), j)}{\omega(s_j, j)}) P(s^j | s_1)$$

3.3 Decision Problems

This paper focuses on two decision problems: the social security (SS) problem and the private information planning problem (PP). These problems have the same objective but different constraint sets. V_{SS} and V_{PP} denote the maximum ex-ante, expected utility achieved in these problems.

$V_{PP} \equiv \max E[\sum_j \beta^{j-1} u(c_j, l_j)]$ subject to $(c, l) \in \Gamma_{PP}$

$\Gamma_{PP} = \{(c, l) : E[\sum_j \frac{(c_j - \omega(s_j, j) l_j)}{(1+r)^{j-1}}] \leq Cost$ and (c, l) is IC $\}$

$V_{SS} \equiv \max E[\sum_j \beta^{j-1} u(c_j, l_j)]$ subject to $(c, l) \in \Gamma_{SS}$

$$\Gamma_{SS} = \{(c,l) : \textstyle\sum_j \frac{c_j}{(1+r)^{j-1}} \leq \sum_j \frac{(\omega(s_j,j)l_j - T_j(x_j,\omega(s_j,j)l_j))}{(1+r)^{j-1}}, \forall s^J \in S^J$$
$$x_{j+1} = F_j(x_j, \omega(s_j,j)l_j, c_j), x_1 \equiv 0\}$$

The constraint set Γ_{PP} for the planning problem has two restrictions. First, the expected present value of consumption less labor income cannot exceed some specified value, denoted $Cost$. Present values are computed with respect to an exogenous real interest rate r. Second, allocations (c,l) must be incentive compatible (IC).

The constraint set Γ_{SS} for the social security problem is specified by a tax function T_j and a law of motion F_j for a vector of state variables x_j. The tax function states the agent's tax payment at age j as a function of period earnings $s_j l_j$ and the state variables x_j. A negative tax is a transfer. The social security problem requires that the present value of consumption is no more than the present value of labor earnings less net taxes for any labor-productivity history.[5] The next section demonstrates that this abstract formulation is able to capture features of the US social security and income tax system.

Ex-ante expected utility can be ordered in these problems so that $V_{PP} \geq V_{SS}$. This occurs when $Cost$ in the planning problem is selected to equal the expected present value of taxes incurred in a solution (c^*, l^*) to the social security problem (i.e. $Cost \equiv E[\sum_j -T_j(x_j, \omega(s_j,j)l_j^*)/(1+r)^{j-1}]$). The argument is based on showing that if the allocation (c^*, l^*) achieves the maximum in the social security problem, then (c^*, l^*) is also in Γ_{PP}. Since (c^*, l^*) satisfies the present value condition in Γ_{SS}, then it also satisfies the expected present value condition in Γ_{PP}. Thus, it remains to argue that (c^*, l^*) is incentive compatible. However, the Revelation Principle implies that if (c^*, l^*) is the best choice of the agent under social security then it necessarily is incentive compatible.

To conclude this section, we raise two issues concerning how to interpret solutions to the planning problem. First, is a solution to the planning problem a Pareto efficient allocation? Solutions to the planning problem are Pareto efficient allocations between a risk-averse agent and a risk-neutral principal with discount factor $1/(1+r)$ when the utility possibility frontier is downward sloping. It is straightforward to show that the frontier is downward sloping when the agent's period utility function $u(c_j, l_j)$ is additively separable and is continuous and strictly increasing in consumption. Second, does a solution to the planning problem also solve the problem of maximizing ex-ante, expected utility of a large cohort of ex-ante identical agents subject to incentive compatibility and to the requirement that the realized present value cost to the planner not exceed some prespecified level? The assumption here is that agents experience id-

[5]The budget set can equivalently be formulated as a sequence of budget restrictions where the agent has access to a risk-free asset, starts life with zero units of this asset and must end life with non-negative asset holding.

iosyncratic but not aggregate risk. The contract theory literature mentioned in section 2 imposes the requirement that a present value condition or a market clearing condition must hold in equilibrium but not necessarily for any conceivable (non-equilibrium) reports that agents could make.[6] Under this requirement, a solution to the planning problem is a solution to the planning problem with a large cohort of ex-ante identical agents.

3.4 US Tax-Transfer System

The tax function and law of motion (T_j, F_j) are now specified to capture features of the US social security system together with the US federal income tax system. Specifically, the tax function T_j is the sum of social security taxes T_j^{ss} and income taxes T_j^{inc}. Note that the state variable $x_j = (x_j^1, x_j^2)$ has two components.

$$T_j(x_j, \omega(s_j, j)l_j) = T_j^{ss}(x_j^1, \omega(s_j, j)l_j) + T_j^{inc}(x_j^1, x_j^2, \omega(s_j, j)l_j)$$

3.4.1 Social Security

The model social security system taxes an agent's labor income before a retirement age R and pays a social security transfer after the retirement age. Specifically, taxes are proportional to labor earnings $(\omega(s_j, j)l_j)$ for earnings up to a maximum taxable level e_{max}. The social security tax rate is denoted by τ. Earnings beyond the maximum taxable level are not taxed. After the retirement age, a transfer $b(x^1)$ is given that is a fixed function of an accounting variable x^1. The accounting variable is an equally-weighted average of earnings before the retirement age R (i.e. $x_{j+1}^1 = [\min(\omega(s_j, j)l_j, e_{max}) + (j-1)x_j^1]/j$). The earnings that enter into the calculation of x_j^1 are capped at a maximum level e_{max}. After retirement, the accounting variable remains constant at its value at retirement.

$$T_j^{ss}(x_j^1, \omega(s_j, j)l_j) = \begin{cases} \tau \min(\omega(s_j, j)l_j, e_{max}) & : \quad j < R \\ -b(x_j^1) & : \quad j \geq R \end{cases}$$

The relationship between average past earnings x^1 and social security benefits $b(x^1)$ in the model is shown in Figure 1. Benefits are a piecewise-linear function of average past earnings. Both average past earnings and benefits are normalized in Figure 1 so that they are measured as a multiple of average earnings in the economy. The first segment of the benefit function in Figure 1 has a slope of .90, whereas the second and

[6] See Mas-Colell and Vives (1991) for a discussion of this issue and for results on implementation in exchange economies with a continuum of agents.

third segments have slopes equal to .32 and .15. Thus, the benefit function bends over. The 'bend-points' in Figure 1 occur at 0.21 and 1.29 times average earnings in the economy. The variable e_{max} is set equal to 2.42 times average earnings. The bend-points and the maximum earnings e_{max} are set at the actual multiples of mean earnings used in the US social security system. The slopes of the benefit function are also set to those in the US social security system.[7]

[Insert Figure 1 Here]

The specification of the model social security system captures many features of the old-age component of the US social security system. Two differences are the following:

(i) The accounting variable in the actual US system is an average of the 35 highest earnings years, where the yearly earnings measure which is used to calculate the average is capped at a maximum earnings level.[8] In the model, earnings are capped at a maximum level just as in the US system, but earnings in all pre-retirement years are used to calculate average earnings.

(ii) In the actual US system the age at which benefits begin can be selected within some limits with corresponding "actuarial" adjustments to benefits. In the model the age R at which retirement benefits are received is fixed.

3.4.2 Income Taxation

Income taxes in the model economy are determined by applying an income tax function to a measure of an agent's income. The empirical tax literature has calculated effective average tax rates (i.e. the empirical relationship between taxes actually paid divided by economic income).[9] We use tabulations from the Congressional Budget Office (2004, Table 3A and Table 4A) for the 2001 tax year to specify the relation between average effective individual income tax rates and income. Figure 2 shows average effective tax

[7]Under the US Social Security system, a person's monthly retirement benefit (i.e. the primary insurance amount) is based on a person's averaged indexed monthly earnings (AIME). For a person retiring in 2002 this benefit equals 90% of the first $592 of AIME, plus 32% of AIME between $592 and $3567, plus 15% of AIME over $3567. Dividing these "bend points" by average earnings in 2002 and multiplying by 12 gives the bend points in Figure 1. The bend points change each year based on changes in average earnings. The maximum taxable earnings from 1998- 2002 averaged 2.42 times average earnings. All these facts, as well as average earnings data, come from the Social Security Handbook (2003). The retirement benefit above is for a single-person household. The US system offers a spousal benefit that we abstract from.

[8]The 35 highest years are calculated on an indexed basis in that indexed earnings in a given year equal actual nominal earnings multiplied by an index. The index equals the ratio of mean earnings in the economy when the individual turns 60 to mean earnings in the economy in the given year. In effect, this adjusts nominal earnings for inflation and real earnings growth.

[9]See, for example, Gouveia and Strauss (1994).

rates on income in the US for households whose head is 65 or older or is younger than 65. The horizontal axis in Figure 2 expresses income in multiples of average individual earnings in the US for the year 200x (correct?). Figure 2 shows that average tax rates increase strongly in income.

In the model economy, we base income taxes $T_j^{inc}(x_j^1, x_j^2, \omega(s_j, j)l_j)$ before and after the retirement age R on the average tax rates in Figure 2. Specifically, income is defined as the sum of labor income $\omega(s_j, j)l_j$, asset income $x_j^2 r$ and social security transfer income $b_j(x_j^1)$. Asset income is calculated as follows: $x_{j+1}^2 = \omega(s_j, j)l_j + x_j^2(1 + r) + -T_j(x_j^1, x_j^2, \omega(s_j, j)l_j) - c_j$.

[Insert Figure 2 Here]

4 Parameter Values

The benchmark results of the paper are based on the parameter values in Table 1. There are $J = 61$ model periods in an agent's lifetime. This corresponds to real-life ages 20 to 80. The retirement age (i.e. age at which retirement benefits are received) occurs in model period $R = 46$ which corresponds to a real-life retirement age of 65. This is the current age at which full benefits are received in the US system. The social security tax rate τ is set to equal 10.6 percent of earnings. This is the combined employee-employer tax for the US old-age and survivor's insurance benefit. The social security benefit function $b(x)$ and the income tax function T_j^{inc} are are given by Figure 1 and Figure 2. The previous section discussed how these functions were selected.

An agent's labor productivity is given by a function $\omega(s_j, j) = \mu_j s_j$. The term μ_j captures the systematic variation in mean labor productivity with age. We set μ_j equal to the US cross-sectional, mean-wage profile for males from Heathcote et al (2004). This is displayed in Figure 3, where we normalize μ_1 to equal 1. We have imposed that the mean productivity profile is zero at a real-life age of 65. Thus, an agent is not able to work at age 65 or afterwards. The term s_j captures idiosyncratic variation in labor productivity. We consider two possibilities for the stochastic structure of shocks: perfectly permanent shocks and purely temporary shocks. In the case of permanent shocks, an agent is "born" at age $j = 1$ with a realization of the permanent shock which remains with the agent over the life cycle. The agent receives no subsequent shocks. In the case of temporary shocks, an agent draws a shock each period independently from a fixed distribution. In both cases the distribution of shocks is a discrete approximation to a lognormal distribution (i.e. $\log(s_j) \sim N(-\sigma^2/2, \sigma^2)$).[10]

[10] We approximate the lognormal distribution with 5 equally-spaced points in logs in the interval $[-3\sigma, 3\sigma]$.

[Insert Figure 3 Here]

Table 1: Parameter Values

Definition	Symbol	Value
Model Periods	J	$J = 61$
Retirement Period	R	$R = 46$
Social Security Tax	τ	$\tau = .106$
Benefit Function	$b(x)$	Figure 1
Income Tax Function	T^{inc}	Figure 2
Labor Productivity	$\omega(s_j, j)$	$\omega(s_j, j) = \mu_j s_j$ $\log(s_j) \sim N(-\sigma^2/2, \sigma^2)$
Mean Productivity Profile	μ_j	Figure 3
Interest Rate	r	$r = 0.042$
Discount Factor	β	$\beta = 1.0/(1 + r)$
Preferences	$u(c, l)$	$\frac{c^{(1-\rho)}}{(1-\rho)} + \phi\frac{(1-l)}{(1-\gamma)}^{(1-\gamma)}$ $\rho = 1, \gamma = 3.1955$ ϕ see text

Heathcote et al (2004) have decomposed the idiosyncratic component of variation of log wages of US males into the sum of permanent, persistent and purely temporary components. They estimate that the variance of the perfectly temporary component of log wage shocks is $\sigma^2 = 0.074$ and that the variance of the permanent component of log wage shocks is $\sigma^2 = 0.109$.[11] These estimates will lie in the range of the variances σ^2 for temporary and permanent shocks that we consider in the next section.

Probabilities are set to the area under the normal distribution, where midpoints between the approximating points define the limits of integration. This follows Tauchen (1986).

[11] The estimates cited in the text are the average values of the variances of the respective shock components. These values come from Heathcote et al (2004, Table 2) after weighting the variance in 1967 by the average factor loadings from 1967-1996.

One important restriction on the utility function $u(c, l)$ is the assumption of additive separability. Most of the theoretical literature on dynamic contract theory with a labor decision referenced in section 2 is based on this assumption. We make use of this assumption when we design a procedure to compute solutions to the planning problem.[12] The discount factor β and the real interest rate r are set so that $\beta(1+r) = 1$. Under these assumptions on preferences and the discount factor, the consumption profile over the life cycle is flat in a solution to the planning problem, when there is no labor-productivity risk. We set the real interest rate equal to 4.2 percent. This is the average real return over the period 1946- 2001 to an equally-weighted portfolio of stock and long-term bonds (see Siegel (2002, Tables 1-1 and 1-2)).

In the benchmark model, we set $u(c, l) = c^{(1-\rho)}/(1-\rho) + \phi\frac{(1-l)}{(1-\gamma)}^{(1-\gamma)}$. This choice implies a constant elasticity of intertemporal substitution of consumption equal to $\epsilon = -1/\rho$ and a constant Frisch elasticity of leisure with respect to the wage equal to $\epsilon_{leisure} = -1/\gamma$.

We now discuss how we set the parameters ρ and γ. We make use of estimates based on micro data and the assumption that the period utility function for consumption and labor is additively separable. The estimates of ϵ surveyed in Browning et al (1999, Table 3.1) range from -0.25 to -1.56. This would suggest a coefficient of relative risk aversion ρ ranging from below 1.0 to 4.0. In the benchmark model we set $\rho = 1$ (i.e. $u(c) = \log(c)$) and later examine the sensitivity of the results to higher values. On the labor side, the literature has focused on estimating the Frisch elasticity of labor supply (see Browning et al (1999, Table 3.3)). For the preferences under consideration, the Frisch elasticities of labor and leisure are related as follows: $\epsilon_{labor} = -\epsilon_{leisure}(1-l)/l$. We set the parameter γ to match an estimate of the Frisch elasticity of male labor supply. Domeij and Floden (2004, Table 6) estimate that $\epsilon_{labor} = 0.49$, using annual data for US males.[13] We choose $\gamma = 3.1955$ to match this estimate of the labor elasticity when labor l in the model equals the average fraction of time worked in the US.[14] The remaining parameter ϕ is set so that, given all other model parameters,

[12]It is used in Theorem A.3 in the Appendix to establish which incentive constraints bind and to develop a two-stage approach to solve the recursive-dual problem. The algorithm to solve the social security problem does not make use of additive separability.

[13]They show, within a model, that this elasticity is biased downward when agents are at or close to their borrowing limits and when standard empirical procedures are employed. Using US data, they find that the estimated elasticity is larger when the data set excludes households with small amounts of liquid assets. The estimate in the text is for households with liquid assets equal to at least one months wages. This estimate is higher than many in the literature but still within the range of estimates surveyed by Browning et al (1999, Table 3.3).

[14]The average fraction of time worked in the US is 0.383. This equals average hours worked divided by available work time. Average hours worked comes from Heathcote et al (2004, Table 1). Available work time

the average fraction of time worked in the model equals the average value in the US economy. When the variances for the permanent and temporary shocks are set to the point estimates discussed above, the value ϕ equals 0.6085 for the permanent shock case and 0.5560 for the temporary shock case.

5 Results

This section quantifies the inefficiency of the US system when labor-productivity shocks are temporary or permanent. The magnitude of inefficiency is the percentage increase α in consumption in the allocation (c^{ss}, l^{ss}) for the social security problem so that ex-ante expected utility is the same as in the private information planning problem, holding the expected present value of resources equal in both problems. This calculation is shown below, where superscripts denote the respective allocations. The results of this section are based on computing solutions to the social security problem and the planning problem. Our computational methods are described in detail in the Appendix.

$$E[\sum_j \beta^{j-1} u(c_j^{ss}(1+\alpha), l_j^{ss})] = E[\sum_j \beta^{j-1} u(c_j^{pp}, l_j^{pp})]$$

5.1 Assessment of Inefficiency

Figure 4 highlights the inefficiency of the US social insurance system in the benchmark model for a range of values for the variance of log labor-productivity shocks. Figure 4 shows that the measure of inefficiency is increasing in the variance of the shocks. To quantify the size of the inefficiency of the US system, one would need an estimate of the variance of the shocks to log wages. As described in the previous section, Heathcote et al (2004) estimate that $\sigma^2 = 0.074$ for the variance of temporary shocks and that $\sigma^2 = 0.109$ for the variance of permanent shocks. Using these estimates, the inefficiency of the US system is about 5.0 percent of consumption in the permanent shock case and 0.25 percent in the temporary shock case. One striking feature of these results is that the inefficiency of the US system is more than 40 times greater with idiosyncratic risk compared to no risk when shocks are permanent.

[Insert Figure 4 (a)-(b) Here]

equals 16 hours per day times 365 days per year.

5.2 Sources of Inefficiency

We now attempt to gain some insight into what lies behind the results presented in Figure 4.

5.2.1 No Idiosyncratic Risk

We first focus on understanding the source of the inefficiency in the US system in the absence of labor-productivity risk. This addresses the location of the intercept in Figure 4, which equals 0.12 percent for the permanent and temporary shock cases. To understand the source of the inefficiency, we compute labor profiles for the US system and for the efficient allocation with the same resources. As Figure 5 shows, both labor profiles are hump-shaped. In the efficient allocation marginal rates of substitution and transformation are equated. Given additive separability (i.e. $u(c,l) = u(c)+v(l)$), these conditions can be rewritten as follows: $u'(c_j) = \beta(1+r)u'(c_{j+1})$ and $-v'(l_j)/u'(c_j) = \omega(s_j, j)$. The first condition and $\beta(1+r) = 1$ implies that the consumption profile is flat. These conditions together imply that the efficient labor profile is hump-shaped because the agent's labor productivity profile is hump-shaped over the life cycle.

[Insert Figure 6 Here]

Why is the labor profile under the US system rotated counter-clockwise compared to the efficient profile? To answer this question, first note that in the model an extra unit of earnings when young or when old increases mean lifetime earnings by the same amount and thus has the same effect on increasing the retirement benefit. This follows directly from the social security benefit function described in section 3.4. However, the present value of these marginal benefits is substantially smaller when an agent is young compared to when old as the interest rate r is positive. Since the social security tax rate τ is constant, the amount of taxes paid for an extra unit of earnings is constant. The implicit marginal social security tax rate equals one minus the ratio of marginal benefits to marginal taxes. Thus, this implicit marginal tax rate decreases as an agent ages. Figure 7 graphs the marginal social security tax rate.[15]

[Figure 7: Marginal Social Security Tax Rate]

An optimizing agent who faces this social security system but no income taxes will equate the marginal rate of substitution between consumption and labor to the after-tax wage which equals the product of labor productivity $\omega(s_j, j)$ and one minus the marginal tax rate. Thus, the counter-clockwise rotation of the labor profile is due to the fall in the tax rate with age. This effect accounts for the inefficiency of social

[15]Clearly, introducing other features of the US social security system (e.g. teh spousal benefit or the fact that benefits are based on the 35 highest earnings years) would affect the tax rate in Figure 7.

security without income taxation. Income taxation acts to depress this marginal rate of substitution even further as well as to distort the intertemporal marginal rate of substitution of consumption.

5.2.2 Idiosyncratic Risk

It is natural to conjecture that when there is idiosyncratic risk a key reason for inefficiency is that the US system and efficient allocations differ strongly in the provision of insurance. To investigate this conjecture, we compute lifetime average net-tax rates for different realizations of lifetime labor-productivity histories. The net-tax rate is computed as the present value of earnings less consumption, expressed as a fraction of the present value of earnings. This is done for the allocation under the US system and the efficient allocation. Figure 5 presents the results, where the horizontal axis measures the present value of earnings and the vertical axis measures the lifetime net-tax rate. Results for the permanent and temporary shock cases are based on the point estimates for the variances previously highlighted in section 4 and in Figure 4.[16]

[Insert Figure 7 Here]

Figure 5 shows that the net-tax rate is increasing in the present value of earnings for both allocations. An interpretation is that both systems transfer more net resources to those with lower lifetime earnings realizations. Recall from the discussion in the introduction that the Economic Report of the President (2004, Ch. 6) highlighted exactly this pattern of resource tranfers as the mechanism by which in practice the US system provides valuable social insurance.

The most striking feature of Figure 7 is that the net-tax rate increases much more sharply in an efficient allocation as compared to the allocation under the US social security system. Consider the agent with the lowest permanent labor-productivity shock. Under the US social security system, this agent has a positive net-tax rate. In an efficient allocation this agent's net-tax rate is about -150 percent. Thus, consumption is about 250 percent of earnings. At the other end of the spectrum, consider an agent with the highest permanent labor-productivity shock. This agent has a net-tax rate of about 7 percent under the US system and about 33 percent in an efficient allocation.

It is interesting to observe in Figure 7 that a positive net-tax rate is much more likely than a negative net-tax rate. In fact, with permanent shocks the net-tax rate is positive

[16] When shocks are permanent, there are exactly 5 possible labor-productivity shock histories (see section 4). Thus, Figure 7a graphs the present value of earnings and the net-tax rate corresponding to these 5 histories. In the case of temporary shocks, there are many possible labor-productivity shock histories. Figure 7b is based on (i) drawing 10,000 shock histories, (ii) simulating consumption and earnings profiles for each of these histories, (iii) creating 8 bins for the present value of earnings and (iv) graphing the average net-tax rate in each of these bins.

under social security for all labor-productivity shock histories. The preponderance of positive net-tax rates reflects the fact that the model social security system extracts resources in expected present value terms under the parameter values considered in Table 1.[17]

How to Decompose Efficiency Gains?

Two methods come to mind.

Method 1 is to calculate the equivalent variation due to changing from (c^{ss}, l^{ss}) to (c, l^{pp}), where c raises c^{ss} proportional to the increased present value of labor earnings. The remaining gain in efficiency is due to changing consumption. Merit of this approach is that it makes the distinction between a larger pie and a smaller pie. The drawback is that the relevant allocation is not guaranteed to be incentive compatible.

Method 2 is to calculate the equivalent variation due to changing from (c^{ss}, l^{ss}) to (c, l^{ss}), where c is chosen to maximize expected utility subject to incentive compatibility and the present value budget constraint. This approach highlights the consumption insurance gain, holding labor fixed. Thus, it separates out any gain from increasing the size of the pie to be distributed, from from the gain due to improved consumption insurance with a fixed pie. This method is given by the calculations below.

$$E[U(c^{ss}(1+\alpha), l^{ss})] = \max_{c \in \Gamma(l^{ss}, Cost)} E[U(c, l^{ss})]$$

$$V_{PP} \equiv \max_{l} \{ \max_{c \in \Gamma(l, Cost)} E[U(c, l)] \}$$

$$\Gamma(l, Cost) \equiv \{c : (c, l) is IC, E[\sum_{j} \frac{(c_j - \omega(s_j, j) l_j)}{(1+r)^{j-1}}] \leq Cost\}$$

6 Discussion

Issues:

(1) Sensitivity Analysis: (i) Tighter borrowing constraints and (ii) greater Frisch elasticity of labor.

(2) Is the public information optimum far away from the private information optimum?

(3) What is the difficulty w/ analyzing a richer labor-productivity process?

(4) How to implement or approx. implement efficient allocations?

[17] Although the model is partial equilibrum, this pattern of taxation is consistent with the pattern of taxation across generations in a steady-state of a general equilibrium model with a pay-as-you-go social security system, when the interest rate is above the aggregate growth rate of the economy.

[To Be Completed]

(1) Hubbard and Judd (1987) have argued that borrowing constraints may make a social security system unattractive even when it is assumed that a social security system has an advantage over private markets in providing annuities. The simple intuition is that, in the absence of social security, a hump-shaped earnings or wage profile will make young agents want to borrowing from the future to smooth consumption. In the presence of borrowing constraints such consumption smoothing over time may be difficult. By taxing young agents to make (illiquid) transfers to these agents in old age, social security can make this pattern of consumption smoothing more difficult.

The benchmark model analyzed in the paper up to this point assumed that borrowing limits were fairly generous. Specifically, an agent could borrow up to the level consistent with paying back loans with certainty by the end of life. Now we analyze the inefficiency of the social insurance system when borrowing limits are less generous. Following Hubbard and Judd (1987), we do not model the source of these tighter borrowing limits but we do explore the consequences.

(2) Preliminary calculations indicate that they differ greatly.

In Figure 5, the inefficiency measure is derived by comparing expected utility in the US system to expected utility in the public information optimum, holding expected resources constant. In the public information optimum a planner is only subject to an expected present value resource constraint and is not subject to choosing allocations that are incentive compatible. Thus, the assumption is that the planner observes both labor productivity and earnings of the agent. Figure 5 also shows the inefficiency of the private information optimum relative to the public information optimum.

[Insert Figure 5 (a)-(b) Here]

Figure 5 is critical for understanding the degree to which welfare in the US system differs from a first-best allocation because of the private information friction or because of the non-optimality in the design of the US system. Figure 5 shows that the inefficiency of the US system is substantially larger when the benchmark for comparison is the public information optimum.[18] An important fraction of this inefficiency measure is due to the informational friction. For example, when shocks are permanent and the variance is $\sigma^2 = 0.109$ about half of this inefficiency measure is due to private information. The results for the temporary shock case when the variance is $\sigma^2 = 0.074$ are similar. Thus, taking the informational friction seriously, one message of Figure 5 is that comparisons to the public information optimum can be quantitatively quite misleading as a guide to possible efficiency gains to improving the design of the social

[18]Thus, the allocation achieving the public information optimum is not incentive compatible. In this optimum all agents share the same consumption but high productivity agents work more than low productivity agents.

insurance system.

(3) An analysis with a richer shock structure (e.g. permanent plus transitory shocks or permanent plus persistent shocks) is much more difficult. This is due entirely to the computational burden of solving the private information planning problem. This occurs because the dimension of the state space in the recursive dual problem is large when shocks are not temporary. For example, in the case of permanent shocks the recursive dual problem would have as a state variable a function $w_{s'}(s)$ describing the promised utility to an agent with shock s that reported shock s'.[19] This formulation is briefly presented in Appendix A.1. For this problem one needs to keep track of $|S|$ continuous state variables, where $|S|$ is the number of possible permanent shocks. This is a computationally daunting task even for a small number of shock values. This paper computes solutions to the primal problem rather than the dual problem when shocks are permanent.

[19]Fernandes and Phelan (2000) analyze planning problems when shocks are not independent. They consider examples where shocks take on two values. The main issues that arise in a recursive approach to such planning problems are clear from their analysis.

References

Albanesi, S. and C. Sleet (2003), Dynamic Optimal Taxation with Private Information, manuscript.

Atkeson, A. and R. Lucas (1992), On Efficient Distribution with Private Information, *Review of Economic Studies*, 59, 427-453.

Auerbach, A. and L. Kotlikoff (1987), Dynamic Fiscal Policy, (Cambridge University Press, Cambridge).

Battaglini, M. and S. Coate (2004), Pareto Efficient Income Taxation with Stochastic Abilities, manuscript.

Browning, M., Hansen, L. and J. Heckman (1999), Micro Data and General Equilibrium Models, in *Handbook of Macroeconomics,* ed. J. Taylor and M. Woodford, (Elsevier Science B.V., Amsterdam).

Congressional Budget Office (2004), Effective Federal Tax Rates: 1979- 2001, http://www.cbo.gov/.

De Nardi, M., Imrohoroglu, S. and T. Sargent (1999), Projected US Demographics and Social Security, *Review of Economic Dynamics,* 2, 575-615.

Diamond, P. (2003), Taxation, Incomplete Markets, and Social Security, (The MIT Press, Cambridge).

Diamond, P. and J. Mirrlees (1978), A Model of Social Insurance with Variable Retirement, *Journal of Public Economics*, 10, 295-336.

Diamond, P. and J. Mirrlees (1986), Payroll-Tax Financed Social Insurance with Variable Retirement, *Scandinavian Journal of Economics*, 88, 25-50.

Domeij, D. and M. Floden (2004), The Labor-Supply Elasticity and Borrowing Constraints: Why Estimates are Biased, manuscript.

Economic Report of the President (2004), (United States Government Printing Office, Washington).

Feldstein, M. (1996), The Missing Piece in Policy Analysis: Social Security Reform, *Papers and Proceedings of the American Economic Review,* 85(2), 1-14.

Fernandes, A. and C. Phelan (2000), A Recursive Formulation for Repeated Agency with History Dependence, *Journal of Economic Theory*, 91, 223-247.

Golosov, M., Kocherlakota, N. and A. Tsyvinski (2003), Optimal Indirect and Capital Taxation, *Review of Economic Studies*, 70, 569-88.

Golosov, M. and A. Tsyvinski (2004), Designing Optimal Disability Insurance: A Case for Asset Testing, NBER Working Paper No.10792.

Gouveia, M. and R. Strauss (1994), Effective Federal Icome Tax Functions: An Exploratory Analysis, *National Tax Journal*, 47, 317- 39.

Green, E. (1987), Lending and the Smoothing of Uninsurable Income, in *Contractual Arrangements for Intertemporal Trade*, editors N. Wallace and E. Prescott, (Minneapolis: University of Minnesota Press).

Heathcote, J., Storresletten, K. and G. Violante (2004), The Cross-Sectional Implications of Rising Wage Inequality in the United States, manuscript.

Hubbard, G. and K. Judd (1987), Social Security and Individual Welfare: Precautionary Savings, Liquidity Constraints and the Payroll Tax, *American Economic Review*, 77, 630- 46.

Huggett, M. and G. Ventura (1999), On the Distributional Effects of Social Security Reform, *Review of Economic Dynamics*, 2, 498- 531.

Imrohoroglu, A., Imrohoroglu, S. and D. Joines (1995), A Life Cycle Analysis of Social Security, *Economic Theory*, 6, 83-114.

Imrohoroglu, A., Imrohoroglu, S. and D. Joines (2000), Computational Models of Social Security: A Survey, in *Computational Methods for the Study of Dynamic Economies*, R. Marimon and A. Scott, eds., (Oxford University Press, Oxford).

Kocherlakota, N. (2003), Zero Expected Wealth Taxes: A Mirrlees Approach to Dynamic Optimal Taxation, Federal Reserve Bank of Minneapolis Staff Report.

Krueger, D. and F. Kubler (2003), Pareto Improving Social Security Reform when Financial Markets are Incomplete!?, manuscript.

Lindbeck, A. and M. Persson (2003), The Gains from Pension Reform, *Journal of Economic Literature*, XLI, 74-112.

Mas-Colell, A. and X. Vives (1993), Implementation in Economies with a Continuum of Agents, *Review of Economic Studies*, 60, 613- 29.

Mas-Colell, A., Whinston, M. and J. Green (1995), Microeconomic Theory, (Oxford University Press, Oxford).

Mirrlees, J. (1971), An Exploration into the Theory of Optimum Income Taxation, *Review of Economic Studies*, 38, 175- 208.

Nishiyama, S. and K. Smetters (2004), Does Social Security Privatization Produce Efficiency Gains?, manuscript.

Press, W., Teukolsky, S., Vetterling, W., and B. Flannery (1994), Numerical Recipes in Fortran: The Art of Scientific Computing, 2nd ed., Cambridge University Press.

Rogerson, W. (1985), Repeated Moral Hazard, *Econometrica*, 53, 69-76.

Siegel, J. (2002), Stocks for the Long Run, Third Edition, (McGraw Hill, New York).

Social Security Handbook (2003), see www.socialsecurity.gov.

Spear, S. and S. Srivastava (1987), On Repeated Moral Hazard with Discounting, *Review of Economic Studies*, 54, 599-617.

Storesletten, K., Telmer, C. and A. Yaron (1999), The Risk-Sharing Implications of Alternative Social Security Arrangements, *Carnegie-Rochester Conference Series on Public Policy*, 50, 213- 59.

Tauchen, G. (1986), Finite State Markov-Chain Approximations to Univariate and Vector Autoregressions, *Economics Letters*, 20, 177-81.

Thomas, J. and T. Worrall (1990), Income Fluctuations and Asymmetric Information: An Example of a Repeated Principal-Agent Problem, *Journal of Economic Theory*, 51, 367-90.

A Computational Methods

Appendix A contains three sections. Section A.1 provides theory for computing solutions to the private information planning problem. Section A.2 describes our general approach for computing solutions to the planning problem and the US social security problem. FORTRAN programs that compute solutions to these problems are available (eventually!) upon request. Section A.3 proves all Theorems from section A.1.

A.1 Private Information Planning Problem: Theory

Theory for analyzing the private information planning problem is laid out in three steps. Step 1 states a dual problem with the feature that solutions to the dual problem are solutions to the original planning problem. Step 2 provides an equivalent formulation of incentive compatibility that is useful for a recursive statement of the dual problem. Step 3 formulates the dual problem as a dynamic programming problem and indicates how to further simplify this problem for computational purposes.

A.1.1 Primal and Dual Problems

Primal Problem: $\max E[\sum_j \beta^{j-1} u(c_j, l_j)]$

subject to (1) (c,l) is IC and (2) $E[\sum_j (c_j - s_j l_j)/(1+r)^{j-1}] \leq Cost$

Dual Problem: $\min E[\sum_j (c_j - s_j l_j)/(1+r)^{j-1}]$

subject to (1) (c,l) is IC and (2) $E[\sum_j \beta^{j-1} u(c_j, l_j)] \geq u^*$

Theorem A1: Assume $u(c,l) = u(c) + v(l)$, $u(c)$ is continuous on R^1_{++} and $u(c)$ is strictly increasing. If (c,l) solves the dual problem, given $u^* > -\infty$, then (c,l) solves the primal problem, given $Cost \equiv E[\sum_j \frac{(c_j - s_j l_j)}{(1+r)^{j-1}}]$.

Proof: See Appendix A.3

Theorem A2 provides conditions which are equivalent to the incentive compatibility conditions.[20]

Theorem A2:

(i) Consider the case of independent shocks.
(c,l) is IC iff $\exists \{w_j(s^{j-1})\}_{j=2}^{J+1}$ such that restrictions (a)-(b) hold:
(a) $u(c_j(s^{j-1}, s_j), l_j(s^{j-1}, s_j)) + \beta w_{j+1}(s^{j-1}, s_j) \geq$
$\quad u(c_j(s^{j-1}, s'_j), l_j(s^{j-1}, s'_j)(s'_j/s_j)) + \beta w_{j+1}(s^{j-1}, s'_j), \forall (s^{j-1}, s_j), \forall s'_j$
(b) $w_j(s^{j-1}) = E[u(c_j(s^j), l_j(s^j)) + \beta w_{j+1}(s^j)|s^{j-1}]$ and $w_{J+1}(s^J) = 0$
where s^j denotes the history of (truthful) reports up to period j.

[20] These results are adaptations of Green (1987, Lemma 1-2).

(ii) Consider the case of permanent shocks.
(c, l) is IC iff $\exists\{w_j(s, s')\}_{j=2}^{J+1}$ such that restrictions (a)-(b) hold:
(a) $u(c_1(s), l_1(s)) + \beta w_2(s, s) \geq u(c_1(s'), l_1(s')(s'/s)) + \beta w_2(s, s'), \forall s, s'$
(b) $w_j(s, s') = u(c_j(s'), l_j(s')(s'/s)) + \beta w_{j+1}(s, s')$ and $w_{J+1}(s, s') = 0$

Proof: See Appendix A.3.

A.1.2 Recursive Formulation of the Dual Problem

Temporary Shocks

A recursive formulation for the Dual problem is provided below for the case of temporary shocks. The function $C_j(w)$ is the minimum expected discounted cost of obtaining utility w. The notation (c_i, l_i, w_i) describes period consumption, labor and future utility delivered when shock $i = 1, ..., I$ occurs and the agent tells the truth.

$C_j(w) = \min \sum_i [c_i - l_i s_i + (1 + r)^{-1} C_{j+1}(w_i)]\pi_i$
subject to $(c_i, l_i, w_i)_{i \in I} \in \{(c_i, l_i, w_i)_{i \in I}$: IC and PK constraints hold $\}$
(PK) $w = \sum_i [u(c_i, l_i) + \beta w_i]\pi_i$,
(IC) $u(c_i, l_i) + \beta w_i \geq u(c_j, l_j(s_j/s_i)) + \beta w_j, \forall i, j$

Theorem A3 establishes some basic properties of the incentive constraints.[21] The following compact notation is used: $C_{ij} \equiv u(c_i, l_i) + \beta w_i - [u(c_j, l_j(s_j/s_i)) + \beta w_j]$. $C_{ii-1} \geq 0$ is called a local downward incentive constraint, whereas $C_{i-1i} \geq 0$ is called a local upward incentive constraint. Theorem A3 says that (a) the local upward and downward constraints convey all the IC restrictions (Thm. A3(ii)), (b) if all the local downward constraints bind then all local upward constraints also hold (Thm. A3(iii)) and (c) in a solution to the recursive dual problem all local downward constraints bind (Thm. A3(iv)). Theorem A3 also delivers the standard insight that the incentive compatibility restrictions alone imply that "earnings" or "output" $l_i s_i$ increases as the shock i increases.

Theorem A3: In the recursive dual problem assume $u(c, l) = u(c) + v(l)$, u and v are strictly concave, u is increasing, v is decreasing and that shocks are independent and ordered so that $s_1 < s_2 < ... < s_I$. Then

(i) Incentive compatibility implies that $l_i s_i$ is increasing in i.

(ii) $C_{ii-1}, C_{i-1i} \geq 0, i = 2, ..., I$ imply $C_{ij} \geq 0 \quad \forall i, j$.

(iii) $C_{ii-1} = 0, i = 2, ..., I$ imply $C_{i-1,i} \geq 0, i = 2, ..., I$ and $C_{i-1,i} > 0$ whenever $l_i s_i > l_{i-1} s_{i-1}$.

(iv) In a solution to the recursive dual problem all local downward constraints bind.

[21] Theorem A3 is parallel to results which hold w/o a labor-leisure decision (e.g. Thomas and Worrall (1990)) and to results in the literature following Mirrlees (1971).

Proof: See Appendix A.3.

To compute solutions to the recursive dual problem it is useful to solve two subproblems: DP 1 and DP 2. These problems reduce the dimensionality of the choice variables by making use of additive separability of the objective. Dimensionality can be further reduced by solving DP 1' in place of DP 1. DP 1' solves out for utility z_i in terms of promised utility w and the labor plan $(l_1, ..., l_I)$ by using the fact, established in Theorem A3, that all local downward constraints hold with equality and that the local downward constraints imply all the restrictions of incentive compatibility.

Subproblems:

(DP 1) $C_j(w) = \min \sum_i [-l_i s_i + \hat{C}_j(z_i))]\pi_i$

(1) $w = \sum_i [v(l_i) + z_i]\pi_i$

(2) $v(l_i) + z_i \geq v(l_j(s_j/s_i)) + z_j, \forall i, j$

(DP 2) $\hat{C}_j(z) = \min_{\{(c,w'):z=u(c)+\beta w'\}} c + (1+r)^{-1} C_{j+1}(w')$

(DP 1') $C_j(w) = \min \sum_i [-l_i s_i + \hat{C}_j(f_i(l_1, ..., l_I; w)))]\pi_i$

The functions $z_i = f_i(l_1, ..., l_I; w)$ in problem DP 1' are constructed in the two equations below. The first equation holds for $i > 1$ by repeated substitutions from the downward IC constraint. This equation says that promised utility z_i to a person with shock i is the utility to the person with the lowest shock z_1 plus the sum of the utility differences when one lies downward one shock. These utility differences are positive by Thm. A.3(i). The second equation holds by substituting the first equation into the promise keeping constraint. This then states z_1 in terms of the labor choices and promised utility w. These two equations define the functions $z_i = f_i(l_1, ..., l_I; w)$.

$$z_i = z_1 + \sum_{j=2}^{i} [v(l_{j-1}(s_{j-1}/s_j)) - v(l_j)]$$

$$z_1 = w - \sum_{i=1}^{I} v(l_i)\pi_i - \sum_{i=2}^{I} [\sum_{j=2}^{i} (v(l_{j-1}(s_{j-1}/s_j))) - v(l_j)]\pi_i$$

Permanent Shocks

A recursive formulation for the Dual problem for the case of permanent shocks is provided below. Although we will not use this formulation for computation, it is helpful to see what would be entailed. In this problem, the choice variables are consumption and labor in each state as well as promised utility $w'_s(\hat{s})$ next period. $w'_s(\hat{s})$ is the promised utility for an agent with true state $\hat{s}$ who reports state s. At a computational level, the dimension of the state space can be quite large for $j \geq 2$ as for any value of s one needs to keep track of a function $w'_s(\hat{s})$. With $|S|$ possible permanent shocks, the state variable in period 2 and beyond has $|S|$ continuous state variables. Fernandes and Phelan (2000) consider recursive formulations of problems from dynamic contract theory where similar issues arise.

$C_1(w) = \min \sum_s [c(s) - l(s)s + (1+r)^{-1} C_2(s, w'_s(\hat{s}))] P(s)$
subject to $(c(s), l(s), w'_s(\hat{s}))$ satisfying (1)-(2)
(1) $u(c(s), l(s)) + \beta w'_s(s) \geq u(c(\bar{s}), l(\bar{s})(\bar{s}/s)) + \beta w'_{\bar{s}}(s), \forall s, \bar{s}$
(2) $w = \sum_s [u(c(s), l(s)) + \beta w'_s(s)] P(s)$

$C_j(s, w_s(\hat{s})) = \min c - ls + (1+r)^{-1} C_{j+1}(s, w'_s(\hat{s})), \forall j \geq 2$
subject to $(c, l, w'_s(\hat{s}))$ satisfying (3)
(3) $w_s(\hat{s}) = u(c, l(s/\hat{s})) + \beta w'_s(\hat{s}), \forall \hat{s}$

A.2 Computation

A.2.1 Social Security Problem

The social security problem is stated below as a dynamic programming problem. This involves reformulating the present value budget constraint as a sequence of budget constraints where resources are transfered across periods with a risk-free asset. Risk-free asset holding must then always lie above period and shock specific borrowing limits: $\underline{a}_j(s)$.[22] The state variable is (a, s, z) where a is asset holdings, s is the period productivity shock and z is average past earnings. The functions T_j and F_j describe the tax system and the law of motion for average past earnings. Labor productivity is a Markov process with transition probability $\pi(s'|s)$.

$V_j(a, s, z) = \max_{(c,l,a')} u(c, l) + \beta \sum_{s'} V_{j+1}(a', s', z') \pi(s'|s)$
(1) $c + a' \leq a(1+r) + \omega(s, j) l - T_j(a, z, \omega(s, j) l)$
(2) $c \geq 0, a' \geq \underline{a}_j(s); l \in [0, 1]$
(3) $z' = F_j(z, \omega(s, j) l)$

This problem is solved computationally by backwards induction. The value function $V_j(a, s, z)$ is computed at selected grid points (a, s, x) by solving the right-hand-side of Bellman's equation using the simplex method. Specifically, we use amoeba from Press et al (1994). This involves a bi-linear interpolation of the function $V_{j+1}(a', s', z')$ over the two continuous variables (a', z'). We set the borrowing limit to a fixed value $\underline{a}$ in each period. We then relax this value so that it is not binding. This is a device for imposing period and state specific limits $\underline{a}_j(s)$. To use this device, penalties are imposed for states and decisions implying negative consumption.[23]

We compute ex-ante, expected utility V_{SS} and the expected cost of running the social security system, denoted $Cost$, by simulation, under the assumption that an agent starts out with no assets. Specifically, we draw a large number of lifetime labor-productivity profiles, compute realized utility and realized cost for each profile and then compute averages.

[22] These limits are the maximum present value of labor earnings plus social security benefits in the worst labor-porductivity history. This assumes that one can borrow against future social security benefits.

[23] We mention two points. First, the backward induction mentioned above takes as given a value for average earnings in the economy. This variable is used to determine the retirement benefit function. Thus, an additional loop is needed so that guessed and implied values of average earnings coincide. Second, we use 1000 evenly spaced grid points on assets a, 25 grid points on average earnings z over the interval $[0, e_{max}]$.

A.2.2 Planning Problem

We describe how we compute the optimized value $V_{Private}$, given the value of $Cost$. The algorithm for the temporary shock case is presented first.

(DP 1') $C_j(w) = \min \sum_i [-l_i s_i + \hat{C}_j(f_i(l_1, ..., l_N; w)))]\pi_i$
(DP 2) $\hat{C}_j(z) = \min_{\{(c,w'):z=u(c)+\beta w'\}} c + (1+r)^{-1} C_{j+1}(w')$
Algorithm:

1. Set terminal value function on grid points $w \in \{w_1, ..., w_M\}$: $\hat{C}_J(w) \equiv u^{-1}(w)$
2. For each $w \in \{w_1, ..., w_M\}$, we use amoeba from Press et al (1994) to solve the right-hand-side of DP 1' to compute C_j. This involves a linear interpolation of $\hat{C}_j$.
3. Given C_j, compute $\hat{C}_{j-1}$ at gridpoints by solving DP 2. This is done by grid search.
4. Repeat steps 2-3 for all ages j back to age 1.
5. Solve the equation $C_1(V_{Private}) = Cost$ for $V_{Private}$. This is done by simulation using the optimal decision rules.

We now indicate how to compute $V_{Private}$ for the case of permanent shocks. The original formulation of the permanent shock problem is stated below.

$$V_{Private} \equiv \max_{(l_j(s), c_j(s))} \sum_s [\sum_j \beta^{j-1}(u(c_j(s)) + v(l_j(s)))]P(s) \text{ s.t.}$$
$$\text{(i) } \sum_s [\sum_j (c_j(s) - l_j(s)s)/(1+r)^{j-1}]P(s) \leq Cost$$
$$\text{(ii) } \sum_j \beta^{j-1}(u(c_j(s)) + v(l_j(s)) \geq \sum_j \beta^{j-1}(u(c_j(s')) + v(l_j(s')s'/s)), \forall s, \forall s'$$

We analyze a "relaxed" problem which is the same as the problem above except that we require that only the local downward incentive constraints hold rather than all the incentive compatibility constraints. It is straightforward to show two results. First, in a solution to the relaxed problem all the local downward incentive constraints bind. Second, if an allocation (c, l) has the property that all the downward incentive constraints bind and $l_j(s)s$ is increasing in s for all j, then all the incentive constraints hold. [The proof of this assertion is similar to the argument in the proof of Thm. A3 (ii)-(iii).] Our computational strategy is therefore to compute solutions to the relaxed problem AND to verify ex-post that $l_j(s)s$ is increasing in s (i.e in a solution to the relaxed problem required output of an agent in any period of life is increasing in the agent's productivity shock).

$$\max_{(l)} \sum_s [\sum_j \beta^{j-1} v(l_j(s)) + g(l, s, cost)]P(s)$$

We compute solutions to the relaxed problem by solving the equivalent problem above. This equivalent problem is useful for computational purposes as it reduces the dimension of the control variables by substituting out all binding constraints. This equivalence follows from two observations. First, additive separability of $u(c, l)$ implies that the intertemporal MRS of consumption is chosen in the relaxed problem without distortion. Using this fact, maximization could then be done over labor and the lifetime utility of consumption $u(s)$.

This eliminates the choice of consumption from the problem. The relevant cost constraint is written below, where $COST(u(s))$ is a known function, derived from the first order conditions to the relaxed problem, describing the minimum resource cost of obtaining lifetime utility of consumption $u(s)$.[24] Second, we can also eliminate maximizing over $u(s)$ by expressing $u(s) = g(l, s, Cost)$ as a function of the labor plan and other data. To do this, we solve out $u(s)$ from all relevant binding constraints. The last two equations below are intermediate steps towards computing $u(s) = g(l, s, Cost)$. The last equation uses the fact that shocks are ordered so that $s_1 \leq s_2 \leq ... \leq s_I$.

$$\sum_s [COST(u(s)) - \sum_j sl_j(s)/(1+r)^{j-1}]P(s) = Cost$$

$$\sum_j \beta^{j-1} v(l_j(s)) + u(s) = \sum_j \beta^{j-1} v(l_j(s')s'/s)) + u(s')$$

$$u(s_n) = u(s_1) + \sum_{i=2}^{n} [\sum_j \beta^{j-1} v(l_j(s_{i-1})s_{i-1}/s_i)) - \sum_j \beta^{j-1} v(l_j(s_i))]$$

We use amoeba from Press et al (1994) to solve the relaxed problem. This involves maximizing over labor choices $(l_1(s), ..., l_{R-1}(s))$. These choices lie in an $R-1 \times |S|$ dimensional space as there are $R-1$ labor periods and $|S|$ possible permanent shocks. Each evaluation of the objective requires the computation of the function $g(l, s, Cost)$. This involves finding a value $u(s_1)$ solving the three equations above, given $(l_1(s), ..., l_{R-1}(s))$ and Cost.

A.3 Proofs of Theorems A1-3

Theorem A1: Assume $u(c,l) = u(c) + v(l)$, $u(c)$ is continuous on R^1_{++} and $u(c)$ is strictly increasing. If (c,l) solves the dual problem, given $u^* > -\infty$, then (c,l) solves the primal problem, given $Cost \equiv E[\sum_j \frac{(c_j - s_j l_j)}{(1+r)^{j-1}}]$.

Proof: Suppose not. Thus, there exists $(\bar{c}, \bar{l})$ that is IC and costs no more than (c,l) but that delivers strictly more expected utility than (c,l). Construct (c^*, l^*) that satisfies constraints (1)-(2) in the Dual Problem but that delivers strictly lower cost than (c,l).

Set $l^*_j \equiv \bar{l}_j, \forall j$ and $c^*_j \equiv \bar{c}_j, \forall j \geq 2$. Set $c^*_1(s)$ to solve $u(c^*_1(s)) = u(\bar{c}_1(s)) - \epsilon$. Thus, $c^*_1(s)$ produces a uniform decrease in utility in period 1 of $\epsilon > 0$. If $\bar{c}_1(s) > 0, \forall s$ [Need an extra assumption! $u(0) = -\infty$ is sufficient.], then by continuity there exists $\epsilon > 0$ such that $c^*_1(s) \geq 0, \forall s$ and $E[\sum_j \beta^{j-1} u(c^*_j, l^*_j)] \geq u^*$. Since $(\bar{c}, \bar{l})$ is IC and the utility decrease is uniform regardless of reports, (c^*, l^*) is also IC. This is a contradiction since (c^*, l^*) costs strictly less than (c,l). □

Theorem A2:

[24] When $\beta(1+r) = 1$, $COST(u(s))$ has a simple form as consumption is constant. When $\beta < 1$ and $r > 0$ then $COST(u(s)) = u^{-1}[(1-\beta)u(s)/(1-\beta^J)][1-(1/(1+r))^J](1+r)/r$.

- (i) (Independent Shocks)
  (c, l) is IC iff $\exists\{w_j(s^{j-1})\}_{j=2}^{J+1}$ such that restrictions (a)-(b) hold:

(a) $u(c_j(s^{j-1}, s_j), l_j(s^{j-1}, s_j)) + \beta w_{j+1}(s^{j-1}, s_j) \geq$
$u(c_j(s^{j-1}, s'_j), l_j(s^{j-1}, s'_j)(s'_j/s_j)) + \beta w_{j+1}(s^{j-1}, s'_j), \forall (s^{j-1}, s_j), \forall s'_j$
(b) $w_j(s^{j-1}) = E[u(c_j(s^j), l_j(s^j)) + \beta w_{j+1}(s^j)|s^{j-1}]$ and $w_{J+1}(s^J) = 0$
where s^j denotes the history of (truthful) reports up to period j.

- (ii) (Permanent Shocks)
  (c, l) is IC iff $\exists\{w_j(s, s')\}_{j=2}^{J+1}$ such that restrictions (a)-(b) hold:

(a) $u(c_1(s), l_1(s)) + \beta w_2(s, s) \geq u(c_1(s'), l_1(s')(s'/s)) + \beta w_2(s, s'), \forall s, s'$
(b) $w_j(s, s') = u(c_j(s'), l_j(s')(s'/s)) + \beta w_{j+1}(s, s')$ and $w_{J+1}(s, s') = 0$
Proof:
(i) ($\Rightarrow$) Backward induction on restriction (b) defines the function w_{j+1} uniquely. Substitute w_{j+1} into restriction (a). The resulting inequality is then a direct implication of (c, l) being IC. Specifically, it is implied by truth telling being superior to a feasible report σ where one reports truthfully at all ages and histories except age-history (s^{j-1}, s_j) where the report is s'_j rather than s_j. [Independence used here.]
($\Leftarrow$) Suppose not. Then restriction (a)-(b) hold but there is a report σ that strictly improves over truth telling, given (c, l). Let σ have the smallest number of false reports at distinct age-histories s^j among those report functions σ that strictly improve over truth telling. This is clearly possible since the number of age-histories is finite. Choose j as large as possible so that σ involves a false report (i.e. $\sigma_j(s^j) \neq s_j$) at some age-history s^j. Then restriction (a)-(b) implies that given that σ has been used in the past, telling the truth in period j and subsequently leads to at least as much conditional expected utility at age-history s^j as using σ. Thus, there is another feasible report function that strictly improves over truth telling and that has a smaller number of false reports. Contradiction.
(ii) ($\Rightarrow$) Set $w_j(s, s') = \sum_{k=j}^{J} \beta^{k-j} u(c_k(s'), l_k(s')(s'/s))$. This satisfies restriction (b). Restriction (a) holds since (c, l) is IC.
($\Leftarrow$) Backward induction on restriction (b) defines $w_j(s, s')$ uniquely:

$$w_j(s, s') = \sum_{k=j}^{J} \beta^{k-j} u(c_k(s'), l_k(s')(s'/s))$$

Insert this into restriction (a) to produce the condition that (c, l) is IC. □

Theorem A3: In the recursive dual problem assume $u(c, l) = u(c) + v(l)$, u and v are strictly concave, u is increasing, v is decreasing and that shocks are independent and ordered so that $s_1 < s_2 < ... < s_I$. Then

(i) Incentive compatibility implies that $l_i s_i$ is increasing in i.

(ii) $C_{ii-1}, C_{i-1i} \geq 0, i = 2, ..., I$ imply $C_{ij} \geq 0 \quad \forall i, j$.

(iii) $C_{ii-1} = 0, i = 2, ..., I$ *imply* $C_{i-1,i} \geq 0, i = 2, ..., I$ *and* $C_{i-1,i} > 0$ *whenever* $l_i s_i > l_{i-1} s_{i-1}$.

(iv) In a solution to the recursive dual problem all local downward constraints bind.

Proof:

(i) Assume that it is feasible to claim to have received shock i when one has shock $i-1$. If not, then $l_i s_i \geq l_{i-1} s_{i-1}$ holds trivially. Thus, we have that $C_{ii-1}, C_{i-1i} \geq 0$. Adding these inequalities and using the fact that $u(c,l) = u(c) + v(l)$ implies the first equation below. The second equation rearranges the first. The second equation and v concave then implies that $l_i s_i \geq l_{i-1} s_{i-1}$ must hold.

$$v(l_i) - v(l_{i-1} s_{i-1}/s_i) \geq v(l_i s_i/s_{i-1}) - v(l_{i-1})$$

$$v(l_i s_i/s_i) - v(l_{i-1} s_{i-1}/s_i) \geq v(l_i s_i/s_{i-1}) - v(l_{i-1} s_{i-1}/s_{i-1})$$

(ii) Show first that $C_{ij} \geq 0, \forall j < i$. As a first step show that $C_{ii-2} \geq 0$. This follows from the three lines below. The first line is $C_{i-1i-2} \geq 0$. The second line follows from line one and the fact that $v(l_{i-1} s_{i-1}/s) - v(l_{i-2} s_{i-2}/s)$ increases as s increases for $s \geq s_{i-1}$. The last fact holds since $l_i s_i$ increases as i increases (Thm. A3(i)) and since v is concave. Line three follows from line two and $C_{ii-1} \geq 0$.

$$u(c_{i-1}, l_{i-1}) + w_{i-1} \geq u(c_{i-2}, l_{i-2} s_{i-2}/s_{i-1}) + w_{i-2}$$

$$u(c_{i-1}, l_{i-1} s_{i-1}/s_i) + w_{i-1} \geq u(c_{i-2}, l_{i-2} s_{i-2}/s_i) + w_{i-2}$$

$$u(c_i, l_i) + w_i \geq u(c_{i-1}, l_{i-1} s_{i-1}/s_i) + w_{i-1} \geq u(c_{i-2}, l_{i-2} s_{i-2}/s_i) + w_{i-2}$$

To show that $C_{ij} \geq 0$ holds for all $j < i$, proceed by induction repeating the three steps above, where the first step is the induction step.

It remains to show that $C_{ij} \geq 0, \forall j > i$ if any of these upward lies are feasible. As a first step show that $C_{ii+2} \geq 0$. This follows from the three lines below for essentially the same reasons as in the argument above. The remainder of the proof follows by an induction which is parallel to that given above.

$$u(c_{i+1}, l_{i+1}) + w_{i+1} \geq u(c_{i+2}, l_{i+2} s_{i+2}/s_{i+1}) + w_{i+2}$$

$$u(c_{i+1}, l_{i+1} s_{i+1}/s_i) + w_{i+1} \geq u(c_{i+2}, l_{i+2} s_{i+2}/s_i) + w_{i+2}$$

$$u(c_i, l_i) + w_i \geq u(c_{i+1}, l_{i+1}s_{i+1}/s_i) + w_{i+1} \geq u(c_{i+2}, l_{i+2}s_{i+2}/s_i) + w_{i+2}$$

(iii) Obvious from v strictly concave.

(iv) (Rough argument) Suppose not. Let (c_i, l_i, w_i) be a solution in which a downward constraint is not binding. Construct (c_i^*, l_i^*, w_i^*) so that labor and future utility are the same as before but consumption is different. Squeeze the consumption distribution so that (a) mean consumption is lower, (b) all downward constraints still hold and (c) mean $u(c)$ unchanged. This lowers the objective and satisfies all constraints. Contradiction.

[Note: Argument involves lowering consumption in some state. Thus, one needs strictly positive consumption. A sufficient condition for this to hold for states $w > -\infty$ is $u(0) = -\infty$.]

□