Links between benefit formula design and pension funding [*]

By

Salvador Valdés-Prieto[**]

Abstract

This paper presents a model that brings together two important labor decisions that are distorted by a mandatory pension plan: participation in the covered market, and the number of hours of work effort in the covered market. It is shown that the design of the benefit formula affects both distortions, while the financing method distorts the participation decision of the average member.

It is also shown that the savings distortions introduced by mandated savings may be empirically more important than distortions to the number of hours worked. This idea is backed by some Chilean data. Then our model is extended to show that as the iliquidity premium on mandated savings grows, the tax equivalent of the mandate approaches the full contribution, regardless of the marginal tax on hours and regardless of the financing method of the plan.

[*] Paper prepared for the IZA Conference *Pension Reform and Labor Markets*, May 19-21, 2001, Berlin, Germany. Financial assistance for travel and expenses is gratefully acknowledged to IZA.

[**] Professor of Economics, Catholic University of Chile, Santiago. svaldes@volcan.facea.puc.cl

1. Introduction

The distortions on the labor market caused by pension plans have received a lot of attention from economists, since it was discovered that an improved design of the benefit formula could achieve substantial welfare gains. These gains can be achieved without changing the financing method of the plan, i.e. while keeping pay as you go finance.

This paper explores the link between funding approach, labor market distortions and progressivity and redistribution. It shows that the design of the benefit formula affects both distortions and redistribution, in an unpalatable tradeoff. As the financing method affects the tradeoff, it might appear that there is a strong case for increasing the degree of funding.

However, it is also shown that Chilean data suggests otherwise, as labor market distortions persist despite full funding. The paper explores the idea that savings distortions introduced by mandated savings may be empirically more important than distortions to the number of hours worked. It is shown that as the iliquidity premium grows, the tax equivalent of the mandate approaches the full contribution, maximizing the distortion on hours worked and the distortion on participation, regardless of the financing method of the pension plan.

2. The basic model

In this simple model, we assume that the active phase of life is consolidated into a single period. This implies that we will be unable to distinguish between benefit formulae according to the way in which they combine contributions made in different years. We also assume that the passive phase of life is consolidated into a single period. This simplification precludes distinctions between plans that pay a single lump sum, and those that pay pensions, possibly with different risk profiles. It also precludes consideration of the distortions to the retirement age and the pension age caused by the mandate to contribute.

To simplify further, we assume that the individual expects to work a fixed number of hours in old age and earn a fixed amount of labor income. Thus, lifetime utility depends of the following three choice variables:

c_y = consumption when young, i.e. in the active phase of life.

c_o = consumption when old, i.e. in the passive phase of life.

n_y = time not worked in the labor market when young, or leisure taken when young. The total number of available hours per day is normalized at 1. Thus, working hours are $(1-n_y)$, and all those hours are worked in the covered sector.

Preferences are represented by $U(c_j, c_v, n_j)$, where U is a function which is increasing in all three arguments and meets the standard conditions of utility theory.

Regarding the pension plan, we assume initially that all earned income is covered by the plan[1]. Another important aspect of the plan is the benefit formula. The pension P may be a function of A, the contribution actually made by the individual in the past, or a function of the number of years of service N and the average of past taxable earnings. These options describe the two main types of benefit formulae, which in our simple setting are:

(1a) $P = A\cdot(1+\rho) = \theta\cdot w\cdot(1-n_y)\cdot(1+\rho)$ actuarial formula

(1b) $P = k\cdot E = k\cdot w\cdot(1-n_y) = (k/\theta)\cdot A$ years of service formula

Where: ρ = real rate of return at which contributions accumulate in act. formula.
$k = a\cdot N$ = product of an actuarial factor (a) and the years of service (N).
$E= w\cdot(1-n_y) = A/\theta$ = labor earnings.

In this simple setting, without uncertainty and a single period, the two benefit formulae are perfectly equivalent when $1+\rho = a\cdot N/\theta$ (compare (1a) and (1b)).

The pension P may also be a function of labor income relative to the national average of labor income, as in plans that attempt to redistribute wealth through the benefit formula. However, due to the existence of other instruments that can be used to redistribute wealth in an individualized basis, such as the tax system (personal income taxes and estate taxes exhibit progressivity) and government expenditure programs targeted to help the poor (assistance pensions and income support also exhibit progressivity), we assume that the pension plan is not used to redistribute wealth[2].

The period budget constraints are:

(2a) $c_y = w\cdot(1-n_y) - A + F_0 - F_1$ where

(2b) $c_o = y_o + F_1\cdot(1+r) - F_2 + P(n_y)$

where:

w = wage rate in the active phase, per hour worked. Earnings E is $w\cdot(1-n_y)$.

---

[1] This will be relaxed below.

[2] We are careful to distinguish between attempting redistribution and achieving it. Detailed country studies have shown that many apparently progressive pension plans achieve little or no actual redistribution (for the Netherlands, see Nelissen 1998).

$A = \theta \cdot w \cdot (1-n_y)$ = mandatory saving into the pension plan.

θ = contribution rate (total, including employer and worker contributions).

F_0 = inherited wealth at the beginning of the active phase.

F_1 = voluntary saving = wealth at the end of the active phase.

y_o = expected labor income earned when old, a fixed value.

r = real interest rate earned by voluntary saving, at the beginning of period 2.

F_2 = planned bequest, to be paid at the end of the passive phase.

$P = P(n_y)$ = pension amount. It is a function of the taxable income earned by the individual in the past, when young, and thus, of leisure taken when young.

Eliminating F_1 from equations (2), the period budget constraints reduce to an intertemporal budget constraint:

$$(3)\quad PVC(n_y) = \left\{ w \cdot 1 + \frac{y_o}{1+r} + \left[F_0 - \frac{F_2}{1+r} \right] - w \cdot n_y \right\} \quad + \quad \left\{ \frac{P(A(n_y))}{1+r} - A(n_y) \right\}$$

where: $PVC \equiv c_y + c_o/(1+r)$ is the present value of lifetime consumption.

Equation (3) shows that the opportunities for lifetime consumption and leisure when young depend of at least three factors:

(i) the individual's wealth or endowment, shown in the first parenthesis. This is the potential human capital if he took no leisure when young, plus the bequest received, minus the bequest left.

(ii) the wealth effect of participating in the pension plan, shown in the second parenthesis. This term may be positive, zero or negative, depending of the benefit formula and of the parameters of the benefit formula, such as ρ or $a \cdot N/\theta$.

(iii) The value of the applicable interest rate, about which we say more later.

The value of both parentheses depends of the amount of work/leisure taken when young, n_y. We emphasize the second dependence, i.e. that the wealth effect of participating in the pension plan depends of the size of contribution A or of earnings E, which in turn depends of the amount of work and leisure chosen when young. The benefit formula $P(\cdot)$ determines the size of this wealth effect. This wealth effect may also be described as the net tax associated to participation in the pension plan.

The traditional method of representing the worker's options in (c_y, c_o) space, is unable to represent choice in this setting, because the map of indifference curves in this space shifts continuously as the hours of leisure change. To produce a bidimensional graphical analysis for this choice problem, it is necessary to reduce the three choice variables c_y, n_y and c_o to just two. To do this, we make an additional assumption: consumption preferences are homothetic, in the sense that the preferred ratio of consumption when old to consumption when young depends solely of the interest rate and the wage rate, but not of wealth. This is somewhat restrictive, as homotheticity fails to apply in important cases, such as workers that are consuming near their perceived subsistence levels (Attanasio and Weber, 1995). However, for other workers it is a natural approximation. Homotheticity in consumption is defined as[3]:

$$(4)\ \frac{c_y}{W}=\mu_1(r,w)\quad \forall W \quad , \quad \frac{c_o}{W}=\mu_2(r,w)\quad \forall W \qquad where\quad W\equiv c_y+c_o/(1+r)+w\cdot n_j$$

Equation (4) implies that $c_o/c_y = \lambda(r,w)$, a ratio that does not depend of wealth. From the definition of PVC it follows that $c_y = PVC\cdot(1+r)/(1+r+\lambda(r,w))$, where PVC is the present value of consumption. PVC is a consumption bundle of fixed composition for given factor prices. Replacing these definitions in the utility function:

$$(5)\ U(c_y,c_o,n_y)=U\left(\frac{(1+r)\cdot PVC}{1+r+\lambda(r,w)},\frac{\lambda(r,w)\cdot(1+r)\cdot PVC}{1+r+\lambda(r,w)},n_y\right)=V(PVC,n_y,r,w)$$

Equation (5) shows that for these preferences, we can draw indifference curves in (PVC,n_y) space, which has only two dimensions. The indifference curves in V have the usual concavity properties in (PVC,n_y) space provided that some conditions are met: the cross partial U_{12} must not be very positive and U_{32} must not be very large in absolute value. We will assume these conditions are met. Of course, the map of indifference curves in V remains fixed only when factor prices remain fixed. This allows graphical analysis as shown by figure 1.

Figure 1: The budget constraint for a given benefit formula

The budget constraint in figure 1 shows that in the absence of a pension plan, the individual can increase lifetime consumption by working more hours and reducing leisure. Participation in a pension plan adds the term $(P(n_y)/(1+r) - A(n_y))$ to the level of lifetime consumption at each level of hours worked.

When the worker chooses an interior optimum, the slope of the indifference curve is equal to the slope of the possibilities curve. The assumption that preferences are

[3] According to Kreps (1990), p. 50, the standard definition of homotheticity is $\partial(x_j/Y)/\partial Y = 0$ for all goods j, where Y = total income and x_j = demand for good j at priced p and income Y. This condition implies equation (4).

homothetic implies that when joining a pension plan that has a fixed negative wealth effect that is independent of hours worked, the increase in hours worked and the reduction of consumption occur at the same ratio. This ratio does not depend of the initial level of wealth. However, if the wage rate increases, the map of indifferent curves shifts and the slope of the budget constraint increases. If the interest rate changes, the map of indifferent curves shifts and the level of the budget constraint falls because the present value of future labor income falls (provided planned bequests are relatively small).

The slope of the budget constraint in figure 1 is obtained by differentiating (3) with respect to the hours of work, and using $A = \theta \cdot w \cdot (1-n_y)$:

$$(6) \qquad \frac{\partial PVC}{-\partial n_y} = \frac{\partial PVC}{\partial (1-n_y)} = w \ - \ \theta \cdot w \cdot \left\{ 1 - \frac{1}{1+r}\frac{\partial P}{\partial A}(A^*) \right\}$$

The first term of the right hand side recognizes that working more hours raises labor income at rate w. The interesting term is the second one, which takes into account that an additional hour of work also raises the amount of contributions and the amount of pension benefits. The parenthesis shows the net present value of each additional dollar contributed to the pension plan. This net present value may be positive or negative depending of the benefit formula and the discount rate.

In equation (6), the second term is the *change* in the wealth effect associated to participation in the pension plan that occurs when the taxable salary increases due to more hours of work. The marginal relationship between contributions and benefits $\partial P/\partial A$ governs the size and sign of this change. This change can create a tax or a subsidy on hours worked[4]. This marginal tax is independent of the level of the tax associated to participation in the plan.

Let us review several possible benefit formulae:

Figure 2: Possible benefit formulae

The straight lines AB and CF show the possibilities for consumption and leisure with two benefit formulae that have a common characteristic: the only impact of working an additional hour is to increase the present value of consumption by w, which is the first term of (6). The difference between these two benefit formulae is the *level* of the wealth effect associated to participating in the pension plan. This wealth effect is zero with the benefit formula represented by AB.

[4] This distortion became well-known in the economics literature at the beginning of the 80's . Its importance was stressed by Auerbach and Kotlikoff (1987).

The wealth effect in benefit formula ACF is equal to a subsidy AC to all participants. An example of ACF is a universal flat pension financed with general taxation, such as a consumption VAT, as in the first pillar pensions of New Zealand and Denmark. If the benefit is granted as a function of residence status only, this plan does not tax the decision to participate in the covered labor market. The benefit formula does not tax nor subsidize additional hours of work. Of course, the VAT taxes the additional hour of work, as long as the extra labor income is eventually consumed[5].

An "actuarially fair" benefit formula is defined as one where the allocation of hours to paid work is unaffected by the presence of the pension plan. AB is actuarially fair according to this definition. Therefore, benefit formula AB also represents the consumption/leisure opportunities in the absence of a mandate to participate in the plan. The Chilean pension plan, appears to be an example of AB, because it pays market rates of return to its members. We will argue otherwise below.

In addition, it is useful to define "actuarial fairness at the margin". The requirement for this is that an additional dollar of contributions generates an increase in expected benefits of $1 plus interest, at the same interest rate that the worker can obtain in the financial market. In this simple model, the requirement for actuarial fairness at the margin is $\partial P / \partial A = (1+r)$. Benefit formula ACF is actuarially fair at the margin only.

ADE represents a benefit formula that is not actuarially fair: benefits are a lump sum (line AD), regardless of the size of the contribution. This is the case of many universal first pillar pensions, such as the Argentinean Basic Pension. The slope of section DE is not zero even though extra contributions do not increase pensions at all (i.e. $\partial P / \partial A = 0$). As shown by (6), when $\partial P / \partial A = 0$, one more hour of work still allows the worker to increase consumption by $w \cdot (1-\theta) > 0$. Even with this benefit formula, most of labor income is not captured by the pension plan and can be spent in higher lifetime consumption.

Curve AG shows a benefit formula where benefits increase rapidly with the first additional contributions, but later increase ever more slowly. This decline implies that the initial contributions get a marginal subsidy but the latter suffer a substantial marginal tax, which grows with the level of covered income. This may represent benefit formulae that attempt to redistribute wealth from high earners to low earners.

The lesson of this simple model is that when analyzing a pension plan, economists should make a distinction between taxes on coverage or participation (wealth effects) and marginal taxes on hours worked. The distinction is obvious in figure 3: that worker gets a subsidy of the first type and pays a tax of the second type.

[5] This tax distorts decisions such as how much time is devoted to home production and child care, including production is small agricultural holdings and in home premises, as in many poor countries today, as compared to production for sale in formal and VAT-taxed markets.

Figure 3: Taxes on coverage and on hours worked.

Curve AG in figure 3 shows a benefit formula and worker V that obtains a net subsidy for participating in the pension plan, because the level of his preferred point is above line AB, which is the actuarial fairness line. At point M, the size of the wealth effect (subsidy) is MN, measured in present consumption units. Simultaneously, the *marginal* relationship between benefits and contributions at that point is a tax to additional hours of work in the covered sector. We know this because the *slope* of the indifference curve at M is smaller than the slope of the actuarially fair line AB. With the same benefit formula, a worker with preferences like U pays a tax on both counts.

3. The benefit formula, the financing method and redistribution

The polar methods to finance a pension plan are full funding (prefunding) and pay as you go. This section examines the statement that it is possible to eliminate marginal taxes even if the financing method is pay as you go, or for that matter, exhibits any degree of partial funding[6]. This statement is important for pension policy, because elimination of marginal taxes is desirable, and it were true, it could be achieved in partially funded plans without facing the distortions created by the high general tax rates associated to a transition to full funding.

However, we show that the financing method imposes substantial restrictions on the benefit formula. The relationship between contributions and benefits for a particular individual may be chosen with some freedom when designing the benefit formula, but at the aggregate level there exists a financial restriction, set by the financing method.

We insert the two-period lifecycle model of section 2 in an overlapping generations setting. The internal rates of return paid by the pension plan are defined by:

(7a) $1+\rho_i = P_i/A_i$ for individual i, and

(7b) $1+\rho_A = P_A/A_A$ for the average member of his generation

Where $A_A = \theta \cdot w \cdot (1-n^*_j)$ = contribution of the representative worker.

$P_A = P(\ \theta \cdot w \cdot (1-n^*_j)\ )$ = benefit of the representative worker.

[6] See for example Auerbach y Kotlikoff (1987, p. 154-161). They suggest that the marginal distortion can be eliminated choosing the interest rate credited by the benefit formula to be equal to the rate of return offered by the financial markets.

Now we review the link between the average rate of return and the financing method. In order to preclude increases to the internal rate of return for the current generation financed by drawing down the pension funds, and thus confiscating future generations of members, we only consider plans where the degree of funding is kept constant over time. Let us define:

$$(8) \qquad \gamma_t \equiv \frac{F_t}{PV_t(N_{ot+1} \cdot P_{t+1})} = \frac{F_t}{(N_{ot+1} \cdot P_{t+1})/(1+r_t)}$$

where γ = degree of funding at the end of period t.

F_t = the financial assets F of the pension plan at the end of period t.

PV_t = present value of the liabilities of the pension plan, at the end of period t, promised to the generation that will be retired next period.

N_{ot+1} = number of pensioners ("old") in the next period.

P_{t+1} = average pension in the next period.

r_t = discount rate applicable between periods t and t+1.

From the cash flow identity in an overlapping generations economy where GDP, the wage bill and all the other aggregates grow at a constant rate g, and where the contribution rate, the degree of funding and the real interest rate earned by pension funds are constant, it can be shown that ρ_A is the harmonic mean of r and g:

$$(9) \qquad \frac{1}{1+\rho_A} = \frac{\gamma}{1+g} + \frac{1-\gamma}{1+r}$$

As dynamic efficiency requires r > g, equation (9) implies that $\rho_A \leq r$. Equality occurs at full funding only ($\gamma = 1$). Specifically, ρ_A moves between g and r as the degree of funding γ increases from 0 to 1. The average internal rate of return offered by the plan to a single generation is given by the degree of funding γ, g and r.

The financial equilibrium of the pension plan requires that the desired point chosen by the representative worker in the consumption/leisure possibilities curve, be such that he generates levels of contributions and benefits that comply with (7b). Because of this, equation (7b) restricts the functional form of the benefit formula P(A).

We replace (7b) in (1a) and this in (3), yielding the following budget constraint at the point chosen by the average worker:

$$(10) \quad PVC(n_y) = \left\{ w \cdot (1-n_y) + \frac{y_o}{1+r} + \left[F_0 - \frac{F_2}{1+r} \right] \right\} \quad + \quad \left\{ \frac{\rho_A - r}{1+r} \right\} \cdot A_A$$

This is the level of lifetime consumption at the point where the average worker chooses his hours of work. When the plan is fully funded, then $\rho_A = r$. In equation (10) the second term is zero, and thus there is actuarial fairness. Thus, we have shown the well known result that a fully funded plan can replicate the financial market, in the sense of avoiding wealth effects, i.e. avoiding taxes or subsidies on participation in the labor market it covers.

However, as the degree of funding falls, the level of lifetime consumption falls according to equation (10), because the last term of the right hand side becomes negative. Thus, the plan must deviate from actuarial fairness, in levels, because a negative wealth effect exists. There exists a tax on participation in the covered labor market. This is valid regardless of the shape of the benefit formula, and is valid only for the point chosen by the average member of the plan.

The previous section showed that taxes on participation are unrelated to taxes on marginal hours of work. We have said nothing about the latter, because the shape of the benefit formula controls them.

Consider figure 4, where line AB represents actuarial fairness. Line AC represents aggregate financial balance or equation (10), and thus depends of the degree of funding. If the representative worker had preferences such that his optimal point is Q, the pension plan would have a permanent financial deficit, because aggregate contributions are smaller than aggregate benefits. This is shown by the fact that Q is above line AC. Therefore, point Q is not viable at the aggregate level.

Figure 4: Aggregate financial restriction for the representative worker

Curve AQRESD in figure 4 represents an interesting benefit formula. It is financially viable because the representative worker chooses point E on it, and point E happens to belong to line AC, which represents aggregate financial balance. But in addition, it turns out that the slope of curve AQRESD *at the point chosen by the representative worker (E)* is the same as the slope of line AB. Therefore, the slope of AQRESD is actuarially fair at the margin at E and there is no marginal distortion of the consumption/leisure choice of the representative worker.

Moreover, this benefit formula is not so special after all. Starting with the indifference curves of the representative worker and the actuarially fair slope AB, it is always possible to draw the locus of consumption/leisure choices where choice of hours worked is not distorted. On the other hand, financial viability and the degree of funding set the slope of line AC. Where both lines cross we set the desired location of E. Then any benefit formula that starts at A, passes through E with the slope of line AB[7], and has

[7] This can be done both with actuarial benefit formulae and with years of service formulae.

a general shape such that the best choice for the worker is E, eliminates the marginal distortion to work effort.

As this can be done for any degree of funding, it seems possible to decouple the choice of financing method from the choice of the benefit formula, as proposed by Auerbach and Kotlikoff (1988). In other words, the financing method would not be linked to the marginal distortion on labor effort created by the benefit formula.

This does not deny that the financing methods where $\rho_A < r$, such as pay as you go, create a negative wealth effect. This effect distorts the decision to participate in the jobs covered by the pension plan. Thus, the financing method remains critical for actuarial fairness and the distortion on participation.

However, figure 4 also suggests that this argument is subject to an important caveat: The members of any pension plan exhibit substantial diversity in their preferences over consumption and leisure. At a minimum, they confront different wage rates and different rates of return on saving at the margin. More deeply, they have different tastes. Thus, when members choose their most preferred point in the benefit formula AQRESD, many will choose a point outside of section RS, and therefore their consumption/leisure decision will be distorted at the margin.

It may appear that this objection may be faced by adjusting the shape of curve AQRESD, i.e. the benefit formula, to extend the range of it in which segment RS applies, to avoid distortions for all possible levels of leisure and consumption. This approach is equivalent to impose the following condition on the benefit formula:

(11) $$\partial P / \partial A = 1+r \quad \text{(r is the interest rate earned on voluntary saving).}$$

As r is independent from P and A, it is possible to integrate (11), from which:

(12) $$P^*(A) = (1+r)\cdot A - K$$

Where $P^*(\cdot)$ = the benefit formula that avoids marginal distortions.
K = constant of integration. It is a lump sum adjustment to benefits.

Although the appearance of (12) is of an actuarial formula, a benefit formula based on years of service can achieve the same shape, and avoid the marginal distortion, by choosing $a = \theta\cdot(1+r)/N$. However, in a multiperiod setting and with uncertainty this equivalence fails, and the years of service benefit formulae distorts the number of hours worked more than actuarial formulae with a fair interest rate.

Another implication of equation (12) is that to eliminate the distortion on hours of work, the interest rate used by the actuarial formula has to be the same as that offered by the financial market. This can be done in fully funded plans. However, it is a difficult requirement for notional account systems such as those introduced in Italy, Poland and Sweden since 1995. These plans use an actuarial benefit formula (see 1), but it is not fair because it does not pay the financial market interest rates r, but another (lower) interest rate linked to ρ_A (< r). As those formulae do not meet condition (12), notional account plans cannot avoid the marginal distortion on hours worked.

Equation (6.A.4) has a constant of integration. To find its value, we use the requirement of aggregate financial balance for the plan. According to (7b), aggregate financial balance assures that for the representative worker, $P_A = A_A \cdot (1+\rho_A)$. Replacing in (12), we find that:

$$(13) \qquad K = (r-\rho_A)\cdot A_A = (r-\rho_A)\cdot\theta\cdot w_A$$

Equation (13) says that when the financing method is full funding, K = 0. However, in the other financing methods K > 0. But this implies that benefits must be subject to a *lump sum reduction*, if the marginal distortion on hours worked is to be avoided.

Figure 5: A formula with no marginal distortion financed with less than full funding

Figure 5 shows that to avoid marginal distortions and achieve a slope like AB for all the values of the benefit formula, segment RS of figure 4 must be extended to the full range of values of leisure. However, to make this compatible with aggregate financial equilibrium when financing is not full funding ($\rho_A < r$), a lump sum tax on benefits is required, whose present value is given by segment AT. The size of segment AT, is the present value of equation (13), i.e. $\theta\cdot w_A\cdot(r-\rho_A)/(1+r)$.

An alternative way to present this design, more amenable to notional accounts, is as follows: the notional accounts are credited every year with a market rate of return, but the state imposes a negative initial account balance on all young members when they join the plan, of size AT.

The actual size of this negative initial account balance is of interest. For a notional account plan whose internal rate of return is 2% real per annum (it may be partially funded), in a financial market that yields 4% real per annum, where the contribution rate is 15% on average taxable wages of 2,500 dollars per month, the initial account balance AT is -32,453 dollars[8]. This allows the notional account plan to credit the financial market rate as the notional account, so the hours of work are not distorted by the plan. This is equivalent to 7.2 years of contribution.

Of course, such a lump sum tax distorts the decision to participate in the jobs covered by the plan. The lump sum tax looms heavier for less well-off workers, for members that work part time, and for members that interrupt their labor market participation to work at home, i.e. raising children. Of course, that outcome would be grossly regressive. Moreover, some of the members that do take jobs covered by the plan would reach old age with very low balances, or even with a negative balance. As that debt would not be collectable in practice, the pension plan would be thrown out of financial balance.

This criticism does not apply to actuarial formulae only. The method used by traditional years of service formulae to incorporate lump sum tax on participation has been to establish a minimum number of years of contribution as a requirement to obtain positive benefits. Workers that do not meet this requirement say because of long spells of work at home or in the informal sector, are denied all benefits, despite their positive contributions to the plan. Thus, the lump sum tax implicit in years of service formulae has been regressive as well.

Tradeoffs with equity

These results suggest the following tradeoff: A design of the benefit formula that eliminates the marginal distortion (on hours of work) for all levels of work, *redistributes* the aggregate tax on participation towards those that work few hours covered by the plan. This happens because the participation tax is reformulated to take the form of a lump sum. This is regressive.

In the efficiency front, this new distribution of the aggregate tax on participation may exacerbate the participation distortion in absolute amounts, because the members that work few covered hours may be more willing to avoid participation than high salary workers that work a large number of hours. Although the aggregate tax on participation cannot be avoided by modifying the benefit formula, the aggregate welfare loss created by that distortion might increase due to the attempt to reduce the marginal tax on hours of work, depending of parameters. Thus, elimination the marginal tax on hours worked may be both inefficient and inequitable.

[8] This amount is derived as follows: the annual contribution is 0,15x2,500x12 = 4,500 dollars. The future value 40 years later of the cumulative value of the difference in rates of return is $4{,}500\cdot\left[\left(\frac{(1+r)^{40}-1}{r}\right)-\left(\frac{(1+\rho)^{40}-1}{\rho}\right)\right]$. This is discounted to the present with $(1+r)^{-40}$.

One may consider escaping this dilemma by requiring the lump sum tax to be different for different members. Of course, the lump sum cannot depend of the actual number of hours worked by each individual, because the tax would cease to be lump sum. If the lump sum depends of observable characteristics at a young age, say when leaving high school, educational decisions would become distorted. Going further back in time and conditioning the lump sum on genetic features seems worrying. Conditioning on parents' wealth may be indefensible in a mobile society.

Another approach is to abandon the requirement that the pension plan be financially independent. When the plan receives regular fiscal subsidies, it becomes possible to credit all individual account with the market interest rate even though the financing method is pure pay as you go (Diamond, 1981). The government subsidy pays for the negative account balances issued to new workers. However, the state must finance the subsidy to the pension plan by raising general tax revenue, or by reducing other valuable public expenditures. In both cases, new distortions appear, either by higher tax rates o by underspending in public goods. It is not clear that there is a net social gain in general, although in specific cases those gains might exist.

I conclude that there is a deep problem with attempts to improve the benefit formula keeping the financing method constant: they redistribute the wealth effect (the tax on participation) towards members of the plan that work the least hours. This is regressive, and it may exacerbate the aggregate welfare loss due to participation distortions. The next section explores the idea of seeking an increase in the degree of funding of the plan, financed by the better off members of the current generation.

4. Some empirical findings for Chile

The previous section suggests that a pension plan that is fully funded and applies an actuarially fair benefit formula avoids both labor distortions: the tax on participation, and the marginal tax on hours. The Chilean labor market provides an interesting test of this proposition because of two features:

First, the main mandatory pension plan during the last 20 years has been fully funded, and its benefit formula has been actuarial. The formula is an individual account with defined contributions, with mandatory annuitization at a variable pension age. The adjustment to average pension ages reduces the variation in the replacement rate originated in the variation of investment returns. Although commissions have been relatively high[9], the gross rate of return has also been exceptionally high from inception

[9] However, commissions in the pension system have been larger than those charged to voluntary saving not intensive in marketing, such as savings passbooks, and for higher amounts, bank deposits and custody accounts with local stockbrokers (for the evidence, see Valdés-Prieto, 1999).

(1981): recently, the net-of-commissions rates of return for 1981-2001 are estimated at 7.4% plus inflation per annum by the regulators[10], which is very attractive.

Second, independent workers have always been legally exempt from contributing. Only "dependent" workers are covered, with dependency defined by the labor law as an individual that accepts in his labor contract to work under the command of somebody else, such as a foreman or administrative superior. Being in the exempt sector in Chile implies avoiding contributions to five social security programs[11]: old age pensions, disability and survivorship insurance, health insurance and workman's compensation. The total contribution to all these programs was 21.5 percentage points of gross salary, of which 13.0 was for the pension package (old age, disability and survivorship pensions).

If we believe in the results of sections 2 and 3, we would expect the Chilean pension plan to be free of distortions to participation and hours of work, because it is fully funded and returns have been good. However, the pension reform of 1981 did not reduce appreciably the share of independent workers in total employment. That share stands in 2001 at 30% of total employment, near its value in the 1960's. New entrants into the labor market are forced by law to join the new system. The coverage of dependent workers by the new pension system has increased towards 100% over time, as the generation that chose to remain in the old system have aged and retired. However, the total coverage of both pension systems taken together has increased slowly and with fluctuations, as shown by Table 1.

Given that economic growth over 1984-1998 averaged 6.6% per year, a portion of this increase in total coverage may be due to the formalization of employment that is associated with economic growth, not to the increase in the coverage by the funded and actuarially fair pension plan.

Torche and Wagner (1998) take advantage of the fact that the workers may have arbitraged their participation in the covered labor market, to measure the size of the tax perceived by Chilean workers in a position to choose between the covered and exempt markets. They assume that migration between jobs in the two sectors proceeds until a marginal worker is indifferent between them. They also take into account that employers choose between workers of different characteristics and adjust wages until they become indifferent as well.

At one extreme, some workers may value pension benefits enough to feel compensated for the reduction in take-home salary associated to contributions. Those workers should be willing to choose between sector only on the basis of gross salaries,

[10] See table "Información histórica rentabilidad real anual de la cuenta individual, fondo tipo 1, Julio 1981-Diciembre 2000", for a taxable income of Ch.$236.549 per month, and for the average of all fund management companies, Circular 1.154, Enero 25 2001, Superintendencia de AFP, www.safp.cl.

[11] In 2001 a sixth program was added: an unemployment insurance scheme.

before contributions are deducted. Worker arbitrage would lead gross salaries to be the same in both sectors, if both coexist. In another extreme, some workers may perceive that the entire contribution is a pure tax. In this extreme, net salaries, after contributions are deducted, should equalized across sectors by worker arbitrage.

With two samples of 17.104 and 21.891 observations for 1990, Torche and Wagner (1998) run wage equations where they control for differences among workers in areas such as experience, education and region of residence. On the employer/job side, they controlled for differences due to sector of the economy and size of the employer. They used a further control regarding the existence of a formal labor contract, which may be interpreted as an indicator of formality.

They find that the Chilean workers were willing to accept a reduction in gross salary of 10.6 percentage points in order to remain in the exempt sector. Comparing with the total contribution of 21.5 points, they find that workers consider 49% of the wage differential a pure tax.

It should be noted that 4 of the 5 programs pay benefits within a short time frame. For example, health insurance, which takes 7.0 percentage points of contributions, covers mainly a portion of the cost of standard medical services. For this reason Chilean observers think that households effectively recover most of the health contribution within a year, and it is unlikely that more than a small part of it is perceived as a pure tax. For example if 20% of the health contribution is perceived to be a tax, the pure tax component of the other contributions would be (10.6-0.2x7.0)/(21.5-7.0) = 63%. This suggests that nearly 2/3 of the contribution for pensions is considered a pure tax in Chile.

As these results obtain despite the fact that the main Chilean pension plan is fully funded, and despite the fact that the benefit formula is actuarially fair and credited interest rates have been very high. I conclude that other missing factors are more important than the labor market distortions identified in sections 2 and 3. The next section presents a candidate.

5. Liquidity equivalent deposits

We suggest a look at the applicable interest rate. Sections 2 and 3 assumed that the savings interest rate was the applicable one. However, mandatory savings can create substantial inefficiencies in savings because of its iliquidity and the legal prohibition of posting it as collateral.

Consider, for example, what would be the premium on deposit interest rates that banks would have to pay, in a bank deposit contract that specified that the money cannot be withdrawn until death, disability or old age, and cannot be used as collateral, i.e. a bank deposit of "equivalent liquidity" to the pension plan. Assuming that conventional (liquid) bank deposits that pay interest rate r_p continue to be available, it is possible that consumers that are provident, patient and prudent and have the median income might require a premium on the equivalent deposit over traditional deposits of 5 or 10 percentage points per year. This is compatible with the interest rates paid on credit cards, which also serve liquidity needs.

One may think that wealthy consumers that have most their needs covered through a stock of voluntary liquid saving, and are provident, may require a negligible premium on "equivalent liquidity" deposits. But even families in the 95% percentile in the income distribution may require some premium, because if they are offered a profitable investment opportunity where minimum entry size is substantial, they may have to choose between losing that opportunity and becoming vulnerable to consumption shocks. The option value of liquidity in the investment side should be taken into account also. Of course, the truly rich would not require a premium, but as their mandatory saving is negligible, their presence does not affect our analysis.

Thus, we suggest modifying the basic model by adding this premium into the definition of the interest rate r earned by voluntary saving, before applying equation (2) and its sequels. This addition is necessary to take into account the difference in liquidity between mandatory saving and voluntary saving in conventional (liquid) saving vehicles. The implication of this small adjustment is obvious: even a fully funded plan exhibits a rate of return ρ much smaller than r, the interest rate on voluntary saving with equivalent liquidity.

The same result at the margin obtains in a different case, where the worker is driven by a high contribution rate to rely on consumer credit to cover part of his consumption. In this case, the alternative interest rate is given by the option to pay back part of the stock of consumer debt. Again a fully funded plan would exhibit a rate of return ρ much smaller than r. This is compatible with full funding in general equilibrium when the pension plan invests part of its assets at the margin, directly or indirectly, in securities issued by the lenders of consumer credit, such as banks and finance houses.

Returning to liquidity equivalent deposits, we review the implications of the models of sections 2 and 3. Section 2 showed that when analyzing a pension plan, economists should make a distinction between taxes on coverage or participation in the plan (wealth effects) and marginal taxes on hours worked. This remains correct, with the proviso that it is likely that all pension plans will exhibit taxes on both participation and hours worked, when the liquidity premium is taken into account.

Recall the definition of "actuarial fairness at the margin": an additional dollar of contributions generates an increase in expected benefits of $1 plus interest, at the same interest rate that the worker can obtain in the financial market. Using the liquidity equivalent deposit as the alternative, it seems very unlikely for any pension plan to be actuarially fair at the margin. An implication of higher implicit tax rates is that the hours distortions caused by the lack of actuarial fairness at the margin create welfare losses that are larger than anticipated in section 2. For the same reason, the welfare gains from a given improvement in the degree of actuarial fairness at the margin are larger than anticipated in that section.

Turning to section 3, consideration of liquidity equivalent deposits means that fully funded plans are not actuarially fair at the margin and distort hours substantially. They also distort participation decisions substantially, which is compatible with the Chilean evidence. Another implication concerns the statement that it is possible to eliminate marginal taxes even if the financing method is pay as you go. This must be reexamined, because the required slope in segment RS in figure 3 is much larger than anticipated (due to the iliquidity premium). A related implication is that in the notional account interpretation of a plan with no marginal taxes, the required initial balance in individual accounts issued to young workers joining the covered labor force would be larger than anticipated.

It is interesting to follow the example of a notional account plan whose internal rate of return is 2% real per annum (it may be partially funded), where the notional interest rate credited to member on their savings (or charged on their debts) is 14% real per annum, the equivalent alternative in the financial market. In this case, and keeping the assumption that the contribution rate is 15% on average taxable wages of 2,500 dollars per month, the initial lump sum tax on the individual that joins the plan that keeps aggregate financial balance is -30,534 dollars rather[12]. This is *smaller* than the -32,453 dollars calculated in section 3.

The explanation seems to be that a 14% real interest rate on debts to the notional account fund makes likely that most workers in the steady state will reach retirement with a small account balance, and their pension benefits will be small. However, the initial members that had positive account balances achieve big gains when the credited interest rate is raised to 14% real. They can get very high pensions due to the change. Therefore, intergenerational redistribution would be massive, unless initial owners of notional capital are subject to heavy notional taxes that reduce their account balances and increase future generation's initial balances. This suggests that a simple increase in

[12] Recall that the initial balance is $4{,}500 \cdot \left[\left(\frac{(1+r)^{40}-1}{r} \right) - \left(\frac{(1+\rho)^{40}-1}{\rho} \right) \right] \cdot (1+r)^{-40}$. When r = 0.14 and ρ = 0.02 the initial balance is -30,534 dollars.

the notional account interest rate eliminates the distortions on hours of work, at the cost of raising the distortions on participation for future generations. Note also that for this discussion, fully funded plans are much like notional account plans.

Finally, recall the tradeoff between elimination of the marginal distortion on hours of work, versus redistribution of the aggregate tax on participation towards those that work few hours covered by the plan. Let us assume that the members that work few covered hours are more willing to avoid participation than high salary workers that work a large number of hours. In this case, the crediting of the high interest rate of "equivalent liquidity" deposits to individual accounts coupled with a negative initial account balance would assure those workers a negative account balance at pension age, and would induce a larger participation distortion than in section 3. Thus, elimination the marginal tax on hours worked would be event more inefficient and inequitable.

Working in the uncovered sector

Let us define a "tax equivalent" of the mandate to join an illiquid pension plan. The tax equivalent of the mandate is the maximum lump sum tax that the worker is willing to bear in a job exempt from participation in the pension plan, that leaves him indifferent with taking a job that is covered with the plan, provided both jobs offer the same wage rate. A positive lump-sum tax means that the mandate causes a distortion of the decision to participate in the covered labor market.

This is different from the case where the worker can choose to work in a sector where he has lower labor productivity, but the job has the advantage of being exempt from the mandate. This implies there exists a proportional tax that reduces the wage rate per hour to w', which we assumed is below w, the labor productivity in the covered jobs. The tax on wage rates perceived by the worker that joins the exempt sector has rate $t = 1-(w'/w)$. In addition to its wealth effect, this tax distorts the choice of hours worked, reducing welfare further for those that join the exempt sector. This implicit tax means that the mandate distorts hours worked *in the exempt sector*, apart from distorting the hours worked in the covered labor market.

The lump sum tax that must be paid to be allowed to join the exempt sector must take into account the existence of this additional distortion. The lump sum tax is smaller when productivity in the exempt sector is smaller, because the exempt sector becomes less attractive. If we denote this lump sum tax as T, then, the optimal decision of the worker is given by:

(12) If $T > 0$, join the exempt sector. Otherwise, join the covered sector.

To define T, let us consider the indirect utility function N(E,w,r). N indicates the maximal utility that the individual can achieve with the data indicated in the arguments of the function[13]. E is the value of the endowment given by equation (3), which includes human capital and the present value of the pension plan for covered workers, w is the marginal wage rate from equation (6), and r is the applicable interest rate. To achieve comparability, we use the "liquidity equivalent" deposit interest rate for workers in the covered sector, so r is the same in both sectors. Of course, the contribution rate is zero in the exempt sector. Using these facts, the tax equivalent T is defined by the following equation[14]:

$$(13) \qquad N^c = N\left(w' \cdot 1 + \frac{y_o}{1+r} - T, w', r \right)$$

where N^c = utility level achieved in the covered sector. In turn, N^c is determined as

$$(14) \qquad \begin{aligned} N^c &= max U(VPC, n_y) \qquad s.t. \qquad VPC \le w \cdot 1 + \frac{y_o}{1+r} + \frac{P(A(n_y))}{1+r} - A(n_y) \\ &= N\left(w \cdot 1 + \frac{y_o}{1+r} + \frac{P(A)}{1+r} - A, w - \theta \cdot \left[1 - \frac{\partial P/\partial A}{1+r} \right], r \right) \end{aligned}$$

These equations imply an interesting new result: the tax equivalent T is similar (A- θ·1), when r grows due to the iliquidity premium, regardless of the marginal tax on hours and regardless of the financing method of the pension plan.

It can be seen in equation (13) that if the contribution rate θ is similar to (w-w'), and thus w' = w - θ, and r grows due to the iliquidity premium, then T meets the equation N^c = N(-T + w - θ, w - θ, ∞). Equation (14) shows that under the same conditions, N^c = N(w - A, w- θ, ∞). Thus, T becomes equal to (A-θ·1). This expression is the full contribution, minus the value of the wage rate differential applied to the full endowment of hours, which is normalized at 1.

The implication is that when the iliquidity premium is high, most of the mandatory contribution A is seen as a tax, regardless of the financing method, may explain the Chilean results.

---

[13] See Varian (1978), p. 85-92.
[14] To reduce notation, we assume inheritances and bequests are zero.

The policy implications are straightforward: mandatory pension plans must improve their liquidity to reduce the distortions they impose on labor market decisions. An exclusive emphasis on reducing or eliminating the marginal tax on hours is misplaces, because iliquidity is likely to be more important empirically.

6. Conclusion

Clever design of the benefit formula is always desirable to minimize the marginal tax on the number of hours worked, but when this is done a difficult tradeoff with redistribution appears: the simple methods to eliminate this tax seem to require a regressive distribution of the burden of less than fully funded pension finance.

This trade-off becomes more favorable and disappears as the financing method approaches full funding. Thus, full funding offers an important advantage: it is the only way to achieve simultaneous minimization of both the marginal distortion and the participation distortion that a mandatory pension plan places on the labor market.

Of course, the transition costs towards full funding must also be considered. If those costs are financed with new taxes on the labor market, transition provoke substantial additional distortions there. If less distortionary forms of taxation are available, then that transition may generate a small welfare gain.

The second important point of this paper is that when the iliquidity premium is high, most of the mandatory contribution A is seen as a tax, regardless of the financing method. The Chilean evidence shown suggests that this may be indeed the case in middle income countries. There seem to exist substantial saving distortions for the great majority of individuals who value the liquidity or accessibility of voluntary savings. Mandatory plans are completely illiquid, as withdrawal of funds is prohibited except in the cases of death, disability and old age, and even in those cases, withdrawal is limited to strict monthly limits.

The policy implication is that economists should work to devise methods that improve the liquidity of mandatory pension plans to reduce the distortions they impose on labor market decisions, without jeopardizing the adequacy of old age savings. This endeavor is as productive for pay as you go plans as for fully funded plans.

References

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

Torche, Arístides and Gert Wagner (1998) "Previsión Social: Valoración individual de un beneficio mandatado", *Cuadernos de Economía*, año 34, Nº 103, p. 363-390.

To be completed

Table 1: Total coverage of all employment in Chile 1984-2000
(total coverage = [non-military contributors to the new funded system + old system]/nonmilitary employment)

Year	Coverage (%)	Year	Coverage (%)
1984	55.7	1993	62.2
1985	56.9	1994	63.3
1986	57.1	1995	65.0
1987	61.5	1996	68.7
1988	60.7	1997	70.4
1989	60.1		
1990	59.6	1998	61.7
		1999	64.0
1991	62.5	2000	63.1
1992	63.5		

Sources: figures up to 1997: Acuña and Iglesias, Cuadro 4 (2000); figures since 1998: INE: employment data for Oct.-Dec. quarter of each year; Superintendence of AFP: total contributors to the new system; the number of contributors to the old system since 1998 was estimated at 200,000.