August 14, 2007

To: Workshop Participants

From: Louis Kaplow

Re: Taxation and Social Security

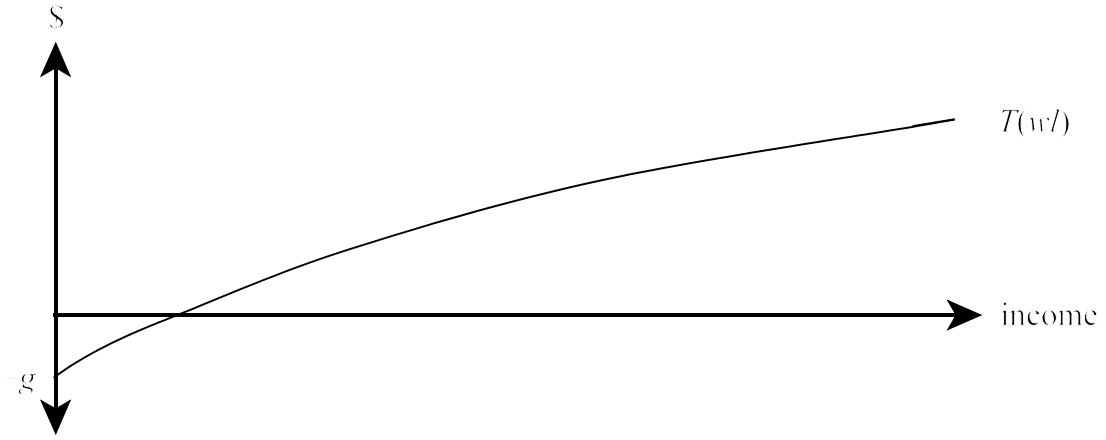
Attached is draft chapter 11 from my forthcoming book, *The Theory of Taxation and Public Economics* (Princeton University Press).

Some parts, especially the first few pages, may be difficult to understand from lack of context. The material on the next page of this cover memo will help to fill in some of the blanks.

In addition, other portions (particularly subsection B.1 and B.2) are more technical, but it is hoped that most of the ideas can be gleaned from the prose.

BACKGROUND FOR CHAPTER 11
Excerpts from Chapters 4 and 9

Figure 4.1
Nonlinear Income Tax and Transfer Schedule



T is the tax/transfer schedule. It is a function of income, wl: w is an individual's wage rate and l the amount of labor supplied. The term g refers to the grant received by an individual earning no income, i.e., $-T(0)$, reflecting that the tax schedule T is taken to represent the entire tax-transfer system. Taxes may include sales taxes or VAT payments in addition to income taxes. Transfers include those through the tax system, such as the EITC in the United States, welfare programs (see chapter 7), and under some interpretations public goods (see chapter 8).

* * * * * * * * * *

Consider a two-period model wherein individuals work only in period 1 and consume in periods 1 and 2. (Period 1 can be thought of as an aggregate of an individual's working years and period 2 as retirement years; the extension to more periods or to a continuous-time model is straightforward.) Suppose further that there is only one type of commodity in each of the two periods, denoted c_1 and c_2. Individuals' utility is $u(c_1, c_2, l)$. Finally, normalize the within-period price of each commodity to one, and let r be the interest rate. The individual's budget constraint is

$$(9.1)\ c_1 + \frac{c_2}{1+r} = wl - T(wl).$$

CHAPTER 11

TAXATION AND SOCIAL SECURITY

A substantial fraction of taxation and expenditure in developed economies is devoted to social insurance, especially to finance consumption during years of retirement (including consumption of medical care). Systems typically impose a labor income tax—such as a flat-rate payroll tax—during working years to finance payments to retirees.[1]

This chapter first analyzes purely redistributive aspects of social security schemes in a setting in which individuals are taken to be rational, far-sighted utility maximizers not subject to liquidity constraints. Then these assumptions are relaxed for purposes of considering a central feature of social security, the forcing of a minimum level of savings. Finally, but briefly, some additional insurance dimensions are noted. A number of other important features of social security are not inherently related to the central themes of this book and therefore are omitted, including broader fiscal issues involving deficits and investment policy as well as political economy considerations, such as those related to pre-funding and the merits of privatization.[2]

A. Redistribution

1. Labor Income Tax Comparison

To examine the purely redistributive element of social security, it is useful to set aside its other features and potential respects in which individual behavior may deviate from rational maximization of a standard utility function. Consider the simple two-period model employed in chapter 9 to analyze taxation of capital: Individuals work in only the first period and consume in both periods, with first-period savings earning interest (here assumed for ease of exposition to be untaxed). In addition to the redistributive labor income tax, now denoted $T^I(wl)$, individuals pay a social security tax of $T^S(wl)$ as well, and in the second period they receive social security benefits of $B^S(wl)$, which depend on their previous earnings. Modifying the budget constraint (9.1) and rearranging terms indicates that second-period consumption is given by

$$(11.1)\quad c_2 = \left[wl - T^I(wl) - T^S(wl) - c_1\right](1+r) + B^S(wl).$$

If social security were actuarially fair, we would have $T^S(wl)(1+r) = B^S(wl)$ for all wl. In allowing for redistributive social security, this equality is not imposed for any particular type of individual. One can, however, assume that $(1+r)\int T^S(wl) = \int B^S(wl)$. (This assumption, as will be

---

[1]Many countries now mandate private retirement schemes in addition to or in lieu of government social insurance. See Bateman, Kingston, and Piggott (2001). A relevant distinction is that private retirement accounts tend to be actuarially fair by design. In any event, much of the analysis of this chapter is applicable to these programs as well.

[2]For broader treatments, see, for example, Diamond (1977, 2002, 2003, 2004), Feldstein (2005), and Feldstein and Liebman (2002b).

apparent, is without loss of generality in the present model; the possibility of intergenerational redistribution in a setting with overlapping generations is considered in subsection 3.)

Define the net tax (transfer, if negative) imposed by social security as $T^N(wl) = T^S(wl) - B^S(wl)/(1+r)$. Using this expression to substitute for $T^S(wl)$ in (11.1) yields

$$(11.2)\quad c_2 = \left[wl - T^I(wl) - T^N(wl) - B^S(wl)/(1+r) - c_1\right](1+r) + B^S(wl)$$

$$= \left[wl - T^I(wl) - T^N(wl) - c_1\right](1+r)$$

$$= \left[wl - T(wl) - c_1\right](1+r),$$

where $T(wl) = T^I(wl) + T^N(wl)$, making use of the fact that the labor income tax and social security tax are both functions of (first-period) earnings. The last line in expression (11.2) is, of course, simply a rearrangement of the budget constraint (9.1) for the problem with no social security.

Accordingly, the existence of a redistributive social security system makes no difference in the present setting.[3] Any redistribution can be incorporated into the labor income tax and transfer scheme. Furthermore, that some of earnings must be set aside in period 1 for consumption in period 2 is immaterial because it is assumed here that individuals can borrow and lend, without constraint, at the rate r and that their decisions are fully rational. Therefore, it is not meaningful to ask how redistributive a social security scheme is or should be, unless political factors distinguish between economically equivalent systems or one introduces other features such as myopia, liquidity constraints, and forced savings, as is done in section B. Even in the latter case, it should not matter what portion (if any) of the payments that individuals at any income level are required to make in period 1, as a function of earnings, is nominally deemed to be part of the income tax or a separate social security tax. Nevertheless, following the practice employed throughout this book, it often will be convenient analytically to hold redistribution (in the entire fiscal system) constant in order to examine the optimal magnitude of an (actuarially fair) social security system, say, when individuals are myopic.

An implication of the foregoing discussion is that familiar claims regarding the efficiency consequences of marginal tax-benefit linkage in social security systems are potentially misleading. Linkage is said to be complete when $T^{S\prime}(wl) = B^{S\prime}(wl)/(1+r)$ (implying that $T^{N\prime}(wl) = 0$) and nonexistent when $T^{N\prime}(wl) = T^{S\prime}(wl)$ (implying that $B^{S\prime}(wl) = 0$, which means that benefits are uniform, independent of earnings). Moving, say, from no marginal linkage to complete linkage does reduce labor supply distortion, assuming that income taxes are unchanged. Note, however, that such a reform accomplished through changing the benefits formula necessarily entails a change from lump-sum benefits to benefits tied to earnings in a

[3]For analyses of the extent of redistribution in the United States social security system and under various reform proposals, see Feldstein and Liebman (2002a) and Coronado, Fullerton, and Glass (2000).

manner that has the same marginal incidence as the social security tax. If instead the tax formula is changed, it would need to be converted to a uniform lump-sum tax, to match the incidence of the benefits. In either case, the resulting reduction in distortion arises as a consequence of a concomitant reduction in redistribution. The increase in linkage is equivalent to reducing marginal tax rates in the income tax, funded by shrinking the lump-sum grant. That one can reduce distortion by reducing redistribution has nothing in particular to do with social security tax-benefit linkage. Furthermore, observe that if one wished to improve tax-benefit linkage within social security without changing overall redistribution, the income tax schedule would need to be adjusted in an offsetting manner. But in that case $T(wl)$ would be unchanged, and there would be no reduction in labor supply distortion.

2. Lifetime Income

Social security retirement benefits are ordinarily a function of individuals' earnings over the course of their working lives, which raises questions concerning optimal redistribution from a lifetime perspective. This problem is naturally analyzed using the optimal income taxation framework. As a first cut, Diamond (1977, 2003) suggests that one might reinterpret Mirrlees (1971) as addressing how lifetime taxes and transfers should depend on lifetime income. If individuals' earning abilities or utility functions vary over time, including importantly cases involving uncertainty (whether of earning ability, utility, or lifespan), further analysis is required.

Subsection 5.C.1 introduces the generalization involving group-specific income tax schedules $T(wl,\theta)$, where in the present setting θ might index individuals' ages. This formulation, represented in expression (5.1), is insufficient for present purposes because, for example, consumption at any given age—and thus both the marginal utility of consumption and the marginal contribution of utility to social welfare, W'—will in general depend on past earnings and consumption as well as on expected future earnings and consumption. Consequently, regarding the tax schedule itself, one may wish to interpret θ as a vector indicating not only age but also earnings history, so that a current period's taxes may depend on prior earnings as well as on current earnings and age per se (implicitly making possible any manner of lifetime income averaging, on which more below).[4]

This substantially more complex problem has received limited attention. As a matter of efficiency, one might suppose that marginal tax rates should be constant over time because distortion rises disproportionately with the marginal tax rate.[5] However, even with utility functions that are time-separable and identical in each period, it need not be true that constant marginal rates minimize distortion in raising a given amount of revenue (in present value) from

---

[4]The analysis in this subsection implicitly assumes that the government commits to a tax schedule so that it is not possible ex post to extract more tax from individuals who, in prior periods, have revealed themselves to have high ability through their earnings. See the brief discussion in subsection 9.C.2 on capital levies and transitions.

[5]The present suggestion brings to mind the well-known result of Barro (1979); however, his analysis simply assumed that the function relating distortion and taxation is the same in each period (the model is a reduced-form pertaining to the economy as a whole), so the issues to be explored here did not arise.

an individual whose earning ability varies over time.[6] As Heckman (1974) shows, individuals will tend to exert greater labor effort in periods in which their *w* is higher: Starting from a point of equal effort in each period, slightly raising effort in a high-*w* period and lowering effort by the same amount in a low-*w* period will have no first-order effect on the disutility of labor effort but will increase earnings.[7]

Given that both *w* and *l* will differ across periods (starting from a base case of identical tax functions in each period), it is hardly obvious that the labor supply elasticity will be the same in each period. In particular, the elasticity may be lower in high-*w*, high-*l* periods (that is, high-income periods).[8] If a given percentage increase in *w* raised *l* by a common percentage even in high-*w* periods, the increase in lifetime consumption would be greater than for other, low-*w* periods, so marginal utility would fall more, requiring a greater reduction in labor effort to restore individuals' first-order conditions. (Note that higher consumption in one period causes labor supply to fall in all periods: Saving and borrowing are used to equate marginal utility across periods, so changes in lifetime consumption can be thought of as changing a common marginal utility of consumption; when that marginal utility falls, individuals will find it optimal to reduce labor effort in all periods. See, for example, Heckman (1974) and MaCurdy (1981).) However, for a given elasticity, a higher *w* implies a lower optimal marginal tax rate because a given reduction in labor effort is more costly. (Recall the discussion of the denominator of expression (4.10).) These two potentially competing effects indicate that the question of the optimal lifetime pattern of marginal tax rates is complex; constancy may not be optimal, but the nature of the optimal deviation is not obvious. A further complication is that, in a system with earnings-history-dependent taxation, labor effort in any period will in general affect expected marginal tax rates in future periods, so the current effective marginal tax rate diverges from the rate nominally indicated by the tax schedule.

---

[6]Note that the present problem is formally quite similar to a version of the problem of taxing a two-earner family considered in subsection 12.B.1.b. There, a case is examined in which two family members jointly choose labor effort and allocate consumption between themselves to maximize the sum of their utilities. The two different family members correspond to two different time periods (imagining now that an individual works and consumes in two periods); the allocation of consumption between the family members is governed by the same principles as the individual's allocation of consumption between the two periods, and the family members' choices of labor effort are governed by essentially the same first-order conditions as the individual's choices of labor effort in the two periods. Accordingly, the potential optimality of differentially taxing the earnings of the two family members is closely related to the potential optimality of tax rates varying across time periods in the present setting.

[7]Discussion in the text will abstract from complications arising from positive interest and (utility) discount rates. For example, a positive interest rate makes present earnings more valuable than future earnings, but one can interpret *w* as a time-adjusted (interest-rate-adjusted) effective wage for purposes of comparing wages across periods.

[8]This conjecture is suggestive at best because the result depends on the form of the utility function, the stipulated simplifying assumptions, and particularly on cross-effects (the extent to which raising marginal tax rates in some periods increases revenue in other periods on account of individuals' raising labor effort in all other periods due to the increase in the marginal utility of consumption).

The discussion to this point does not exploit systematic patterns in age-earnings profiles. As noted in subsection 5.C.1, Kremer (forthcoming) suggests that lower marginal rates on the young (particularly the very young, from ages 17–21) may be optimal. First, he offers evidence that their distribution of earnings is quite different: The ratio $(1-F)/f$ is much lower at low income levels because earnings are more concentrated there. (Compare the discussion of categorical assistance in subsection 7.C.1.) Second, he offers some evidence that labor supply elasticities are higher. Both factors indicate that lower marginal income tax rates on young low earners may be more efficient. Finally, there may also be some distributive benefit, which reflects the low correlation between early income and lifetime income. Although to a lesser extent, some of these factors may also apply to older workers (and to women and some racial minorities). Thus, independent of whether a higher or lower tax rate would be optimal in one or another year for a particular individual considered in isolation, the fact that age is a signal of the distribution of abilities and other factors implies that age-dependent taxation can raise social welfare.

Additional issues are presented by the introduction of uncertainty concerning earnings ability, utility, and lifespan. This problem, which is considered briefly in subsection 5.E.2 and in section C below, can now be imagined in a setting with many periods. Analyzing this case can be seen as encompassing unemployment insurance, disability insurance, medical insurance, and annuitization—the relevance of each depending on the availability of private insurance, as noted previously.

To examine how these various considerations relate to social security in particular, it is useful, as suggested in subsection A.1, to restate social security tax and benefit schemes as net taxes (or transfers, if negative) on labor income, which in turn can be viewed as part of the labor income tax per se. In the case in which benefits are a separable function of each year's earnings, this task is straightforward: the expression $T^N(wl) = T^S(wl) - B^S(wl)/(1+r)$ can be subscripted to refer to each period's earnings and taxes, where $B^S(wl)$ would refer to the component of ultimate benefits attributable to the corresponding period's earnings (and r could be restated to reflect the number of years of discounting required).

More generally, benefits may depend in a nonseparable manner on prior earnings. For example, they may be a nonlinear function of average earnings, or more weight may be given to years with higher earnings. In such cases, it would still be possible to state a function $T^N(wl)$ for each year, reflecting the difference between that year's social security–designated taxes and the net (present value) increment to benefits on account of that year's earnings—perhaps assuming hypothetically that the individual would have no future earnings, or perhaps instead that future earnings would be constant at the current level. Clearly, the definition of $T^N(wl)$ in each year (except the last) would not be unique; moreover, the function would now need to be stated as $T^N(wl,\theta)$ because, in general, the net tax (transfer) would also depend on prior years' earnings. These two points are interrelated: An individual in a given year, when choosing labor supply, would take into account not only current taxes but also how current earnings would affect future taxes. In a world of certainty with known, fixed future tax schedules and benefit formulas, it would not matter which of the nonunique specifications was chosen because, as long as the net (present value) tax (or transfer) as a function of any annual earnings pattern for the individual was the same, behavior, utility, and revenue would be the same. The relevant point is that, even when adding the complication of nonseparable benefits, one can view a social security system as tantamount to an adjustment to the labor income tax, in this case a time-dependent labor income tax that may depend on prior as well as present earnings.

It follows, therefore, that, just as in the two-period model in section 1, there is little meaningful that can be said about the optimal social security system with regard to income redistribution, now viewed in terms of lifetime income. The foregoing analysis suggests that the optimal income tax problem in this setting is complex. Whatever solution emerges, it does not matter in the present setting what part of that scheme, if any, is designated as the social security system. Even if one introduces myopia or liquidity constraints, considered in section B, there will be no inherent relationship between the optimal social security system and lifetime redistribution, for the extent of redistribution is determined by the combined scheme, including the income tax and transfer system. Any degree of redistribution that incidently is optimal in the social security system can be offset with the income tax.

Consider briefly some features of the existing social security retirement system in the United States (features shared, in varying degrees, by some other countries' systems). The use of a payroll tax (that is, a wage or labor income tax) that is constant over time might be viewed as a desirable feature because of the idea that time-invariant rates tend to minimize distortion. Qualifications to this view have already been noted, but in any event the description is inapt because it reflects an unintegrated view of the fiscal system in two important respects. First, the relevant marginal tax rate is not that of any specific tax but rather the aggregate net marginal rate from all taxes, including notably the income tax and phaseouts in transfer programs. With a nonlinear income tax-transfer system and income that varies over time, aggregate marginal tax rates are not constant.

Second, as subsection 1 emphasizes, the pertinent tax rate in viewing the social security system is not $T^S(wl)$ but $T^N(wl)$. Although the former applies a constant marginal tax rate (for earnings below the payroll tax ceiling), the latter does not because different periods' earnings have widely varying effects on future benefits. Some low-earning years contribute nothing to benefits (lowest-earning years are dropped), so $T^N(wl) = T^S(wl)$ in such years. But some high-earning years might contribute more to benefits than taxes paid, so not only does $T^N(wl) \neq T^S(wl)$, but we have $T^N(wl) < 0$ in those years. Thus, the implicit values of $T^N(wl)$ and therefore probably the values of $T(wl) = T^I(wl) + T^N(wl)$ vary substantially across years, with significantly higher effective tax rates applied in low-earning years (assuming that marginal rate graduation in the explicitly designated income tax is insufficient to offset the effect of varying $T^N(wl)$'s). Although the preceding analysis did not firmly endorse a presumption that marginal tax rates should be constant over time, in which case the existing scheme would be far from optimal, it also does not suggest that the existing pattern is likely to be appropriate. In particular, the direction of deviation from constancy, with higher marginal rates in low-w years typically involving very young workers, may not be correct. Furthermore, if one introduces uncertainty, it may be optimal to tax earnings in high-earning years at a higher rate than those in low-earning years, rather than employing the opposite pattern that is implicit in the current social security system.

Interestingly, although the pattern of rising marginal tax rates that is a common feature of observed income tax systems may not be optimal in a one-period setting, it may be beneficial in the present setting if it turned out to be optimal to tax individuals more in higher-earning years or if, as just discussed, social security systems viewed in isolation produce the opposite result. Nevertheless, an optimal overall system would allow taxes to depend on earnings history. Then there would be no need for the differentiation in marginal tax rates applicable to high versus low earnings across individuals in a given year to match the differentiation in marginal rates applicable to high versus low earnings by the same individual in different years.

Finally, it is instructive to consider how some of these considerations pertaining to the taxation of lifetime income relate to existing and proposed income and cash-flow consumption tax systems, which typically are based on annual earnings or consumption. If the system is linear, a constant marginal rate is applied in every period regardless of fluctuations in earnings (or expenditures), which as already suggested may not be optimal. In such cases, income averaging schemes—which, say, treat some income in high-earning years as though earned in low-earning years—are moot.

Furthermore, with a cash-flow consumption tax, to the extent that individuals smooth consumption over time, averaging also becomes irrelevant because, even with a nonlinear tax schedule, individuals would be at the same point on the schedule every year. In the hypothesized case, individuals use saving and borrowing to make it true in fact that consumption is equal every year, so there is no need for the tax system to undertake adjustments designed to treat individuals as if they had equal consumption every year.[9] The result is that individuals with different earnings patterns but the same present value of lifetime earnings—who in a no-tax world would enjoy the same utility and have the same marginal utility of consumption—will pay the same tax and face the same effective marginal tax rate on labor effort in each period (which, as noted, may not be optimal).

When a nonlinear, age-independent labor income tax is employed and wage rates vary over time (consider a simple, certainty case with a rising wage profile), individuals will face different marginal rates in different years. The efficiency consequences may be better or worse than with a constant marginal rate; they could be better, for example, if higher marginal rates apply to higher earnings and that indeed is optimal. Consider also that, under such a tax, individuals with different earnings patterns face different marginal tax rates in different periods and also will not generally realize the same level of utility or have the same marginal utility of consumption. In a commonly hypothesized case with no uncertainty, one individual is taken to have a constant lifetime wage and another a fluctuating wage with the same average level.[10] If marginal rates are rising (falling), the latter pays more (less) tax and also faces higher (lower) marginal tax rates when wages are high.[11] On distributive grounds, this outcome would not

---

[9]This statement, like the previous analysis in this subsection, abstracts from the effects of the interest rate and utility discount factors that, if not offsetting, would make the optimal consumption path nonconstant. In addition, in a nonlinear tax with falling marginal rates, individuals may fail to smooth consumption because unequal consumption over time would reduce the present value of tax payments. In either case, however, individuals' marginal tax rates on labor effort would be the same each period (because the allocation of incremental earnings, which determines the effective marginal tax rate, does not depend on the timing of the earnings).

[10]As the earlier analysis suggests, the individual with fluctuating wages would actually be better off. This individual would be equally well off if he worked the same amount each period as the individual with a constant wage (abstracting, as is done throughout, from interest rate effects), but the individual will choose to work more (less) when wages are high (low), thereby achieving a higher level of utility.

[11]Discussions of averaging usually consider the case of rising marginal rates, but falling rates may be optimal and, as chapter 7 notes, are common at lower income levels due to the phasing out of transfers.

appear to be optimal; the efficiency consequences are ambiguous, as already noted.

Vickrey (1939) proposed lifetime income averaging as a solution.[12] The general sympathy for this approach is due to distributive considerations. The effect of such a scheme on marginal tax rates and thus on labor supply distortion is not usually considered. On reflection, it should be apparent that averaging in a nonlinear income tax regime has a similar effect to self-averaging (consumption smoothing) under a nonlinear cash-flow consumption tax. In a simple world without uncertainty, the marginal dollar earned in any period is subject to the same effective marginal tax rate. That is, even if a current marginal dollar is taxed at a higher or lower rate, future adjustments will produce the result that the marginal distortion is the same in all periods. Once again, however, such uniform treatment may not be optimal.

3. Intergenerational Redistribution

Social security retirement systems in developed economies commonly operate largely on a pay-as-you-go basis rather than being pre-funded. Specifically, schemes were implemented and benefits were increased so as to provide to older living generations significant net transfers that ultimately must be financed by succeeding generations. For generations still alive, this ongoing intergenerational redistribution could be partially or completely reversed through benefit cuts. Likewise, it would be possible in theory to tax some generations to produce a surplus out of which benefits could be paid to subsequent generations, producing intergenerational redistribution toward younger and future generations.

The subject of the optimal distribution between generations is considered in subsection 14.B.2. It may be noted that, as a practical matter, it is difficult to identify the extent of intergenerational redistribution on account of the baseline issue examined in section 8.E (implicitly in the intragenerational context) with regard to the redistributiveness of the entire existing fiscal system. Notably, it has been observed that although social security has redistributed from younger and future generations to those retiring in the latter half of the twentieth century, those recipients had previously engaged in substantial implicit redistribution toward future generations by fighting, funding, and making other sacrifices during wartime, creating infrastructure, undertaking research, providing for younger generations' education, and so forth. Of course, the extent of such redistribution depends on whether expenditures were financed currently or through issuing debt, a subject to be considered further below.

The main point for present purposes is that, like those dimensions of redistribution considered previously in this section, the use of social security is not distinctive. Since the net social security tax is equivalent to an income tax schedule, income taxation could accomplish a similar result. This potential is most apparent if the income tax schedule is dependent on age or varies with birth cohort. In addition, even with an income tax schedule that in any given year depends only on current earnings, one could accomplish intergenerational redistribution by

---

[12]Under such a scheme, an individual's annual taxes are computed, for each year through the present, as if lifetime income to date had been earned evenly, and from (the present value of) this total tax obligation one subtracts (the present value of) all prior tax payments to determine how much tax is owed. A major complication involves changing family status over an individual's lifetime if tax schedules depend on family status, as they often do and, as chapter 12 indicates, they optimally would in general. For a discussion of other averaging schemes and of the merits of long-term averaging, see Goode (1980).

running current deficits or surpluses. In the short run, however, if one wished to redistribute more to a retired generation, something akin to a pay-as-you-go social security system would be required; specifically, benefits would have to be directed at current retirees. It should also be noted that, to the extent that the use of social security entails other effects, such as forced savings, it is useful to separate the differing objectives and utilize appropriate instruments. For example, if forced savings is undesirable (say, due to liquidity constraints), using the financing mechanism common for social security may not be the most efficient way to accomplish intergenerational transfers to existing retirees. Likewise, if forced savings is desirable but an intergenerational transfer is not, one could use a pre-funded social security system.

The direct efficiency costs of intergenerational redistribution should also be considered. To an extent, society could accomplish such redistribution without distortion, for example, by imposing uniform lump-sum taxes (equivalently, reducing $g = -T(0)$ in the income tax) on individuals in one generation and providing uniform subsidies (raising g) to members of another. However, if the optimal amounts paid and received depend on individuals' incomes, distortion would be involved in the ordinary fashion.

It is worth emphasizing that, despite the likely distortionary costs inherent in accomplishing whatever intergenerational redistribution is desired, intergenerational redistribution does not inherently raise questions of Pareto inefficiency. It is sometimes imagined that somehow everyone can be made better off, but such suggestions (and analyses) typically focus only on the steady state, ignoring the effect on transition generations.[13] The basic point is that, if society does wish to make a transfer to an existing, older generation, this payment must be funded in some manner. It can be paid for either by the current generation, presently, through reduced consumption (which would make them worse off), or through increased debt, which would make worse off the future generations who must then pay interest on the debt. If the debt could simply be extinguished, all future generations would gain, but obviously at a cost to those who held the debt.[14] As demonstrated by Breyer (1989), debt issued to finance the initial transfer cannot be retired without the generations who do so experiencing reduced consumption to that extent.[15] By analogy, an individual undertaking a spending spree will need to reduce future consumption to the extent of increased current consumption, and there is no way to save one's way out of the situation; one can save more in lieu of current consumption, which reduces utility from consumption presently.

Note that the absence of a Pareto improvement through pre-funding (just as the

---

[13]Compare İmrohoroğlu, İmrohoroğlu, and Joines (2003, p. 769 n. 23), who explain that their results on time-inconsistent preferences show unfunded social security to be less attractive than is found in prior work because they ignore the transition and examine only the steady state. This problem is analogous to that identified in subsection 10.C.1 of assessing the policy toward private (typically intergenerational) transfers by examining only the recipient generation or the steady state, ignoring opposing effects on the original donor generation(s).

[14]Likewise, one might benefit future generations by cutting benefits to current retirees, as suggested for example in Smetters (2005), an approach that in some respects is analogous to the use of a one-time capital levy. On transitions and capital levies generally, see subsection 9.C.2.

[15]See also the discussions in Diamond (2002) and Sinn (2000). The idea that moving to a higher-utility steady state is not ordinarily a pure matter of efficiency, due to the need for transitional sacrifices by some generations, was originally emphasized by Samuelson (1975).

impossibility of the individual raising utility from consumption in some periods without facing reductions in others) does not indicate that any given pattern is the social welfare (or lifetime utility) maximum. For example, it may be that societies with substantial unfunded social security retirement commitments (including for medical care) would benefit from increasing national savings (which may be suboptimal due to capital taxation or other factors).[16] In that case, increasing the extent of pre-funding—say, through some mix of benefit reductions and current tax increases—may be desirable, assuming that neither the government nor individuals would undertake offsetting actions, such as through increased government spending and reductions in other taxes or through reduced private savings.

Once again, however, it does not matter in theory whether this is accomplished through changes in social security financing or through other action, notably, raising current taxes or curbing present spending to reduce national debt. Also, as has been previously noted, it may not be best to use social security (for example, entailing forced savings that may or may not be optimal) to implement policies that can be accomplished independently. Some of the debate about social security reform seems to reflect the belief that one or another approach is more likely to be successful on account of political economy considerations; for example, it may be easier politically to raise taxes to fund social security than to run a surplus, or it may be that creating private social security accounts would make it less likely that the government would subsequently increase spending or cut other taxes, undercutting the attempt to increase national savings. Likewise, if individuals are myopic or their behavior otherwise deviates from that in the simple model employed in this section, otherwise equivalent actions may have different effects, which may bear on how social security should be formulated.

An additional intergenerational issue concerns risk-sharing. Given the incompleteness of futures markets, there is a potential role for government to spread risk across generations, as examined in Gordon and Varian (1988), Gale (1990), Shiller (1999), Campbell and Nosbusch (2006), and Krueger and Kubler (2006). Risk-sharing might be accomplished through social security if retirees' benefits are a function not only of their own prior earnings but also of earnings by adjacent generations. Nominally, this is not done, in which case the effects of pay-as-you-go social security systems can be perverse. Specifically, if the obligation charged to younger generations is fixed by the benefit formula based on the earnings of the retiree generation, then when earnings are systematically low, younger workers need to pay higher tax rates since the base is small while the obligation is unchanged. Likewise, when earnings are high, tax rates fall. In practice, however, de facto risk sharing may be accomplished if there is a tendency to increase retiree benefits (retrospectively) when current workers' earnings are high and to trim benefits when they are low. In any event, a social security system is not the only means of spreading risk intergenerationally, and greater analysis is necessary to identify how the optimal intergenerational arrangement depends on different generations' annual earnings and consumption.

---

[16]See, for example, Feldstein and Liebman (2002b). Observe that if savings are distorted, then in principle (in a model with identical individuals and other simplifications) a Pareto improvement is possible intragenerationally, through eliminating the distortion or otherwise achieving the results that would arise if the distortion was not present.

4. Redistribution Across Family Types

Social security systems may also redistribute across family types. In the United States, for example, spousal and other benefits result in substantial redistribution from single individuals to married couples and from two-earner families to one-earner families.[17]

As with other dimensions of redistribution through social security, it is helpful to separate the redistributive component ($T^N(wl)$ in the notation of subsection 1) and view it simply as part of the income tax. Here, however, benefit rules and thus T^N depend on family status, not just income. Nevertheless, the redistributive component can be assimilated into an income tax schedule that itself depends on family status, an approach that will be pursued in chapter 12 on taxation of families. It remains to consider whether social security's objectives are dependent on family status (for example, whether myopia leading to inadequate savings is a particular problem for married couples and whether the extent of any problem differs when there is only one earner). Nevertheless, even if they are, social security need not be redistributive on that account: The extent of forced savings could depend on family type, with future benefits funded on an actuarially fair basis. In any event, any distributive effect of social security can be augmented or offset through the income tax system to produce whatever overall distributive result is desired. Accordingly, the question of optimal redistribution across family types is largely separable in principle from that of the optimal design of social security.[18]

It is sometimes suggested that non-income-based intragenerational redistribution through social security is distortionary because effective marginal tax rates ($T^N(wl)$'s) differ significantly across individuals. It is generally correct that, ceteris paribus, distortion is greater when different individuals face different tax rates (since distortion rises disproportionately with marginal tax rates). If there is to be (income-based) redistribution between types, however, this cost is inevitable. As with standard redistribution, such distortion should in principle be traded off against whatever redistributive benefits are believed to result. Of course, if this sort of redistribution is believed to be undesirable, eliminating it would be doubly beneficial.

B. Forced Savings

Social security retirement provisions can be tantamount to schemes that force individuals to save a certain portion of their earnings to finance consumption during retirement. To focus on this feature, it is helpful to abstract from any redistribution and thereby consider actuarially fair systems, the only effect of which is to place a floor on savings. Such a floor is only interesting to the extent that it exceeds what (at least some) individuals would otherwise choose to save and that individuals do not offset the requirement through increased borrowing (either because that is impossible due to liquidity constraints or on account of aspects of their behavior that can generate inconsistencies, as will be mentioned).

This section first will analyze two primary rationales for forced savings—combating myopia and the Samaritan's dilemma—with particular attention to the effects of social security on labor supply and how those effects depend on the behavioral assumptions that may rationalize

---

[17]For explanations and documentation, see, for example, Boskin et al. (1987), Feldstein and Liebman's (2002b) survey, Feldstein and Samwick (1992), and Leimer (1999).

[18]As always, political economy considerations, perhaps reflecting misunderstanding of how the system actually operates, may influence what sorts of redistribution are incorporated in a social security scheme rather than in the income tax and explicit transfer programs.

forced savings.[19] Subsequent discussion will consider liquidity constraints, the importance of heterogeneity in savings behavior, and finally the relationship between forced savings and redistribution. A complete normative analysis of forced savings requires that other instruments also be considered. See, for example, subsection 9.A.2 on how myopia may bear on the optimal taxation of savings.

1. Myopia

It has long been suspected and recent work has investigated the possibility that individuals may not save adequately because of myopia. See, for example, Laibson (1996, 1997). Alternatively, if as Bernheim (1994), Diamond (2004), and others suggest, the complexity of the retirement problem combined with inexperience puts it beyond the reach of typical workers, a substantial fraction may err by saving too little (others might save too much, but that will not help those who save too little). Empirical evidence is mixed regarding the extent to which individuals' savings upon reaching retirement either are inadequate or would be so but for the forced savings through social security.[20] In any event, it is widely accepted that paternalistically motivated forced savings constitutes an important, and to some the most important, rationale for social security retirement systems.[21]

Assume that all individuals are identical and, on account of myopia, save too little for retirement. That is, in determining how much to consume, their decisional utility overweights the present. Savings is understood to be inadequate normatively because this weighting deviates from their utility as actually experienced. In such cases, the direct effect of forced savings, such as through social security, is to raise welfare by reducing this intertemporal misallocation of resources, and this beneficial effect will grow until the point at which the level of forced savings equals the level of savings individuals would have chosen if their savings decisions were rational.

This conclusion, however, ignores how forced savings would affect labor supply, a matter of particular concern since the labor income tax that finances forced savings exists on top of the distortionary labor income tax used for redistribution. A conjecture is that forced savings in the present setting would reduce labor supply: Considering a scheme that is actuarially fair,

---

[19]To simplify the analysis and focus on the identified issues, this chapter largely relies on a two-period model in which individuals work only in the first period, thereby abstracting from endogenous retirement decisions, which might be incorporated by adding one or more intermediate periods in which individuals may choose to work and considering various ways in which subsequent benefits could depend upon earnings in different periods. This problem is examined extensively in Diamond (2002, 2003) and in a series of papers by Diamond and Mirrlees, summarized therein and in Feldstein and Liebman (2002b) and also noted in section C.

[20]See, for example, Kotlikoff, Spivak, and Summers (1982), Banks, Blundell, and Tanner (1998), Engen, Gale, and Uccello (1999), Moore and Mitchell (2000), Bernheim, Skinner, and Weinberg (2001), Scholz, Seshadri, and Khitatrakun (2006), Aguiar and Hurst (2005), and Smith (2006).

[21]If myopic individuals are also subject to standard-of-living effects under which present consumption influences the utility of future consumption (perhaps by reducing utility but raising marginal utility at any level of consumption), as in Diamond (2003), the welfare benefit of forced savings may be greater.

individuals would excessively discount the future benefits whereas taxes are paid presently; hence, although in fact $T^N(wl) = 0$, individuals are imagined to behave as if $T^N(wl) > 0$. Furthermore, given the preexisting labor supply distortion due to explicit income taxation, the hypothesized behavioral response would generate substantial additional distortion.

It is important to examine this conjecture explicitly. To do so, the analysis follows Kaplow (2006b). Consider again a simple two-period model in which individuals work only in period one and allocate their after-tax earnings between the two periods. Two subcases will be distinguished: When individuals' labor supply decisions are subject to the same myopia that determines the allocation of consumption between periods, and when these decisions are rational in the sense that individuals not only understand that they will allocate their earnings myopically but appreciate what their realized utility will actually be (that is, that such an allocation involves too high a level of first-period consumption, c_1). Both cases are of potential interest because myopic behavior is not very well understood and is context specific. For example, some individuals employ commitment devices (automatic contributions to retirement accounts, not purchasing types of food they know they will overeat), many fail to borrow (fully or at all) from increased home equity despite their tendency to consume all of their paychecks, and savings behavior may be influenced by modest changes in framing (such as when individuals' contributions to 401(k) retirement plans depend on what contribution, if any, is specified by the employer as the default).[22] Note further that, in practice, the effect of myopia on labor supply may depend on the nature of the decision in question: Decisions about whether to pursue higher education or what job to choose from among many that require different effort levels may perhaps be made nonmyopically, whereas the same individuals may forgo overtime opportunities because of the immediate temptation to spend time with friends or watch favorite television shows. For simplicity, the analysis below focuses on the two pure cases.

To examine the effect of social security on labor supply, consider the following simplified model:

$$(11.3)\quad u(c_1, c_2, l) = \frac{{c_1}^{1-\rho}}{1-\rho} + \delta \frac{{c_2}^{1-\rho}}{1-\rho} - z(l),$$

where ρ is the coefficient of relative risk aversion (utility from consumption taking the constant-relative-risk-aversion form from expression (3.3), where it is understood that, when ρ = 1, utility from consumption c_i is instead given by ln c_i), δ is the actual subjective discount factor, and z measures the disutility of labor effort, where $z' > 0$ and $z'' > 0$. (To clarify, δ is taken here to be a real trait of individuals' utility, for purposes of assessing social welfare; myopia will be introduced separately below.) Individuals are subject to a linear income tax, so their budget constraint is

$$(11.4)\quad c_1 + \frac{c_2}{1+r} = (1-t)wl + g.$$

[22]On the latter, see Madrian and Shea (2001) and Choi et al. (2004).

To introduce myopia in a simple manner, suppose that, in allocating disposable income between c_1 and c_2, individuals behave as if they are maximizing the following variant of the utility function given by expression (11.3):

$$(11.5)\ u(c_1, c_2, l) = \beta \frac{c_1^{1-\rho}}{1-\rho} + \delta \frac{c_2^{1-\rho}}{1-\rho} - z(l),$$

where the weight on first-period consumption β is taken to exceed 1.[23] The first-order condition for consumption by a myopic individual (which can be determined by solving the budget constraint (11.4) for c_2, substituting it into (11.5), and differentiating) is

$$(11.6)\ \frac{\partial u}{\partial c_1} = \beta c_1^{-\rho} - \delta(1+r)c_2^{-\rho} = 0.$$

Because social security is assumed here to be actuarially fair, it has no effect on the budget constraint (11.4). The only effect of social security in this model, therefore, is to force savings, so social security can be represented as placing an upper bound on c_1. For concreteness, this bound will be a stated fraction of disposable income, so the social security policy may be denoted by χ, the minimum required fraction of savings.[24] Accordingly, there is now the

---

[23]It would also be natural to weight the disutility of labor, z, by β because labor is supplied in the first period; including such a weight, however, would not materially affect the results (β would weight the z' term in expression (11.8) and appear implicitly in d^2u/dl^2 in the denominator of expression (11.9); nothing else, including any of the interpretations, would change). Note also that, instead of weighting first-period sources of utility by β, one could weight second-period utility from consumption by a fraction less than one. The results would be nearly identical, the difference between these two formulations being in the cardinalization of utility as a function of consumption. (Recall that the discount factor δ does not already reflect such a downward weighting of second-period consumption because δ is taken to be the true subjective discount rate, a feature of the normatively relevant utility function.)

[24]Two other formulations of social security can be considered, one in which the bound is a fraction of earnings, wl, and another in which the bound is a fraction of after-tax earnings, $(1-t)wl$ (which differ from disposable income on account of g). For any given tax rate t, these two formulations are equivalent to each other, and the effect of changing the bound differs between the two cases only in magnitude because raising the latter bound (on after-tax earnings) one unit has a smaller effect than raising the former bound (on before-tax earnings) when $t > 0$. Analysis of these two cases yields results qualitatively similar to those of the case considered in the text. Note that including g in the quantity that is subject to forced savings, as is done here, has the effect that forced savings are a constant fraction of disposable income at all income levels, whereas the other formulations would make forced savings a rising fraction of disposable income (because the grant g, exempt from the forced-savings requirement in the alternative formulations, is a greater share of disposable income for low-income individuals). Alternatively, if the grant payment was divided between the two periods (and borrowing against it was

additional constraint

$$(11.7)\; c_1 \le (1-\chi)\left[(1-t)wl + g\right].$$

If the constraint is binding, this expression is satisfied as an equality, which in turn for any given level of l dictates the allocation of disposable income between c_1 and c_2.

a. Myopic labor supply. Consider the effect of social security, thus defined, on labor supply when the labor supply decision is also myopic in the sense that it is determined by maximizing utility as defined in expression (11.5) rather than as defined in expression (11.3).[25] The first-order condition for labor supply (when the forced-saving constraint (11.7) is binding) is

$$(11.8)\; \frac{du}{dl} = \beta c_1^{-\rho}(1-\chi)(1-t)w + c_2^{-\rho}\delta(1+r)\chi(1-t)w - z' = 0.$$

Differentiating this expression with respect to χ, rearranging terms, and making appropriate substitutions yields the following expression for the derivative of labor supply with respect to the forced-savings requirement:

$$(11.9)\; l_\chi = \frac{(1-\rho)(1-t)w\left[\beta c_1^{-\rho} - \delta(1+r)c_2^{-\rho}\right]}{d^2u/dl^2}.$$

To interpret this expression, note first that the denominator must be negative at the individual's optimum (and it can readily be shown to be strictly negative for all l in any event). Second, observe that the bracketed expression in the numerator equals $\partial u/\partial c_1$ from expression (11.6). Therefore, as the constraint just begins to bind, the effect on labor supply (in whichever direction it may be) will be negligible. (No matter how strong is the extent of myopia, individuals' consumption allocations decisions already reflect it; hence, at their unconstrained consumption optimum, they are indifferent to a marginal reduction in first-period consumption.) This result indicates that an actuarially fair tax on present disposable income to finance forced savings does not affect labor supply in a manner qualitatively or (in general) quantitatively similar to that of further raising the marginal tax rate on current earnings, even though individuals are assumed to be myopic.

---

impossible), these two alternative models would be even closer to the present model.

[25] Another variant would be to assume that individuals do not, when choosing labor supply, anticipate that they will misallocate consumption. In this case, however, they would not expect the social security forced-savings constraint to be binding (assuming that the constraint is not so strong as to force more second-period consumption than the nonmyopic optimum), so tightening that constraint would have no effect on labor supply. As a result, social security would raise welfare through improved consumption allocation and there would be no further effect on labor supply to be taken into account.

As the constraint becomes tighter (as χ increases once the constraint binds), the term in brackets, $\partial u/\partial c_1$, becomes (more) positive. The reason is that this derivative reflects individuals' *perceived* marginal utility from raising c_1 rather than their actual marginal utility. When the social security forced-savings constraint is binding, c_1 is less than what individuals would choose, so the perceived marginal utility of raising c_1 further would be positive. Accordingly, l_χ is negative (positive)—that is, tightening the forced-savings constraint reduces (increases) labor supply—if $\rho < 1$ ($\rho > 1$).

The intuition for this result regarding labor supply can be understood by decomposing two effects, indicated respectively by the "1" and the "$-\rho$" in the leading term $1-\rho$. A direct effect arises from more forced savings. When labor supply decisions reflect the same myopia as do individuals' first-period consumption allocation decisions, forcing an incremental reallocation of consumption toward period 2 is viewed as undesirable. Hence, the perceived return to labor effort falls.

An indirect effect is due to changes in relative marginal utilities of consumption in the two periods. As χ is increased, the consumption reallocation toward period 2 makes the marginal utility of consumption higher in period 1 and lower in period 2. Because first-period rather than second-period consumption is perceived as too low, this reallocation changes the (perceived) marginal utility of consumption more in the first period than in the second. (Stated precisely, the third derivative of utility as a function of consumption is positive, so the magnitude of the second derivative is greater when consumption is (perceived to be) low, as it is here in the first period, than when it is high, as in the second period.) When relative risk aversion is low, specifically, when $\rho < 1$, this latter effect is smaller than the direct effect due to consumption being perceived to be less well allocated between the two periods, so the overall perceived marginal benefit of increasing labor effort falls. However, when risk aversion is high, $\rho > 1$, the latter effect dominates so labor effort rises.[26] In other words, when $\rho > 1$, the fact that social security makes earnings seem less attractive is outweighed by the fact that the forced reduction in c_1 greatly increases the perceived marginal value of first-period consumption, which can only be raised, partially restoring it to its unconstrained level, by working more. When $\rho < 1$, this latter effect is present but is insufficient to outweigh the direct reduction in the value of consumption.

Further illumination can be gleaned by comparing l_χ to l_t, the derivative of labor supply with respect to the income tax rate. Sparing fairly tedious detail, as a very crude statement l_χ will for most values of ρ have the same sign as l_t, but its magnitude will tend to be closer to zero.[27] Recalling that $l_\chi = 0$ until after the point at which the forced-savings constraint begins to

---

[26]Formally, this condition and analysis are close to that offered in subsection 12.A.1 on the question of how generous allocations should be to a two-person family when resources are shared unequally. On reflection, this coincidence is unsurprising: With myopia, it is often stated that individuals behave as if there are two selves (a present self and a future self); in the case of unequal sharing in the family, there literally are two persons, one of whom is given more weight than the other.

[27]See Kaplow (2006b). The expression for l_t differs in a number of respects from that for l_χ. The sign of the former does not depend simply on the sign of $1-\rho$; instead, ρ is weighted in each period by a fraction less than one, namely the period's consumption minus the pertinent forced-savings share of the tax system's grant component g, all divided by the period's

bind, one can now see more fully the respects in which increasing forced savings affects labor supply differently from the way that raising the tax rate does.

b. Nonmyopic labor supply. These results may be contrasted with the case in which labor supply decisions are not myopic in the sense that they maximize utility as defined by expression (11.3) rather than expression (11.5), although labor supply decisions take into account that, when it comes time to decide upon consumption, the allocation will be given by expression (11.6), except to the extent constrained by social security, expression (11.7).[28]

The analysis of this case is straightforward from the above derivation, although the conclusions differ. The first-order condition for labor supply is again given by expression (11.8) and l_χ by expression (11.9), except that $\beta = 1$. The interpretation of (11.9) changes on account of the bracketed term. It still corresponds to $\partial u/\partial c_1$ from expression (11.6), now with $\beta = 1$. But when labor supply decisions are nonmyopic, and assuming that the social security parameter χ is in the range in which there is still overconsumption in period 1, the value of this expression is negative because c_1 has been raised past the point at which the nonmyopic first-order condition for consumption is satisfied.

One implication is that l_χ now has the same sign as $1-\rho$, rather than the opposite sign. That is, as more savings are forced, labor supply will rise (fall) if ρ is less (greater) than 1. As before, the intuition has two components. First, when χ is increased, the proceeds of labor effort are better allocated, which encourages labor effort. Because the labor supply decision is taken to be rational, the value of social security—as a substitute device making it possible for individuals de facto to commit to consume less in period 1—is positive in fact and is perceived as such. On the other hand, as χ is increased, the reduction in consumption misallocation changes the marginal utility of consumption in each period, making it higher in period 1 and lower in period 2. Because of the curvature of utility as a function of each period's consumption, the latter effect is greater. When $\rho < 1$, this indirect effect is less than the direct effect due to consumption being better allocated between the two periods, so the overall marginal benefit of increasing labor effort rises. However, when $\rho > 1$, the indirect effect dominates, so labor effort falls. (One can also, as above, compare l_χ to l_t in the present case. Sparing the details, it is crudely true that l_χ will for most values of ρ have the opposite sign as l_t, and its magnitude will tend to be closer to zero.)

There is another notable difference between the present case and that with myopic labor supply. There, $l_\chi = 0$ as the forced-savings constraint just began to bind, whereas here this is not true: At that point, the marginal gain from consumption reallocation toward the future, which is taken into account in individuals' labor supply decisions, is at its greatest, and thus forced

---

consumption. As a result, there will be a range of ρ somewhat in excess of one for which l_t is negative even though l_χ for the myopic case is positive. (See Chetty (2006) on how unearned income influences the values of ρ for which labor supply is upward sloping.) The tendency for the magnitude of l_χ to be closer to zero than that of l_t is due to the fact that, in the former case, the effects on each period's consumption are opposed to each other whereas, in the latter case, the effects on both periods' consumption work in the same direction.

[28]Paralleling the comment in note 25, one could also consider the case in which myopia is not anticipated when choosing labor supply, but then social security would have no effect on labor supply because the constraint would not be expected to bind.

savings affects labor supply nontrivially (except when ρ or β is close to 1) from the moment the constraint begins to bind. This factor and, accordingly, the effect of forced savings on labor supply will equal zero not when the constraint just begins to bind but rather when χ reaches the point at which the magnitude of forced savings just equals its optimal (nonmyopic) level (that is, when c_1 equals the value that satisfies the first-order condition (11.6), evaluated for $\beta = 1$). If χ were increased further, $\partial u/\partial c_1$ would reverse sign, becoming positive, and the sign of l_χ would reverse from whatever it had been when χ was lower.

To summarize, the effect of forcing additional savings on labor supply is qualitatively different in a number of respects from that of raising the (ordinary) marginal tax rate on labor income. Perhaps the clearest indication is that the sign of the effect (whatever it might be) is opposite for myopic and nonmyopic labor supply decisions. Additionally, in each case, the magnitude of the effect is determined differently. For myopic labor supply, the initial marginal effect is zero, whereas adding a small ordinary tax on top of a preexisting tax has a first-order effect. For nonmyopic labor supply, the effect on labor supply tends to fall as forced savings increases (rather than rising), reaching zero when forced savings equal the nonmyopic savings optimum.

An important part of the explanation for these differences between forced savings and taxation is that, even with significant myopia that leads to substantial overallocation of disposable income to c_1, the result is an increase in the actual and perceived marginal utility of c_2, sufficiently so that, at the unconstrained myopic optimum, the individual is indifferent between reallocations of consumption between the two periods (as noted above). Hence, forcibly reallocating some consumption to period 2 does not act at all like a tax, at least initially, even in the myopic labor supply case—and is viewed as a benefit in the nonmyopic labor supply case. Although these features change as the constraint tightens, they do not immediately vanish.

Finally, as a contrast with the foregoing analysis, consider briefly the possibility raised at the outset that individuals, rather than being myopic although nevertheless fully informed and calculating in their behavior, instead find the problem of intertemporal maximization too complex to solve effectively. To proceed further, it would be necessary to state explicitly how such individuals do behave, in particular when faced by a social security system. Perhaps some individuals overestimate and others underestimate the value of the benefits financed by their current social security tax obligations, in which case the distortion in labor supply due to income taxation would be reduced for the former and increased for the latter. Or, because benefits are far in the future, they might be undervalued fairly generally, in which case an actuarially fair social security system would tend, to that extent, to have the same effect as an additional tax.[29] (It is important to distinguish the previously analyzed case involving the discounting of future

[29]This possibility motivates proposals for providing taxpayers clearer annual statements explaining benefit accruals and for private accounts (which in the simplest case would be actuarially fair by construction and thus, one might suppose, would not be perceived as involving the imposition of any present effective tax). The latter and other forms of pre-funding have also been suggested as a response to perceived political risk that leads individuals not to believe that promised future benefits will actually be paid, on which see Dominitz, Manski, and Heinz (2003) (indicating that a substantial fraction of younger cohorts state that they do not expect to receive social security benefits upon retirement). See also Shoven and Slavov (2006), who compare political risk to investment risk.

utility from the present one of discounting future dollar benefits.[30]) Another possibility is that individuals who are unable to calculate for themselves behave as if the government has (approximately) solved individuals' optimization problems and chosen the forced-savings rate accordingly; in this case, raising forced savings in an actuarially fair system would tend to increase labor supply.[31] Or individuals might not appreciate marginal effective tax rates and respond instead to average rates.[32] Clearly, further empirical work on individuals' behavior is necessary to determine the actual effects of social security on labor supply.

2. Samaritan's Dilemma

Social security's forced retirement savings has also been rationalized by what Buchanan (1975) describes as the Samaritan's dilemma. See also Bernheim and Stark (1988), Bruce and Waldman (1990), Feldstein (1987), and Lindbeck and Weibull (1988). The notion is that individuals who anticipate receiving transfers during retirement if, but only if, they are sufficiently destitute will have an incentive to undersave.

As a simple model of this phenomenon, suppose that in period 2 the government will pay a transfer equal to the shortfall between an individual's available resources, c_2, and a modest target level of consumption, c^*.[33] That is, the period 2 transfer is $\max(0, c^* - c_2)$. Confronted by such a scheme, individuals who would receive any period 2 transfer would always receive the full c^*, for once in the range of eligibility for some transfer, savings are implicitly taxed at 100% so individuals would not save at all. Moreover, some individuals who would otherwise have saved such that $c_2 > c^*$ will also choose not to save anything. Specifically, all individuals will compare their utility given the levels of l and c_1 that would be optimal in a world with no period 2 transfers to their utility if they consume all disposable income in period 1, c^* in period 2, and work the level of l that would be optimal conditional on that plan. In the present simple model, there will be some critical level of w below which individuals save nothing and above which individuals save as they would if there were no transfers.

Because we are focusing on actuarially fair schemes, it follows that individuals whom the government anticipates will save nothing based on their level of earnings (that is, those whose

---

[30]In the latter case, if individuals were otherwise informed and rational, in the present model the effect on period 1 labor supply would be the same as that arising if one reduced the return to labor on the right side of the budget constraint (11.4). For example, if there were an additional tax of s on labor income and the proceeds financed actuarially fair benefits of which only $\alpha < 1$ were perceived, the right side of (11.4) would become $(1-t-s)wl + g + \alpha[swl(1+r)]/(1+r) = [1-t-(1-\alpha)s]wl + g$.

[31]The intuition is that higher forced savings reduces present consumption, which both raises the marginal utility of present consumption and also, in the hypothesized setting, raises the perceived marginal utility of future consumption. This case and others are explored in Kaplow (2007c).

[32]Liebman and Zeckhauser (2004) refer to this possibility as well as individuals' ignoring how present behavior affects future prices or tax rates as "schmeduling."

[33]The Samaritan's dilemma extends to private transfers, notably from family or charities. That is, individuals may undersave because they anticipate support, say, from altruistic relatives. Forced savings likewise may be effective in this circumstance.

earnings are below those of the critical type) would be charged $c^*/(1+r)$ in additional taxes.[34] (Although for individuals of very low ability, such a charge may seem too high to be optimal, keep in mind that any redistribution may be accomplished in the tax-transfer system; this degree of redistribution is taken to be unchanged in the analysis to follow.)

Now compare this period 2 transfer scheme to one that offers no transfers and forces individuals to save at least $c^*/(1+r)$ in period 1. Individuals who would have saved less than this amount in a world with no transfers or social security will be indifferent between this savings requirement and the transfers, for both regimes charge them $c^*/(1+r)$ in period 1 and allow them to consume c^* in period 2. For individuals who would not be subject to the tax and transfer, this constraint will not be binding (their chosen savings exceeds $c^*/(1+r)$ by a nontrivial amount), so they too will be indifferent.

However, for individuals in the intermediate range, who would have saved at least $c^*/(1+r)$ but whose w's are below the critical level, a forced-savings regime produces higher utility than does the tax and transfer regime. This utility gain arises because, under the transfer regime, these individuals were induced to consume less in period 2 than they would have found optimal in an unconstrained world. That is, their ideal level of savings exceeds $c^*/(1+r)$ but is not so high as to deter them from saving nothing if this will qualify them for the period 2 transfer of c^* (which, once they have dominion over their disposable income, is available to them for free, the additional tax of $c^*/(1+r)$ being a sunk cost when they make their savings decisions). The only effect of the transfer regime on them, given that it is actuarially fair, is to distort their savings behavior in the direction of too little savings. Switching to a regime of forced savings eliminates this distortion. Moreover, since by hypothesis these individuals would have saved more than $c^*/(1+r)$, the forced-savings requirement is not binding on them. Note that this benefit from substituting forced savings for period 2 transfers involves an efficiency gain (a weak Pareto improvement), for no redistribution is involved.[35]

Consider now the effect of this sort of forced-savings requirement on labor supply, compared to a regime with no social security and no period 2 transfers. For all but the group for whom the forced-savings requirement of $c^*/(1+r)$ is binding, there is no effect on labor supply. For individuals for whom the constraint is binding, period 2 consumption is simply c^* and all incremental after-tax earnings are spent on period 1 consumption, that is, $c_1 = (1-t)wl + g - c^*/(1+r)$. Accordingly, the first-order condition for labor supply is

---

[34]There is not a unique actuarially fair scheme. The government could, as suggested in the text, charge $c^*/(1+r)$ to all whose earnings are below those predicted for the critical type w in a world with no transfers. However, if the government also imposed the charge on individuals with slightly higher earnings (w's), such individuals then would find it optimal to switch between the two types of optima, choosing to save nothing and receive the transfer c^* in period 2; note that in so doing the result is actuarially fair for them. This complication about uniqueness does not alter the fundamental character of the analysis and thus will be ignored.

[35]Of course, in the present setting without myopia, forced savings is itself inefficient, assuming, that is, that the government (and others) could commit not to make the period 2 transfers. Note further that the analysis in the present setting is analogous to that in others, such as when Medicaid, free health care, or other transfers are available only to individuals with no remaining assets (see chapter 7, note ?) or when disaster relief or a tax deduction mitigates losses but only to the extent that they are not covered by insurance.

$$(11.10)\ \frac{du}{dl} = c_1^{-\rho}(1-t)w - z' = 0.$$

Differentiating this expression with respect to c^*, after some manipulation, yields

$$(11.11)\ l_{c^*} = \frac{\rho c_1^{-\rho-1}(1-t)w/(1+r)}{-d^2u/dl^2}.$$

The numerator is the marginal utility of consumption in the first period, where all incremental after-tax earnings are expended, weighted by the net-of-tax wage and divided by $1+r$ because c^* is denominated in period 2 dollars. Clearly, l_{c^*} is positive, the intuition being that tightening the constraint reduces first-period consumption, the effect of which is to raise the marginal utility of consumption and thus the benefit of increasing labor supply.

Alternatively, one might, as in subsection 1, consider a constraint like that in expression (11.7) in which forced savings rises with disposable income (perhaps the pull of the Samaritan's dilemma is greater when individuals' pre-retirement consumption was higher so that their fall in standard of living in the event of insufficient savings is greater). Again, the comparison is to a regime with no social security and no period 2 transfers. The first-order condition for labor supply is given by (11.8) and the effect of tightening the forced-savings constraint on labor effort is given by (11.9), each considered now for the case in which $\beta = 1$ (because in this subsection forced savings is analyzed under the assumption of no myopia). The interpretation of expression (11.9) parallels that in the case of myopic labor supply (because in that case individuals' consumption and labor supply decisions would, in the absence of social security, be governed by the same utility function, as is the case here). Specifically, recall that the bracketed expression in the numerator equals $\partial u/\partial c_1$ from expression (11.6), again now for the case in which $\beta = 1$. Therefore, tightening the constraint makes this term more positive as c_1 is pushed further below its optimal level. Thus l_χ is negative (positive)—that is, tightening the forced-savings constraint reduces (increases) labor supply—if $\rho < 1$ ($\rho > 1$), with the intuition decomposing the two effects paralleling that given previously.

Because in the present case no myopia is assumed to exist, there is no rationale for imposing forced savings at a level where the constraint is binding, except for fear of the Samaritan's dilemma, but the effects on labor supply in the preceding cases were by comparison to a regime without period 2 transfers. Although such a regime for remedying savings shortfalls is dominated by a pure forced-savings program in terms of the efficiency of consumption allocations, it is appropriate to consider the labor supply effects of this sort of transfer program as well. There are two sets of effects of raising c^* in such a scheme. First, for those who are already constrained, the effect on labor supply is positive and, indeed, is the same as in a regime of forced savings that finances c^* in period 2. (Recall that the difference between period 2 transfers and forced savings is that the former induces an additional, intermediate-ability group to reduce savings so as to become subject to the regime. Accordingly, this larger group is subject to the positive labor supply effects of raising c^*.) Second, as c^* increases, additional individuals will shift from their privately optimal levels of c_1 and c_2, as dictated by their first-order condition, to saving nothing beyond the implicit minimum savings mandate entailed by the

actuarially fair funding requirement of $c^*/(1+r)$. Their labor supply falls by a discrete amount. The reason is that the return to labor in terms of the marginal utility of consumption falls. Previously, the marginal utility of consuming in each period was equated, so it suffices to consider the marginal utility of period 1 consumption. After they shift regimes, if one supposes that labor supply remains the same, it must be that more is expended on period 1 consumption; hence, the marginal utility of consumption falls. Because the marginal disutility of labor supply is taken to be the same, the first-order condition does not hold and can only be restored by reducing period 1 consumption and labor effort.

In each of these regimes, therefore, there may exist costs or benefits due to labor supply effects, which can be significant given the preexisting labor supply distortion caused by redistributive labor income taxation. To assess the regimes, one must combine this consideration with the direct effects of the Samaritan's dilemma and its mitigation regarding individuals' allocations of consumption between periods in order to determine what sort of forced-savings regime is optimal.

3. Liquidity Constraints.

The existence of liquidity constraints and their possible relevance to the redistribution problem were noted in subsection 5.E.1. Liquidity constraints also bear on the need for and desirability of forced savings. See, for example, Hubbard and Judd (1987). Liquidity constraints may inhibit undersavings on account of myopia (see Laibson 1997) or the Samaritan's dilemma. To the extent that individuals cannot borrow against future earnings, they are implicitly forced to save, at least until the time at which such earnings are realized. On the other hand, forcing liquidity-constrained individuals not subject to any infirmities to save a minimum amount reduces their long-run well-being. Of course, in the absence of externalities or irrationalities, forcing more savings than otherwise would be chosen is always distortionary; however, with liquidity constraints, the initial (say, zero) level of savings is already too high, so the marginal distortion from forced savings is greater than otherwise. Moreover, even minimal savings requirements will necessarily be binding on individuals who already are liquidity constrained.

Additional considerations further complicate the problem of determining the optimal extent (if any) of forced savings in the presence of liquidity constraints. One difficulty is empirical: Even if one can identify liquidity constraints (for references, see subsection 5.E.1), it may not be apparent whether such constraints are forcing individuals to underconsume relative to the optimum or instead are preventing even greater overconsumption that would otherwise occur. (Heterogeneity, considered in the next subsection, adds yet another dimension.) Another issue concerns externalities. Greater consumption by workers when they are relatively young may produce positive externalities to other family members, particularly children, relative to what would be produced by consumption in retirement. See chapters 10 and 12, on voluntary transfers and on taxation of the family. On the other hand, the Samaritan's dilemma, which as mentioned in note 33 may also involve family members, suggests that raising retirement consumption, especially when it is otherwise too low, may generate positive externalities.

Liquidity constraints and related factors bear on the optimal timing, as well as level, of savings and thus on the optimal structure of any forced-savings program. Many individuals have hump-shaped or rising wage profiles during their working years although optimal consumption paths may be smooth; there exist early needs for funds to purchase consumer durables and housing (down payments) and also to finance investments in human capital; and consumption expenditures may produce greater benefits when children are present. Accordingly, it would be

optimal for many to engage in little or no savings or to borrow in early to middle-age years and to save substantial amounts (including debt repayment) in later years. Typical forced-savings programs, by contrast, tend to take a constant fraction of earnings every year.

In principle, however, forced-savings regimes could vary by age. Other factors could also be reflected in tax and benefit rules. For example, for different earnings levels or occupational categories that are known, on average, to be associated with different lifetime earnings profiles, different schedules could be applied. Also, forced-savings requirements could be adjusted over the life cycle to reflect variations over time in the number of dependents. If liquidity constraints and these other factors are significant, substantial welfare gains could result from even modest departures from constant contribution rates (keeping in mind that if, say, consumption was already distorted downward in early years due to liquidity constraints, small additional savings requirements would impose first-order losses).

It is also important to consider how such changes in forced-savings requirements would affect labor supply. One instinct is that constant contribution rates are presumptively ideal because they minimize distortion. However, as noted in subsection A.2, the tax burden on account of social security is not given by the stated contribution schedule, $T^S(wl)$, but rather the net, $T^N(wl)$, which takes into account how the year's earnings affect future benefits. (Specifically, it was noted that, in the United States, flat contribution rates result in highly uneven net tax rates, often with the highest rates when individuals are youngest and low or negative rates in peak earning years.) In this section, we have been considering only actuarially fair schemes. As emphasized in subsections 1 and 2, however, when individuals may be myopic and, in any event when their savings behavior is constrained, labor supply responses may differ (in either direction, depending on the behavioral assumptions and parameters) from those of rational, unconstrained individuals. That analysis would need to be extended to a model with more than one period of work and time-varying wages, utility, or myopia to analyze the question properly. The main point, already noted, is that since forced savings, not pure taxation, is contemplated, it is not obvious a priori that constant (contribution) rates would tend to minimize labor supply distortion.

Because forced-savings regimes may be especially appealing when individuals are myopic, it is useful to contemplate additional effects that myopia may have on labor supply in a model with multiple periods of work. One possibility is that individuals' early investments in human capital will be influenced disproportionately by net earnings in early years. (This effect could be a product of myopia or limited information.) In that event, lower tax rates when young may be optimal because, in addition to reducing the current-period distortion in labor supply, there would be the additional benefit of reducing the downward distortion in human capital, which affects the wage rate in subsequent periods. Another possibility is that labor effort is subject to habituation: What effort seems normal, how one learns to live one's life outside work, and so forth may become set or at least shaped by early experiences.[36] In that case, there also may be benefits to reducing taxation in early years.[37] (Note that the magnitude of such effects

---

[36]There are many possible channels. How exertion is experienced may itself reflect the extent to which effort deviates from one's norm. The enjoyment of some leisure activities is enhanced by investments of sorts; certain tastes are acquired and others remain undeveloped.

[37]The present argument depends on the uncompensated labor supply elasticity being positive. If it were zero, for example, lower present tax rates would not affect labor effort and

may be influenced by shortsightedness; to the extent that higher future taxes are not anticipated, individuals will be more inclined to make investments in lifestyle that are best suited to their current level of labor effort.) This argument, in contrast to the previous one, refers to taxes rather than to contributions, which as discussed may have different effects on labor supply. Accordingly, these factors may bear more on the optimal redistributive income taxation problem than on that of optimal forced savings. In the taxation setting, the present argument does offer a reason to deviate from constant marginal tax rates over time.

This subsection has considered liquidity constraints and related matters that are relevant to forced savings viewed from a lifetime perspective. For a more complete understanding, this analysis should be combined with that in subsection A.2, which considers how social security retirement provisions are connected to the redistribution problem (which is abstracted from in the present section) when it is viewed over the life cycle rather than in the conventional one- (or two-) period setting.

4. Heterogeneity

Heterogeneity is likely to be important with regard to myopia and liquidity constraints, as well as savings preferences that reflect different wage profiles over the life cycle and differences in family composition. Moreover, many of the effects of forced savings are nonlinear and even asymmetric, so identifying average tendencies is insufficient in the determination of optimal policy. As previously noted, savings requirements will force upward the savings of those who save too little but not force downward the savings of those who save too much (which would not include the myopic but may include those who err when attempting to solve the lifetime maximization problem). Among the former, labor supply responses have opposite signs for individuals whose labor supply decisions are also myopic and for those whose labor supply decisions are not. It was suggested that myopia may differentially affect different types of labor supply decisions (investment in human capital versus momentary overtime decisions), but obviously myopia may also vary greatly across individuals for the same types of decisions. The importance of liquidity constraints may depend on occupation, stage in the life cycle, family configuration, and other factors.

In trading off competing effects, the optimal forced-savings policy will also reflect variations in how marginal welfare consequences change with the tightness of the savings constraint. For individuals subject to myopia, the first dollar of consumption moved to the future will produce a first-order welfare gain, while the benefit falls to zero when the optimal intertemporal allocation is reached and becomes negative thereafter. As described in subsections 1 and 2, however, labor supply effects have different patterns. For individuals who are liquidity constrained—that is, assuming rational rather than myopic savings and borrowing behavior (and no externalities, notably, relating to the Samaritan's dilemma)—the first dollar moved to the future imposes a first-order welfare cost. For those who are not subject to infirmities or constraints, the first dollar of forced savings will not generally be binding and, as the constraint begins to bind (which it will at different points for different individuals, depending on preferences), there will initially be no first-order welfare loss.

---

thus would not affect habit formation, except that different average tax rates, or forced-savings rates, would influence present consumption levels, another channel by which habit formation may be influenced.

Because some of the possible costs of forced savings involve first-order effects even at the outset, it is possible that the optimum involves no forced savings. Indeed, in such a case, most plausibly produced by significant liquidity constraints, making available additional borrowing may well be optimal. Regarding all of the identified effects (except possibly some regarding labor supply), marginal costs rise and marginal benefits fall as the forced-savings constraint is tightened, a typical pattern. It is worth emphasizing, as suggested in subsection 3, that the optimal scheme may be age-dependent (and also a function of wage level, occupation, and family composition), perhaps with little or no forced savings for the young or those with children but with a positive savings requirement for others.[38]

5. Relationship to Redistribution

The present section has abstracted from any redistribution that may be embedded in forced savings through social security retirement schemes because, as discussed in section A, the redistribution problem is to a substantial extent separable in principle from other, more distinctive features of forced savings. Nevertheless, there are some considerations at the intersection of forced savings and redistribution.

First, as noted in subsection 2, if the Samaritan's dilemma takes a form under which the extent of second period transfers is substantial relative to the low first-period earnings of low-skilled individuals, then actuarially fair funding of a second-period consumption floor could leave such individuals destitute in the first period. Put another way, if individuals (through the government or private transfers) have considerable sympathy for the elderly poor, addressing this concern through second-period payments or through forced savings may in itself make it optimal to depart from actuarially fair finance unless transfers to low-income individuals in the first period are already sufficiently generous.

Second, the analysis of liquidity constraints and of heterogeneity in subsections 3 and 4 suggests that the optimal degree of forced savings may well depend, among other factors, on individuals' levels of earnings (perhaps adjusted for stage in the life cycle and occupation). For example, myopia may be a greater problem for low-income individuals and may even contribute to low earning ability if it affects decisions to invest in human capital.[39] To be sure, merely

---

[38]Externalities, notably arising in years in which children are present, introduce another dimension of heterogeneity that may favor differential forced-savings requirements for different family types. This feature of program design also interacts with the question of optimal redistribution among families, the subject of chapter 12. Nevertheless, the forced-savings issue has independent force: Holding the present value of redistribution across family types constant, forced reallocation of lifetime consumption by certain family types might raise social welfare. Note that a significant component of transfers that depend on the presence of children is the provision of free public education, a transfer that is received at a particular time in the life cycle and, because provided in kind, may well have the result of increasing the portion of a family's lifetime consumption (inclusive of government transfers of all types) that is expended on children. See section 7.E.

[39]Lawrance (1991) provides evidence that low-income, less-educated individuals appear to have a greater subjective discount rate, but evidence on consumption behavior does not distinguish between different underlying (true) preferences and differential susceptibility to myopia (or differential influence of the Samaritan's dilemma). The belief that undersavings is a

allowing the relative degree of forced savings to vary by income does not entail redistribution. However, due to heterogeneity and the presence of liquidity constraints, forced-savings requirements inevitably impose welfare losses on some individuals and, more broadly, affect individuals' marginal utilities of consumption and utility levels. As a result, the optimal degree of redistribution—more precisely, the optimal income tax schedule $T(wl)$—may well be affected by the problems examined in this section and the extent to which forced savings is employed to address them.[40]

C. Insurance

Social security schemes, in addition to incorporating various sorts of redistribution and forcing retirement savings, often contain additional features that involve insurance in some form. Annuitization, whether explicit or implicit (the latter through defined-benefit formulas), is often required. Annuitization tends to be desirable because of uncertainty over longevity. See Yaari (1965). Even when a bequest motive is present or incompleteness of other types of insurance makes precautionary savings rational, some degree of annuitization typically is optimal. See Davidoff, Brown, and Diamond (2005).

One rationale for mandating annuitization is the problem of adverse selection in annuity markets.[41] Another, emphasized by Diamond (2002, 2003), is that individuals may fail to appreciate the benefits of annuitization.[42] Furthermore, myopic individuals might consume their retirement savings too soon if not forced to annuitize. A different sort of justification for compulsory annuitization is that the Samaritan's dilemma is also applicable to the timing of

---

greater problem for low-income individuals may explain why some social security programs, such as that in the United States, provide higher replacement rates (levels of funded retirement consumption relative to lifetime earnings) for lower-income individuals, although redistribution offers another possible explanation.

[40]This possibility can be illustrated using a natural extension of the two-period model with earnings and labor income taxation in only the first period. Consider a model in which there are many periods and also overlapping generations, with the income tax-and-transfer schedule applicable in every period. With no forced savings and many individuals subject to myopia, there will be many retired individuals living largely on the transfer g ($-T(0)$ in the nonlinear income tax scheme), and many of them may have very low abilities and labor supply elasticities, which could favor a higher g but even higher marginal tax rates at low income levels (assuming that the tax schedule is not age-dependent). Compare the discussion in chapter 7. On the other hand, if the poorest individuals are not liquidity constrained because they live largely off transfers whereas a significant fraction of higher-ability individuals are liquidity constrained and a substantial portion of their earnings is subject to forced savings, such individuals' marginal utilities would be higher as a consequence, which would tend to favor a less redistributive scheme.

[41]See, for example, Eckstein, Eichenbaum, and Peled (1985). For empirical evidence on adverse selection in annuities markets, see Brown, Mitchell, and Poterba (2002) and Finkelstein and Poterba (2004).

[42]Diamond argues that existing opportunities for annuitization appear to be underutilized despite substantial utility gains and that individuals generally fail to annuitize at the optimal point in time, when young (when the extent of adverse selection may also be smaller).

consumption during retirement; that is, individuals who rapidly deplete their savings may expect subsequently to receive greater public or private transfers as a consequence.

To a substantial extent, the merits of forced annuitization are unrelated to the optimal degree of redistribution across income levels. Annuitization does entail some horizontal redistribution among individuals with different life expectancies. When such differences are unknown at the time of annuitization, this is simply the provision of insurance; when differences are known (the source of adverse selection), there is no Pareto improvement but still a social welfare gain from forced annuitization. Life expectancies may be correlated with income, in which case forced annuitization would redistribute, on average, across income groups, but contribution rates or benefit levels—or the background income tax and transfer schedule—could be adjusted to offset any such effects.

Likewise, some of the important factors relating to the desirability of annuitization, particularly adverse selection, are unrelated to those bearing on forced savings. Annuitization could be desirable when forced savings is not, and conversely.

As with forced savings, the effects of annuitization on labor supply have not been fully analyzed. Many similar considerations seem applicable. This is especially clear to the extent that myopia, misunderstandings, or the Samaritan's dilemma are relevant to annuitization. Even looking solely at adverse selection, the provision of otherwise unavailable insurance will in general affect the return (measured in expected marginal utility) to labor effort.[43] Although not a distortion considered in a vacuum, labor supply effects are relevant in light of the preexisting distortion due to redistributive labor income taxation.

Disability insurance is another significant feature of many social security systems.[44] Many of the same issues arise as with annuitization, including asymmetric information, myopia and other decisionmaking infirmities, and the Samaritan's dilemma.[45] These considerations in principle are largely independent of redistribution (the insurance can be actuarially fair). Forced savings for retirement is also largely distinct, as reflected in the fact that public disability insurance and forced retirement savings are typically run as separate programs even if administered by a common government agency.

There are, however, important synergies, particularly since some disabilities are permanent, inducing retirement of sorts, and the skills and disutility associated with labor effort in many occupations tend to degrade during the years individuals contemplate retirement. (Indeed, absent changes in utility or opportunities, many individuals might never retire.) This subject has received substantial attention, especially in a series of papers by Diamond and Mirrlees (1978, 1986, 2000).[46] It is understood that full insurance is not optimal because of moral hazard. Furthermore, it is necessary to carefully structure benefit levels for those not working and taxes and future benefits for those who continue to work in order to maintain work

---

[43]The direction and magnitude of the effect will depend on the curvature of individuals' utility functions.

[44]Unemployment insurance has similar characteristics, although the labor market search process and systematic risk may be more important than is the case with disability insurance.

[45]The prevalence of private disability insurance is substantially greater than private annuitization, suggesting that these problems may be less significant for many individuals.

[46]For summaries and further discussion, see Diamond (2003) and Feldstein and Liebman (2002b).

incentives over time. The failure to do so, especially in many European countries, is thought to contribute to the significant rise in earlier retirement. See Gruber and Wise (1999).

Another major feature of some social security systems, notably in the United States, is the provision of medical insurance for individuals beyond a certain age. In many other countries, the government provides health care for everyone, although one can conceptualize such systems as forced insurance paid for by workers currently, with supplemental payroll taxes constituting prepayment for insurance during retirement years. Due to the high and rising cost of health care, particularly for older individuals, medical expenditures for retirees are quite large. Such health care provision raises essentially all of the issues considered in this chapter pertaining to redistribution, forced savings, and insurance and thus can be analyzed similarly.