December 2, 2003

Social Security Rules that Vary with Age[1]

Peter Diamond
Massachusetts Institute of Technology

Some workers enjoy their work and want to continue working beyond what many consider a suitable retirement age. Others no longer enjoy their work (if they ever did) and are eager to stop working as soon as they can afford a decent retirement.[2] A good retirement income system will not overly discourage the first group from continuing to work at ages at which the second group will indeed retire.[3]

Some workers would save adequately and work long enough for a comfortable retirement even if there were no government- or employer-provided retirement income. Other workers would do little or no saving for retirement and would choose to retire with insufficient assets to finance a reasonable retirement for themselves and their spouses. A mandatory public system of retirement income provision needs to recognize the presence of both types of workers when designing detailed rules.

Public systems can address diversity in the working population by having rules relating *benefit eligibility* to some combination of the age of a worker and the level of the worker's continuing earnings and by relating *benefit levels* to the age at which the benefits start as well as the history of past earnings.[4]

---

[1] Paper prepared for 2003 CeRP Conference, 'Is Mandatory Retirement an Outdated Feature of Pension Systems'? September 16, 2003, Real Collegio Carlo Alberto, Moncalieri, Turin. I am grateful to Elsa Fornero for helpful comments.

[2] Both stopping work and starting retirement benefits are often referred to as "retiring." Yet these are distinct concepts and I will avoid referring to either as the start of retirement.

[3] This essay considers rules for national systems that apply to many different industries. To an economist it seems obvious that such a system should not have a single mandatory retirement age. Whether such a rule is appropriate and useful in some particular industries requires different analysis.

[4] This essay considers rules for retirement benefits without also considering rules for disability benefits. The interaction between disability and retirement rules is an important subject. See, for example, Diamond and Sheshinski, 1995. I also do not consider programs that provide minimum income guarantees to poor

This essay begins by considering how a sensible worker might do retirement planning. This shows the considerations behind some workers' retirement decisions. It also shows that many workers would not want to start retirement benefits until they stop working (or cut back significantly) while other workers would sensibly start retirement benefits while continuing to work. Then we consider some principles of the design of a mandatory public system of retirement income. These relate to setting benefit eligibility rules and benefit levels (relative to the age at which they start) to balance concerns coming from workers showing different bases for decisions and experiencing different work opportunities.[5]

The focus is primarily on rules that explicitly relate to age around the age at which benefit eligibility starts. Thus there are several (implicit) age-varying rules that I will not consider. Any method of benefit determination that is based on the history of earnings in different years is implicitly adjusting for different ages by how it weights the earnings in different years. For example a notional defined contribution system (NDC) aggregates taxes on earnings in different years using some measure of wage growth. This treats different ages differently than a defined contribution (DC) system that aggregates taxes on earnings in different years using the rate of return on assets in the account. I will not address issues raised by differences in weights given to different years.[6] Similarly, I will not address the time shape of benefits after retirement which is also an issue that varies implicitly with age. And I will not consider issues raised by the families of workers (particularly widows).[7]

---

elderly. These also interact with retirement income rules, particularly if retirement income is not annuitized. Not that these issues are unimportant – it is just that this is a limited essay.

[5] This essay does not consider the optimal overall level of benefits, just the relative levels for different ages. On the former, see Diamond, 1995. It is also limited in considering retirement income systems without acknowledging the presence of other institutions, such as annual income taxes, that affect the implications and interpretation of behavioral responses to the retirement income system.

[6] Existing systems tend to count or not count earnings in individual years. A more complex weighting scheme could likely do better in insuring the risk associated with the earnings trajectory.

[7] For some discussion of issues in the treatment of widows, see Diamond, 1995, 2004.

The focus is on the impact on individual workers, not on others. For a policy design applying to many years it is a major mistake to discourage work by older workers in the hope of generating more work for younger workers. The economy adapts (through wage adjustments and labor demands) to the supply of workers. It is simply not the case the countries that most discourage work by older workers have systematically lower unemployment rates among younger workers.

There is a large difference in approach between US Social Security, which has a nonlinear (progressive) benefit formula and the typical European (and Canadian) system which has a linear benefit system. While all of these countries have minimum income systems for the elderly, that in the US is far less generous than those in Europe (and Canada). Thus the US system has been viewed as an important tool in fighting elderly poverty, while the higher level of support of the poor elderly through the minimum income system makes this unnecessary in many other countries. But there is more to the difference between linear and nonlinear systems than just fighting poverty. Progressivity in the benefit formula can be seen as part of a country's general redistributive system since it does redistribution on a lifetime basis, while typically tax-transfer systems (e. g., progressive income taxes) are based on annual evaluations and thus can be usefully supplemented by redistribution with a lifetime basis.

While a linear system is not overtly redistributive, it needs to be kept in mind that this does not imply that the system is not redistributive. In any system providing annuitized benefits there is a redistributive component coming from the systematic differences in life expectancy in the population. In particular, women tend to live longer than men, making a gender-neutral system redistributive toward women collectively. It is also the case that within each gender high earners tend to live longer than low earners, with large differences in some countries. This makes a linear system regressive within each gender. A proper analysis of this aspect of redistribution needs a proper counterfactual of what annuitization, if any, would happen without the mandatory program. The simplest setting is to imagine everyone would have annuitized anyway and the mandatory program substitutes uniform pricing for the risk-adjusted pricing that

would occur in a private market. In this case the pattern of redistribution is easy to see, being based on price differences. Once we recognize that there is very little voluntary annuitization in the world, then the discussion needs to incorporate the insurance value of annuitization (which varies with income level) as well as the differences in life expectancy.[8] For this essay, I will simply assume a linear benefit formula. This still leaves open the determination of additional benefits from additional years of work and the issue of how benefits should vary with the age at which they start. This also leaves the important issue of the use of system design to provide insurance of earnings risk. That is, even if we consider a population with the same ex ante wage opportunities, but different stochastic realizations of both opportunities and difficulty of work, we have an insurance issue that closely resembles analyses from a redistributive perspective. Providing insurance to those who have shorter careers because of poor stochastic realizations (assuming as I do that they have greater unmet needs) is an important dimension of the public provision of retirement income, one that remains present even if one constrains the basic benefit formula to be linear. Thus I think of the typical European approach as a constraint to a linear benefit formula rather than a choice to ignore insurance opportunities.

Individual choice without a mandatory system

In a world where all retirement decisions were unconstrained by public or private programs, each worker would decide how much to save (for retirement and other future expenditures), when to stop working, when to start drawing down accumulated balances, and whether to make use of annuities, in some form. A worker choosing to annuitize has choices about when to purchase the annuities, when to have the benefits start, and what time shape and indexing structure to give to benefits.[9]

---

[8] For a systematic study of this issue, see Brown, 2003.

[9] The latter includes choices such as whether to index, whether to have a flat benefit or a sloped benefit relative to the index, whether to use single-life or some form of joint-life annuities, and whether to include a contingent bequest in the annuity contract (often referred to as a guarantee).

In modern advanced economies, workers are required to pay taxes while working (to help finance retirement benefits), required to receive benefits in a given form (commonly as an annuity), and restricted as to when they can receive benefits. The restrictions on benefit receipt may depend on age or a combination of age and earnings levels. Workers may or may not have some choice of assets during accumulation.

Individuals with complete freedom of choice (subject to market availabilities) who follow the standard life-cycle model under uncertainty would start drawing down accumulations for retirement under three circumstances – at the same time they stop working, when work has become lucrative enough or life expectancy short enough that the retirement accumulation is best spent over a longer period than just after stopping work, and when there are some particularly important expenditure opportunities, perhaps for a consumer good, such as world travel, or perhaps for a need, such as medical expenses. For simplicity I assume that the end of work is a once-and-for all decision (although many workers in the US do return to work after self-described retirement) and that the work decision is a zero-one choice – full-time work with no variation in hours or full retirement (although many workers in the US use "bridge jobs" in a transition between a career job and full retirement). Moreover, I will assume that the opportunity set and the arrival of information are such that the choice problem is well-behaved so that one can consider first-order conditions for decisions without needing to compare two distinctly different planned retirement dates as alternatives.

Individuals planning their own retirement and sensibly looking ahead would contemplate the end of work by comparing the utility from consuming the earnings from additional work with the disutility from additional work. Such individuals continuing to work need to decide whether to save or dissave out of earnings, that is, whether to add to their retirement accumulation or subtract from it (in addition to interest earnings). This savings decision would contrast the marginal utility of consumption while continuing to work with the marginal utility from the consumption they could afford after retirement. This is similar to the decision for a worker in a retirement income system between starting retirement benefits now or waiting to start them later. In either setting a worker

would recognize the implications of a delay in the start of benefits in terms of a higher flow of per-period benefits that would start later. How much higher depends on the arrangements for retirement income flows and the behavior of the market. For example, an individual contemplating purchase of a real annuity would recognize that a delay of a year in the purchase of an immediate annuity would mean another year of stochastic accumulation on the existing balance, and a change in the pricing of annuities reflecting both being a year older and the changes in the bases of annuity pricing in general – developments in cohort mortality estimates and in interest rates. Of course, not all of retirement benefits need to be consumed – some could be saved. Alternatively, not starting benefits may be accompanied by further savings from earnings. Implicit in this formulation is the assumption that individuals do not engage in complex asset market transactions to separate out decisions that are made contingent on a retirement decision from decisions that could be made separately. That is, I assume that workers accumulate balances in some portfolio (with unchanging balance) and annuitize at retirement. While economic theory makes it clear that it is advantageous to annuitize earlier (or on a rolling basis) (Brugiavini, 1993, Sheshinski, 2001) this does not seem to be commonly done. Similarly, most workers do not engage in sophisticated hedging actions. Thus I do the analysis around a simple behavioral model, which seems more useful than considering what super-sophisticated workers might find to do, although the latter analysis can also help inform the design of mandatory systems.

In the familiar life-cycle model under certainty, a forward-looking worker retires when the disutility of continued work equals the marginal utility of the additional life-time consumption that could be financed as a consequence of additional work. Adding uncertainty about life-expectancy does not change this analysis if the individual fully annuitizes as may make sense in a setting with no other uncertainty (Yaari, Davidoff, Brown and Diamond). Randomness in interest rates and in annuity pricing would complicate the story, adding reality and additional interactions, but adding little to the basic character of the retirement decision. In a certainty setting the start of drawing on assets for consumption and the end of work occur at the same moment provided the wage

is nondecreasing. (With a constant interest rate, drawing on asset income and decumulating occur at the same time; with varying interest, this need not be the case.)

However, it is straightforward to construct models where assets are drawn on before retirement. What can accomplish this is the arrival of opportunities or information. Either a surprisingly good wage opportunity or learning that life expectancy will be short can result in this pattern. Of course, the arrival of a temporary spending need (e. g., medical expenses) or opportunity (e. g., a good time for world travel) can do the same (in the absence of insurance covering this event).

From the perspective of a mandatory retirement income policy, there are two issues here. One, not discussed but apparent in practice, is that without strict mandates discouraging it, different sensible people want to retire at different ages since they differ in many ways. Second, there may be a difference in timing between when workers want to stop working and when they want to start drawing on their "retirement" benefits.

Age-varying payroll taxes

A mandatory system starts (chronologically) with a requirement to pay taxes. In practice taxes are levied on earnings and taxes are proportional to earnings for any worker, possibly between a floor and a ceiling. The tax rate can vary with age (and does in Switzerland). The advantage of an age-varying tax is to have the tax relatively lower when more workers are having liquidity problems than they will have later in life. That is, typically younger workers have lower earnings than they will have later, and may have higher needs to prepare for home ownership and to buildup precautionary balances, and possibly young children to finance. When older they are more likely to have an easier time setting aside a larger fraction of earnings for retirement purposes. The disadvantage of an age-varying payroll tax is added administrative burden. There is a potential here for a useful age-varying policy, but I do not know empirical work that spells out in detail the extent to which liquidity problems vary by age and so the ideal shape of such a policy.

At the other end of the life cycle, one could lower the payroll tax rate (or drop it totally) for workers who have not yet stopped working by the age at which they can start collecting retirement benefits. In Chile, compulsory savings in individual accounts stops on reaching the point of being able to start drawing benefits at age of 65. In a system without insurance related to earnings or redistribution across workers, it is not clear what is gained by continuing taxes on workers who have access to their retirement accumulations. That is, if a country lets workers start benefits while continuing to work, there is little apparent gain (apart from simplicity) from continued taxation of earnings in a system without redistribution across workers based on earnings histories. In a system that does explicitly redistribute (as does the US and do rules that contain implicit taxes that depend on the length of a career), then continued earnings do represent an addition to lifetime resources and so a greater ability to pay taxes on average. Note I refer to explicit redistribution – there is no way to design a compulsory system that does not involve some redistribution, as noted above.

For the continued analysis, I simply assume a single payroll tax rate that does not vary with age and continues either until benefit eligibility or as long as work has not stopped, depending on which assumption helps make the analysis clearer.

Annuitization

Different countries have different rules as to whether benefits must be received as an annuity or how large a fraction of accumulations must be annuitized or with strong incentives encouraging a particular degree of annuitization. I will not explore this choice of public design of benefit structure, which generally does not vary with age, although the UK has rules requiring annuitization by a particular age.[10]

Different countries have chosen different rules for the determination of the change an annuity benefit level that is already in the payment phase with the passage of time. Some vary benefits with prices, some with wages, some with a combination of prices and

[10] In the US, withdrawals from tax-favored retirement accounts must start by age 70 and 1/2.

wages, and some include a tilt relative to the chosen index. There are good arguments for each of these choices and I will not explore the alternatives. Note that the greater the increase in benefits with age (time) the lower the initial benefits for a given availability of finances. A steeper pattern with a lower initial benefit can be particularly valuable in influencing decisions to stop working by short-sighted workers.

Different countries also vary in the protection offered a surviving spouse, either in the form of a benefit for a spouse who never worked or a change in benefit level for a spouse who also receives a benefit from past work. Protection of surviving family members, particularly widows, is an important part of the design of a good retirement income system. Again, I will not explore this issue here.

And some countries allow some choice in some of the details of annuitization.

To continue the analysis, I will simply assume that once benefits start, they are paid as a real annuity. A critical issue is how much benefits increase if they start later, assuming workers are free to delay the start of benefits. Before discussing this issue, we need to consider what age or combination of age and earnings results in eligibility to start benefits.

Conditions for the start of benefits

There are multiple possibilities for a system to determine whether a worker is eligible to start benefits. The simplest rule is an age at which benefits start – no choice, no requirement of stopping work. A variant on this approach would allow a worker to defer the start of benefits in order to receive larger benefits once they start. Alternative to this approach, the start of benefits might require both a minimum age and a test of retirement (perhaps measured as some allowed level of earnings), with benefits increased for workers who do not stop work and so do not start benefits.[11] This approach could

[11] For those earnings just a little more than the exempt amount, benefits would be reduced to offset some fraction of those earnings, and alter benefits would be increased to counterbalance the decrease.

also be modified to allow a deferral of the start of benefits for a worker who could start benefits, again associated with larger benefits. And, the above possibilities can be combined into an age-varying rule, as in the US, where low earnings are required for the start of benefits between two ages, benefits are allowed to start despite continued work beyond the upper age and deferral of the start of benefits despite eligibility is allowed up to a third age.[12] In addition to rules relying on age and earnings, there are rules that also include a measure of years of service. Offhand, I find it hard to think of a good reason for a national system to include a rule based on years of service, however sensible that might be in some industries. I return to that issue below.

This list raises three questions: How to choose among approaches, for each approach, how to choose the age or ages that make up the rule, and how much to vary benefits with the age at which they start.

Age only eligibility rules

The simplest case to consider is where the benefit eligibility rule is an age-only rule. Assume there is an earliest entitlement age, EEA, at which benefits start, whether continuing to work or not. Assume there is no opportunity for a worker to delay the start of benefits in order to receive larger ones. Then the remaining individual decisions are when to stop working and how much savings, if any, to do in addition to payments to the mandatory program. In thinking about how to set EEA, there are two dimensions to the effects of alternative levels of EEA – what happens to the cash flow to different workers and how workers have different behavior for different levels of EEA.

Assume that the level of resources flowing to retirement incomes is independent of EEA (in present discounted value). Then, the higher EEA the higher the benefits per

[12] In the US benefits are subject to an earnings test between age 62 and what is called the Normal Retirement Age (but would more accurately be called the Age for Full Benefits), which is in the process of changing from 65 to 67. There is no earnings test after the Normal Retirement Age, and the ability to defer benefits in order to receive larger ones stops at age 70, when benefits simply begin.

month once they start.[13] This has three implications. With the PDV of benefits the same, a higher level starting later implies fewer earlier payments and more later payments. With the same timing on tax collection, that would be an increase in national savings. Second, every worker would be receiving a higher level of annuitized benefits. If workers are not overannuitized, that represents providing more insurance against the risk associated with length of life. Third, there is a redistribution of life-time benefits in that those with longer expected lives gain at the expense of those with shorter expected lives. Or, assuming that everyone gains from more annuitization at an individual break-even, those with longer lives have a further gain, while those with shorter lives have an offsetting loss. Since higher earners (of each gender) tend to have longer expected lives, the higher is EEA the more regressive the system is within gender. On the other hand, since women tend to live longer than men and women tend to have lower earnings than men, a higher EEA is more progressive on this basis. A full examination is more complex given the presence of marriage and the tendency of couples to share their resources somewhat.

There are two aspects to this increase in EEA – a disappearance of benefit cash flow at some ages and an increase in the level of annuitized benefits at higher ages. Some workers have greater needs earlier, for example, those with poor or no earnings opportunities and little other income or wealth. Other workers have greater needs later, for example, those with jobs paying well (relative to past earnings) and a pension benefit replacement rate well below one. Thus, a critical question is which EEA does better in providing a larger cash flow when it is needed more. Plausibly the relative weight in the two concerns shifts monotonically with the level of EEA, making analysis of this aspect of the choice of EEA straightforward.

The next step is to think about how the change in flows affects individual worker behavior. To focus on the nonredistributive aspect of this, we analyze the case for a

[13] I will assume that there is some variation in benefits with the change in EEA accompanied by a uniform proportional change in individual benefits.

worker who has the same present discounted value of benefits independent of EEA.[14] To proceed we need to consider different models of how workers might react to different choices of EEA.

Forward-looking workers who behave along the lines of the life-cycle model are mostly unaffected by an increase in EEA if the increased EEA is earlier than the date at which they stop working. The exceptions would be the workers who will work so long relative to life expectancy that they would like to use some of their retirement benefits for consumption while still working. In contrast, a life-cycle worker who would retire at or before the original EEA may find it optimal to work longer than otherwise (unless savings levels were already adequate to finance consumption from the end of work to the start of benefits). Overall, for such sensible workers, the presence of limited access to retirement benefits is a form of capital market imperfection. By itself, the earlier the EEA the better from this perspective. However, insofar as individual savings do not get annuitized, or get annuitized with a heavier administrative load than is provided by the public system (at the margin), then, there is a gain from delay through the mechanism of providing more insurance. That is, with a market imperfection of poorly priced or nonexistent private annuities, the larger annuities from a later EEA can have value for offsetting this market imperfection. Thus this consideration tends to support an EEA interior to the set of times when sensible forward-looking workers would choose to stop working without an EEA restriction.

This analysis would be different if we were considering offering a particular level of monthly benefits and varying the tax rate to finance different levels of EEA. In that case the focus would be on the level of retirement savings that was optimal for different workers rather than the level of annuitization. Again, we would balance those who wanted more savings against those who wanted less, again assuming that the private market did not offer as attractive an alternative (for example through poor functioning of the annuities market from either the supply or demand side).

---

[14] If a worker works longer as a consequence of a change in EEA, there is also a change in benefits and taxes as a consequence. We assume these just balance, not including an implicit tax.

The analysis takes on a different flavor when we consider some alternative models of behavior, models that show plausibility for some workers given empirical studies of consumption and retirement. Some workers may retire as soon as benefits become available even if that implies inadequate resources for the full remaining expected life. Some indirect evidence for workers like this appears in the large spike of workers retiring as soon as benefits become available even in systems that provide a generous increase in benefits for a delayed start, as in the US. Direct evidence would focus on consumption trajectories given a belief that some patterns of declining consumption are not plausibly ex ante optimal. Studies of consumption and resources at different ages are suggestive. Increasing EEA then protects some workers (and their spouses) from retiring sooner than would be sensible given their resources and the remaining life expectancy of the worker and spouse. On the other hand, raising EEA too high would force some workers who do not save to continue working longer than may be optimal given the PDV of their future benefits.

A second form of inefficiency arises for workers who do continue working if they consume too much from the combination of ongoing wages and the start of retirement benefits relative to their later needs. Again one would look at the pattern of consumption over lifetime to identify the presence of this concern.

Formal modeling of an optimal EEA would evaluate the gains to different types of workers associated with different levels of EEA. While variation in life expectancy is measurable and the correlation of life expectancy with incomes is measurable, there are aspects of this issue that are very hard to measure. In particular, it is hard to know how many workers gain from working longer (because they have too few resources for a fully comfortable retirement), as opposed to how many lose (because they would sensibly retire earlier if they had earlier access to retirement resources.

Note that insofar as the choice of EEA affects the extent of work, and insofar as the extra benefits as a result of extra work differ from the extra payroll taxes paid, the

system will gain or lose additional net revenues from an increase in EEA that induces more work. It is important to recognize that delaying the EEA may be of little (or even negative) consequence to social security finances depending on how much benefits increase with an increase in the age at which they start. Similarly, inducing more work by actually subsidizing a longer career hurts finances (unless packaged with other elements that more than offset this).

Assume that an EEA is chosen based on some evaluation of the issues discussed above. Then, it is natural to ask how that EEA should change over time. Note that a change in EEA is likely to have little impact on system net fiscal balance if the adjustment of benefits for a delayed start is roughly actuarial. This is very different from a change in the age at which benefits are paid subject to the basic formula (an age for what are sometimes called full benefits). The latter is equivalent to some pattern of benefit cuts in that benefits are lower at any given retirement age.

Allowing deferral of the start of benefits

One could readily amend the system described above by allowing larger benefits for a later start for workers who choose to delay the start.[15] Workers might choose to delay the start of benefits for several reasons.[16] One is that they are still working (and so not liquidity constrained) and value either the additional insurance that comes with higher monthly benefits or are concerned about their self-control with a larger cash flow and choose to have higher benefits starting later to avoid the temptation of spending too much earlier in their lives.

The other incentive comes from believing one has a long enough life expectancy that one sees profit in expected present discounted value (i. e., ignoring risk aversion) from a delay. This adverse selection effect implies that there is a difficult actuarial estimation in designing a system that is revenue neutral (actuarially fair). If an

[15] I assume that the adjustment factor is the same for everyone.
[16] For evidence that some workers delay benefits that could be claimed in the US, see Coile, Diamond, Gruber, and Jousten, 2002.

equilibrium can be found with this property, then the additional choice given to workers is purely a gain (apart from some workers who perversely save too much – which is not normally seen as a widespread concern).[17] Of course the result that with breakeven pricing there will be a Pareto (or near-Pareto) gain in expected utilities, if delay is priced actuarially for those delaying, does not imply that this is the optimal pricing. By being slightly less favorable in benefit increases it would discourage some who would defer (an efficiency cost) but would permit a redistribution from those who defer to those who do not, based on the gain from those still deferring. Since those deferring are plausibly likely to be better off than those not deferring, such a deviation from actuarially fair pricing may be worthwhile in social welfare terms (Sheshinski).

Age and earnings eligibility rules

Starting from an age-only rule, we can consider adding a requirement (at least for a period of time) of a stop to work (or a very low level of earnings to allow part-time work and work by those with very low earnings) in order to start receiving benefits. The effect of adding such a retirement test (or earnings test) runs differently through different behavior models.

For most people following the life-cycle model, requiring an end to work in order to start benefits would be of little consequence since mostly they do not start using the benefits to finance consumption until they do stop work. If everyone were like this, then adding a requirement of stopping work would open up the possibility of improved insurance (Diamond and Mirrlees, 1978, 1986, forthcoming). That is, some people stop working because of poor realizations of random outcomes such as work opportunities or the utility cost of continuing work. By providing a less than actuarial increase in benefits for a delayed start associated with continued work, the system can provide higher benefits for those who stop work earlier out of the resources freed up by lower benefits (in PDV) for those working longer. In this way more insurance is provided against the poor realizations that shorten working life. That is, one can use earnings as an observable

[17] On having more choice generating a Pareto gain with all workers rational, see Diamond, 1992.

variable correlated with unobservables such as earnings opportunities or work disutilities. Of course, there is some labor market disincentive associated with this, but as is always the case with second-best insurance, some distortion of the labor market in order to have more insurance is worthwhile in simple settings. This is likely to carry over quite generally.

This analysis does need adjustment to recognize that some life-cycle savers might want to start drawing on retirement funding to finance higher consumption while continuing to work, as noted above. Limiting this availability would then have an efficiency cost in two senses – it would make consumption less efficient for those who do continue working and it would add to the responsiveness of premature retirement in order to start receiving benefits.

Also interesting is the response of workers who are less forward looking. For those who do not alter their work experience, providing benefits while they continue to work runs the danger of greater early consumption than is optimal, resulting in declining living standards after they do stop working.[18] Thus a requirement of retirement helps these workers on a lifetime basis. Offsetting this is the discouragement of further work since stopping work is needed to start getting these benefits. This may be particularly costly if these workers are prone to retire "too soon" for their own good anyway – so that encouraging further decreases in working life has a sizable effect right form the start.

In other words, we expect a spike in the retirement hazard at the EEA. Part of this spike comes from life cycle workers who either wanted to retire earlier but waited for liquidity reasons and from life cycle workers with short life expectancies so that the increase in benefits for a delayed start involves an implicit tax for them even if it is actuarially fair on average. (Offsetting this spike is the set of people who would have retired by EEA but delay retirement because of the implicit subsidy in actuarially fair

[18] There have been estimates of the implications of consuming benefits while continuing to work at different ages in the US as part of the discussion leading up to the legislation that removed the earnings test between the Normal Retirement Age and age 70 but did not remove between age 62 and the Normal Retirement Age (Gruber and Orszag, 2000).

benefit increases when one has above average life expectancy.) Part of the spike comes from workers who stop work shortsightedly as early as they are able. Adding to this spike by requiring a stop of work in order to collect benefits has a social cost that is not second order.

Consideration of these short-sighted workers raises the issue of whether there is some design of benefit provision that would result in more work without as large a resource cost as simply paying much more for additional work. That is, is there a way of providing benefits different from the rule that they are real annuities starting with the end of work that would result in more efficient decisions? One possibility is to have benefits that start smaller and grow faster than real annuities (with the same PDV). In addition to influencing career-length decisions, this has income distribution implications based on length of expected life. This is clearly a question lying in the realm of behavioral economics. There have been proposals for addressing this, although I have not seen a derivation of such proposals from basic psychological premises that have been empirically supported. Henry Aaron suggested offering a sizable lump sum benefit after a period of time (three years, say) for those continuing to work beyond EEA (Fetherstonhaugh and Ross, 1999). The argument is that a lump-sum in the future would be perceived as a larger incentive than an increase in monthly benefits with the same PDV. If the lump sum payment is age related then one has to recognized that it is provided before some people would retire anyway. A smoother incentive does not play off the difference in perceptions between lump sums and monthly flows, but in the tendency to pay too much attention to short run considerations. This proposal (Diamond, 1981) is to pay a steadily growing fraction of benefits independent of stopping work, while holding back the remaining fraction, with an increase in future benefits financed by the withheld sum. The idea is the perception of steadily growing total income while continuing to work makes continued work more attractive. These ideas, and others that might come from further analysis based on behavioral analysis should be explored. Whether we can have the sort of success that has happened in encouraging contributions to individual pension accounts (Benartzi and Thaler, 2001) is unclear, and any experiment may be harder to organize.

My bottom line here is that it is plausible that some creative use of withholding benefits in response to continued work can make a well-designed social insurance program better. It is plausible that the choice of EEA would be different with and without an earnings test. A system that withholds benefits until work stops but imposes very large implicit taxes through inadequate or nonexistent increases in benefits for a delayed start is a very poor design, and one that has historically been far too common.

Age-varying eligibility rules

It is natural that the mix of worker types being influenced by rules governing the payment of benefits would change with age. Thus it seems plausible that the rules for benefit eligibility should vary systematically with age. Without in anyway endorsing the particular ages used in the US (which have come historically from legislative parallels, not a detailed policy study), it seems to me plausible that having an earnings test over some range of ages and then paying benefits without regard to continued work make sense.

Varying benefits with the age at which they start

With a defined contribution system, assuming no change in interest rates or cohort mortality expectations or degree of price competition, the change in level of monthly benefits from a delayed start in benefits would be roughly "actuarially fair" in the sense that the expected present discounted value of benefits would be roughly the same whether there was a delay in starting benefits or not for the pool on annuitants who are combined in a single risk classification. Continued contributions while working as well as a delayed start in benefits would raise the benefit level. A public defined benefit system could adjust benefits on the same basis (and an NDC system is designed in parallel with a DC system).

It should be noted that a system that is actuarially fair for a group of workers will give a better return than would be fair for some and a lower return than would be fair for others. This is true whether or not there is risk classification with separate pricing for separate groups. Moreover, workers would be aware on average of their differences in life expectancy (Hurd and McGarry, 1995). Generally public systems use the same annuitization factors for individuals, independent of both easily measured factors that are correlated with life expectancy, such as sex and earnings level, and the possibility of trying to measure health or health practices (e. g. special price for smokers).

There are two questions about the social desirability of break-even pricing for each risk classification. One is whether even with uniform life expectancy within a risk class, social goals across groups are optimized by actuarial fairness. The other is whether the heterogeneity in the population within each classification is a basis for pricing that varies from actuarial fairness. There are research papers that have addressed both issues (Sheshinski, 2001).

But in order to make sense of the question we need to put it in the context of retirement decisions. Workers have 3 options assuming that retirement is required for the start of benefits (at least over some range of ages). One option is to retire and start benefits. A second option is to continue work and delay benefits. The third option is to retire and still defer the start of benefits. In the US, while most workers do one of the first two, a small but significant fraction of workers do the last option. I will not run through how workers conforming to different behavioral models contribute to different choices for both the EEA and the size of the increase in benefits for a delayed start.

Benefit eligibility based on length of service

It is common in industry or firm based defined benefit plans to make use of years-of-service as a key variable in determining benefits. Some national systems do this as well. I can see how this may make sense in a single type of job, but have trouble seeing that it makes sense in a national system applying to a wide range of different types of

work. Of course the level of benefits should vary with the length of career (as happens with DC and NDC systems and some other DB systems) but should the age for full benefits or the EEA vary with length of career?. This raises a question of whether to think of the ability/difficulty of work in terms of age or in terms of years of service – with those starting later (for additional education or time spent with children) significantly more able to continue working than others of the same age who had an earlier start.[19] The ability to hold jobs that make it sensible to work to later ages is plausibly correlated with length of education. Thus this is a vehicle for explicit or implicit redistribution. I do not know of research sorting this issue out, but I am skeptical of the value of relying on years of service.

[19] Such a distinction is defeated if credits are given for time in higher education.

References

Benartzi, Shlomo and Thaler, Richard "**Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving**." 2001.

Brown, Jeffrey. "Redistribution and Insurance: Mandatory Annuitization with Mortality Heterogeneity." *Journal of Risk and Insurance.* 70, 1, 2003. pages 19-43.

Brugiavini, A. "Uncertainty Resolution and the Timing of Annuity Purchases," Journal of Public Economics 50 (1993), 31-62.

Coile, Courtney, Diamond, Peter, Jousten, Alain, and Gruber, Jonathan. "Delays in Claiming Social Security Benefits," Journal of Public Economics, 84 (2002) 357-385.

Coile, Courtney, and Gruber, Jonathan. "Social Security Incentives for Retirement." In David Wise, ed., *Themes in the Economics of Aging.* Chicago: University of Chicago Press, 2001. pages 311-341.

Davidoff, Thomas, Brown, Jeffrey and Diamond, Peter. "Annuities and Welfare." MIT Working Paper, 2003.

Diamond, Peter. "Social Security: A Case for Changing the Earnings Test But Not the Normal Retirement Age." Unpublished. 1981.

_____. "Organizing the Health Insurance Market," Econometrica 60 (November 1992), 1233-1254.

_____. "Government Provision and Regulation of Economic Support in Old Age," in Bruno and Plesovic (eds.), Annual Bank Conference on Development Economics, 1995, Washington DC: The World Bank, 83-103.

_____. "Social Security," American Economic Review, March 2004

Diamond, Peter and Mirrlees, James. "A Model of Social Insurance with Variable Retirement", Journal of Public Economics 10, 1978, 295-336.

_____ and _____. "Payroll-Tax Financed Social Insurance with Variable Retirement." Scandinavian Journal of Economics 88 (1), 1986, 25-50.

_____ and _____. "Social Insurance with Variable Retirement and Private Saving", forthcoming in Journal of Public Economics.

Diamond, Peter and Sheshinski, Eytan. "Economic Aspects of Optimal Disability Insurance," (with E. Sheshinski), Journal of Public Economics 57 (1), May 1995, 1-23.

Fetherstonhaugh, David and Ross, Lee, "Framing Defects and Income Flow Preferences in Decisions about Social Security," in *Behavioral Dimensions of Retirement Economics*, Henry J. Aaron, editor, Brookings and Russel Sage Foundation, 1999, pp. 187-208.

Gruber, Jonathan and Orszag, Peter. "Does the Social Security Earnings Test Affect Labor Supply and Benefits Receipt?," NBER Working Paper #7923, September 2000.

Hurd, Michael, and McGarry, Kathleen. "Evaluation of the Subjective Probabilities of Survival in the HRS," with, *Journal of Human Resources,* 30, 1995, S268-S292.

Sheshinski, Eytan. Optimum and Risk-Class Pricing of Annuities, unpublished 2001.

Yaari, Menahem. "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer." *Review of Economic Studies* 32(2), 1965. pages 137–150.

Starting benefits

	Benefit eligibility rule	
Benefit start rule	**Age only**	**Age and earnings**
Must start at EEA	**1**	**3**
Can defer	**2**	**4**

The size of the increase in benefits for a deferral forms an effective continuum from must start to and beyond actuarially fair.

Appendix: Models of Retirement Decisions

Individual choice

Consider a n+3 period model. Assume that everyone works in the first n periods which are identical and referred to collectively as period 1, no one works in periods 3 and 4, and work in period 2 is the choice variable of concern. These assumptions could be derived from more fundamental assumptions about preferences and opportunities. Assume also that everyone survives to period 3 and there may be a positive probability of death prior to period 4.

Certainty Model

For convenience I assume a zero utility discount rate and zero interest rate. Denote the utility of consumption when working at age z by $u[c] - a_z$. Denote the utility of consumption when retired by $v[x]$. Using separate notation for consumption whether working or not is convenient. Denote the wage at age z by w_z, with the same wage (and same labor disutility) in the first n periods, and assume no lump-sum income. Then utility maximization can be written as the choice between an n period career and an n+1 period career. In both cases consumption will be the same in periods of work and in periods of retirement. With an n-period career, using a superscript 0 to denote choice variables, utility maximization is

$$\text{Maximize}_{c,x} \quad nu[c^0] - na_1 + 3v[x^0] \tag{1}$$

$$\text{subject to:} \quad nc^0 + 3x^0 = nw_1$$

With n+1 periods of work, using a superscript 1 to denote choice variables, we have

$$\text{Maximize}_{c,x} \quad (n+1)u[c^1] - na_1 - a_2 + 2v[x^1] \tag{2}$$

$$\text{subject to:} \quad (n+1)c^1 + 2x^1 = nw_1 + w_2$$

Given the utility-of-consumption functions, u and v, the choice of length of career depends on the disutility of labor and wage in period 2. Consumption in period 2 might be higher or lower than w_2, with a sufficient condition for it to be lower (to not decumulate) being $w_1 \leq w_2$.

To see the outcome most simply, let us assume that the two utility-of-consumption functions, u and v are the same. Then, in this certainty setting, consumption is the same in every period. With an n-period career, lifetime utility is $(n+3)u\left[\frac{nw_1}{n+3}\right] - na_1$. With an n+1-period career, lifetime utility is $(n+3)u\left[\frac{nw_1+w_2}{n+3}\right] - na_1 - a_2$. Comparing these two expressions makes clear the role of both the wage and the disutility of labor in choosing the length of a career - w_2 is added to lifetime resources and a_2 is subtracted from lifetime utility.

Note that depending on the size of these two parameters, consumption in period 2 might be higher or lower than earnings in period 2 for someone with an n+1-period

career. That is, a worker might want access to asset accumulation for retirement while continuing to work. With a zero interest rate, having consumption exceed the wage is equivalent to having asset values decline. With positive interest, this equivalence does not apply. The focus is on the sign of the wage less consumption.

Uncertainty Model

Adding uncertainty about a_2 and w_2 would add realism and more readily allow for decumulation while working. That is, with certainty, earlier consumption adapts to planned career length as influenced by a_2 and w_2. With uncertainty, prior consumption is necessarily independent of the realization of these variables in period 2. Thus there is less scope for adaptation to a better career opportunity (higher wages or lower disutility), resulting in a larger impact of work on consumption in period 2 while continuing to work.

Similarly, adding the arrival of information about life expectancy at the start of period 2 (assuming no insuracne) opens up a similar opportunity as can be seen by assuming the information changes the probability of survival to period 4 from1 to 0. The realization of the shorter life and so altered budget constraint may result in a prompt retirement or may result in work and asset decumulation, even if there would have been continued accumulation if the news was of survival to period 4. A similar possibility would arise with stochastic interest rates - a surprise increase in rates could also lead to decumulation while continuing work.