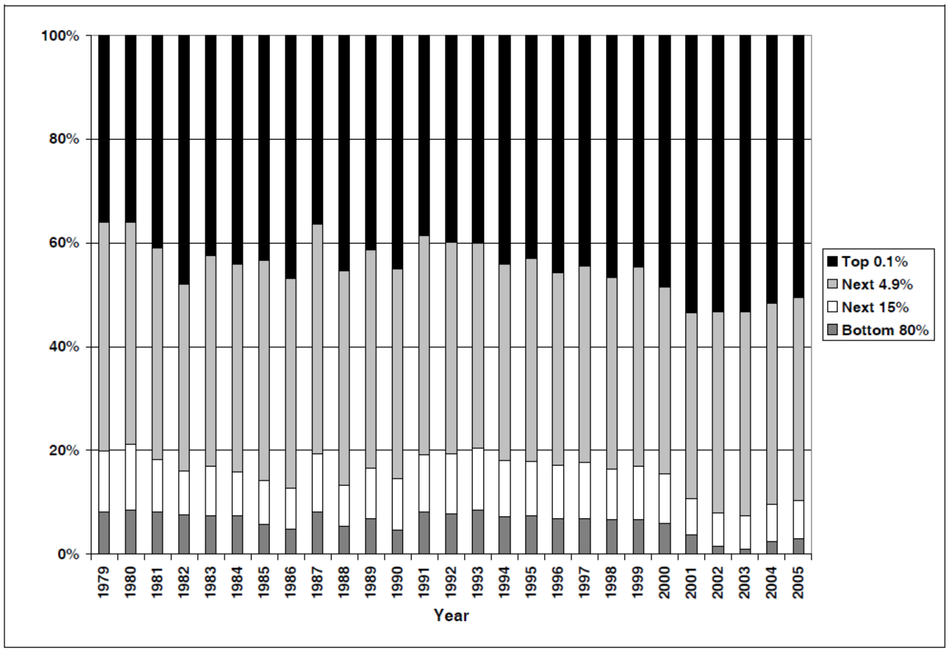
Economics 230a, Fall 2014
Lecture Note 11: Capital Gains and Estate Taxation

Two taxes that deserve special attention are those imposed on capital gains and estates.

Capital Gains Taxation

Capital gains taxes are of particular interest for a number of reasons, even though they do not account for a large share of revenue for a typical government, including the United States. According to Hungerford (Congressional Research Service, 2009), “Since 1954, revenue from the capital gains tax as a share of total income tax revenue has averaged 5.2%.” One reason for the interest in capital gains is their concentration at the top of the income distribution, as shown in this figure (from the same paper).

Figure 5. Distribution of Realized Capital Gains by Income, 1979-2005



Source: CBO, *Historical Effective Tax Rates, 1979-2005: Supplement*, December 2008.

Another important aspect of capital gains is that they are taxed upon *realization* rather than on accrual. This factor makes capital gains taxation complex and subject to a variety of potential taxpayer responses.

What does realization-based taxation do? Consider a two-period model in which an investor has an asset purchased in an earlier period for $1, which has already appreciated in value by an amount g. The investor can either hold the asset for another period, earning an additional return r, or sell and earn the market rate of return i. Suppose all income is taxed at rate t, but only when assets are sold. Also suppose that the investor’s objective is to maximize terminal wealth.

If the investor sells the asset and reinvests, terminal wealth is:

$$W_R = (1+g(1-t))(1+i(1-t)) = (1+g)(1+i) - t[g(1+i(1-t) + (1+g)i]$$

If the investor holds the asset until the end of the second period, terminal wealth is:

$W_H = (1+g)(1+r) - t[(1+g)(1+r) - 1] = (1+g)(1+r) - t[g + (1+g)r]$

Comparing the terms in brackets in the second version of each expression, we can see that the "hold" strategy enjoys a tax advantage over the "realize" strategy – first period gains, *g*, are taxed one period earlier under the latter, and hence the tax liability has a higher accumulated value at the end of the second period because it is multiplied by $1+i(1-t)$. It follows that if $i = r$, the investor will choose to hold rather than to realize, and indeed that there is a range of values of $r < i$ for which it will still be optimal to hold rather than to sell. This phenomenon is known at the lock-in effect – in order to defer tax on previously accumulated gains, individuals have an incentive not to sell assets even when, for non-tax reasons, they would prefer to sell. In this example, the lock-in effect is associated with the investor's willingness to accept a lower before-tax rate of return, but in a realistic setting the major distortion comes from an inefficient allocation of assets across investors. That is, when an individual realizes a capital gain by disposing of an asset, that asset does not typically disappear, but instead ends up in someone else's portfolio. Thus, it is unlikely simply to have a below-market rate of return, because asset prices adjust. Rather, in a setting with risky assets, other investors may be willing to pay more for the asset than the individual currently holding it. For example, suppose that there are two investors, one holding appreciated stock in Apple and the other holding appreciated stock in General Electric. As returns on these two assets are not perfectly correlated, a combined portfolio would offer a better risk-return trade-off than either specialized position. Thus, absent taxation, each investor could be made better off by trading with the other, but if each faces the capital gains tax, the gains from trade may not be realized. Even with no change in overall assets, there is deadweight loss.

The lock-in effect is exacerbated by two other provisions found in the US tax system and typical of others as well. First, gains on assets held for at least one year are taxed at a lower rate (in United States at present, a maximum of 20% vs. a maximum tax rate on ordinary income of 39.6%). Second, gains held until death are not taxed at all. On the other hand, the lock-in effect is reversed when an asset has gone down in value ($g < 0$ in the above example), since deferral of tax in this case means deferring a tax *refund*. Thus, individuals have an incentive to hold gains and realize losses, meaning that those with large numbers of distinct positions in different assets could, on a regular basis, achieve liquidity by "harvesting" losses without having to realize gains. This possibility, in turn, is largely responsible for another tax provision, which limits the annual value of deductible losses (in excess of realized gains) to $3,000. Unfortunately, as discussed in Lecture 10, a limit on the deductibility of losses also discourages risk-taking.

Empirical Evidence on Responses to Capital Gains Taxation

There has been a substantial literature relating capital gains realizations to capital gains tax rates. One of the key issues is the need to distinguish between short-run and long-run responses. We would expect that a change in tax rates could have a large impact on the timing of realizations, because individuals can adjust the timing of their asset sales. For example, after the October, 1986 passage of the Tax Reform Act of 1986, which increased the capital gains tax rate on high-income individuals from 20% to 28% effective January 1, 1987, there was such a surge in realizations in the remainder of 1986 that realizations for that year were approximately twice as high as those in 1985 or 1987. But that doesn't mean that we would expect realizations to be permanently twice as high under a 20% tax rate as under a 28% tax rate. One standard approach developed using panel data by Burman and Randolph (1994; hereafter B-R), and discussed in Poterba, section 3.2, estimates the specification:

(1) $$\ln g_{it} = X_{it}\beta + \gamma_1\tau_{it}^p + \gamma_2(\tau_{it} - \tau_{it}^p) + \gamma_3(\tau_{it} - \tau_{it-1}) + \varepsilon_{it}$$

where g is capital gains, X is a vector of individual attributes, τ is the individual's capital gains tax rate, and τ^p is a measure of the individual's "permanent" tax rate. There are three econometric issues that must be confronted in estimating (1). The first is that realized gains may be zero; a Tobit estimator is used to deal with this. The second issue is that the capital gains tax rate may depend on the level of gains realized, since tax rates rise with income. To deal with this, which is a common problem in empirical analysis of behavioral responses to taxation, B-R use as an instrument for τ a so-called "first-dollar" tax rate – the capital gains tax rate the individual would face on the first-dollar of capital gains realized. The third issue is how to define the individual's "permanent" tax rate. B-R represent this by regressing τ on the maximum federal plus state capital gains tax rate in year *t* in the state where individual i lives, as well as other individual attributes *X* (but not the first-dollar tax rate in year *t*). The rationale is that if an individual's tax rate fluctuates over time due to changes in individual circumstances, such as other income or deductions, then this will affect τ but not τ^p. From their estimates, B-R find a long-term elasticity (based on the coefficient γ_1), taking account of both extensive and intensive responses in the Tobit (i.e., to realize gains and how many gains to realize), of close to zero, and a short-term elasticity (based on the sum of the coefficients $\gamma_1 + \gamma_2 + \gamma_3$) of larger than 6 in absolute value. They conclude that virtually all observed responses of capital gains to tax rates involve timing of realizations, rather than changes in the underlying frequency of realizations. This has important implications for considerations of policy changes, for it means that even though revenues may increase in the short run in response to a reduction in capital gains tax rates (since the short-run elasticity exceeds 1 in absolute value), the opposite is true in the long run.

One critique of the B-R specification is that the use of the maximum federal plus state tax rate to identify τ^p does not correctly distinguish timing and permanent responses. On the one hand, the maximum state and federal tax rates change over time during the sample period, so some of the responses to the B-R measure of τ^p may be timing responses, which would tend to overstate the estimate of the long-run response. On the other hand, individual tax rates may persistently vary from the B-R measure of τ^p, meaning that some of the response classified as temporary may actually be permanent, which would tend to understate the estimated long-run response. To deal with this issue, Auerbach and Siegel (A-S) replace τ_{it}^p in the above specification with τ_{it+1}, the tax rate the individual will face the following year, which is generally known at time *t*, and add the maximum federal plus state tax rate from year *t*+1 as an instrument, also including the first-dollar tax rate as an instrument for τ_{it+1}. The notion here is that next period's tax rate for the individual is a better measure of the individual's "long-run" tax rate than the current year's maximum tax rate. (Viewing next year's tax rate as a good measure of the long-term tax rate makes sense if tax rates roughly follow a random walk.) A-S find a temporary elasticity that is much higher than the permanent elasticity, but their estimated permanent elasticity is substantially greater than 1 in absolute value – much higher than that found using the B-R methodology for the same sample. One other finding by A-S is that individuals who are wealthier or more sophisticated (based on the nature of their transactions) have a higher temporary elasticity and a lower permanent elasticity. The first of these results, especially, may indicate more careful tax planning. In principle, one would expect the timing of capital gains realizations to respond to the *second* moment of the distribution of future capital gains tax rates as well as the first. That is, given the expected value of the future tax rate, greater volatility of the future rate should increase the value of the "real option" embedded in the decision not to realize an accrued gain, as the individual has a greater prospect of realizing the gain in the future at a low rate. However, the literature has not, as yet, uncovered such a relationship.

A further empirical finding of interest is by Ivković, Poterba and Weisbenner, who consider differences in capital gains realizations by individuals who hold both tax-favored and taxable accounts. According to standard theory, there should be no lock-in effect for assets in tax-favored accounts, so that gains should be realized sooner, and losses later, than in taxable accounts. Indeed, the authors find that, *relative to assets in their tax-favored accounts*, investors are less likely to realize gains and more likely to realize losses in their taxable accounts (Figure 3B**)**. However, they also find that investors are more likely to realize taxable gains than taxable losses (Figure 1). There are a variety of possible explanations for this latter finding, including a belief that stock prices are mean-reverting (so that those with gains are expected to fall and those with losses are expected to rise), a need to rebalance portfolios (and hence to shed those stocks that have gained and as a result occupy a larger portfolio share), and the presence of a "disposition effect," by which individuals perceive losses more fully if they are realized.

Reforming the Capital Gains Tax

Some changes in the capital gains tax (such as taxing capital gains at death) could serve to reduce the lock-in effect, but other problems remain as long as the basic approach to taxing capital gains upon realization is followed. Some arguments for keeping the capital gains tax rate lower than other capital income taxes, including the potentially higher behavioral response elasticity and the importance of capital gains in fostering venture capital investments, relate to the realization-based nature of the tax (in the latter case because risky venture-capital investments face serious limitations on their ability to deduct losses, which as discussed earlier is a necessary feature of a realization-based system).

What other alternatives exist? One simple idea would be to tax capital gains as they accrue, rather than upon realization (perhaps combined with a reduced rate to offset the increased present value of taxes). But there are two problems with this approach: (1) taxpayers may lack liquidity to pay taxes until assets are actually sold; and (2) the government may not know the value of some assets until they are actually sold. One proposal for dealing with the liquidity problem, by Vickrey (*JPE*, 1939), amounts to keeping an account of accruing gains and the associated tax liability and charging interest on this accruing unpaid balance until asset sale. That is, the tax liability as of date *s* would evolve according to:

$$(2) \qquad T_{s+1} = [1+i(1-t)]T_s + tr_sA_s$$

where r_s is the rate of return at date *s*, A_s is the value of the asset at date *s*, *i* is the safe rate of interest and *t* is the tax rate. A problem with Vickrey's approach is that r_s and A_s may be unobservable, but Auerbach (*AER* 1991) argued that one can generalize Vickrey's approach to:

$$(3) \qquad T_{s+1} = [1+i(1-t)]T_s + tiA_s + t^*(r_s-i)\, A_s$$

where t^* can take on any value, since (as discussed in Lecture 10), a tax rate on a risky asset's return in excess of the safe rate has no effect on the investor's opportunities. Auerbach then showed that a tax liability of the form:

$$(4) \qquad T_{s+1} = \left[1 - \left(\frac{1+i(1-t)}{1+i}\right)^{s}\right] A_s$$

satisfies (3). Note that the only information needed to assess the tax in (4) is the sale price, A_s, the holding period, s, the safe rate of interest, i, and the tax rate, t, all observable. Auerbach and Bradford generalize this result and show how it can be implemented using a tax system based exclusively on observed cash flows, without having to keep track of individual assets and holding periods.

Estate Taxation

Taxation of estates (or inheritances, if levied on the recipients rather than decedents) is interesting for many of the same reasons that capital gains taxes are. Estate taxation hits only individuals near the top of the income and wealth distribution (in the United States historically around 1-2% of decedents each year) and is also subject to tax planning that can make the tax base very responsive to tax rates. And, like annual capital income taxes, estate taxes discourage saving by reducing the after-tax rate of return. But, there are a number of issues that arise particularly in the case of estate taxation, including intergenerational transmission of wealth, the nature of the social welfare function, and motivations of individuals who leave bequests.

One approach to thinking about future generations is with a dynastic model, in which future generations are simply extensions of current ones. Under this approach, familiar from the Ricardian equivalence argument, the optimal taxation of bequests (leaving aside special tax avoidance opportunities) is simplified by thinking of the consumption of heirs is just another component of future consumption; the Chamley-Judd logic favoring a long-run tax rate of zero would seem to apply. But, if we treat heirs as distinct individuals whose well-being should enter separately in the social welfare function, then the bequest decision has a positive externality, for the individual leaving the bequest takes account only of his own well-being, and not the benefit of the recipient(s). The standard Pigouvian solution is to subsidize bequests and other interpersonal gifts (Kaplow 2001). In a model where the well-being of recipients depends monotonically on the size of bequests received, the externality is declining in the size of the bequest, because of the concavity of social welfare with respect to the consumption of recipients, so the bequest subsidy should decrease with the size of bequests, converging to zero (Farhi and Werning, *QJE* 2010).

A second relevant consideration is that bequests received influence individual decisions. Here, the insights of the New Dynamic Public Finance literature carry over: reducing bequests received helps loosen incentive compatibility constraints on the income tax schedule facing heirs. Thus, whether the marginal tax rate on bequests should be positive or negative depends on the relative strength of this factor and the positive externality, as analyzed in Kopczuk's short paper. A useful observation here is that when the economic circumstances of heirs are not fully predictable from the size of bequests (as would be true if the abilities of parents and children are not perfectly correlated), then inheritance taxation (which, unlike estate taxation, can take account of the heirs' economic circumstances) can improve the performance of the wealth transfer tax. Assuming the use of estate taxes, though, this intergenerational correlation of well-being will influence the optimal tax rate; a stronger correlation points toward more progressive estate taxation, because of both the declining externality and the increasing value of relaxing the incentive compatibility constraint.

The preceding discussion presumes that bequests result from an optimizing decision in which those leaving bequests balance the benefits of own consumption and the benefits of leaving a bequest. But the precise nature of the bequest motive matters for the design of the optimal tax schedule, and bequests may result even without an explicit bequest motive. Without complete annuity markets, individuals may engage in precautionary saving to substitute for the lack of annuities, to avoid outliving their assets. Even with annuities that insure against uncertain mortality, there are other important uncertainties in old age, such as the costs of health and long-term care, for which complete insurance may be very difficult to obtain. Thus, individuals may leave "accidental" bequests. The magnitude of accidental bequests affects the optimal estate tax. Since there are no behavioral distortions involved in taxing accidental bequests, their presence would tend to increase the optimal estate tax rate. A way of thinking about this is that the taxation of bequests, relative to other taxes not conditional on mortality, acts as a kind of annuity, providing more resources to those who survive than those who die (Kopczuk, *JPE* 2003). However, the optimal estate tax rate would not be 100 percent even with only accidental bequests, since the well-being of heirs would still need to be taken into account.

As to the nature of the bequest motive, two standard assumptions are the dynastic motivation already discussed, and the "warm- glow" motive where the donor's utility derives from the net bequest. For some types of analysis the distinction is not that significant, but one potentially important difference relates to the long-run elasticity of bequests with respect to the tax rate. Under the dynastic motivation, this elasticity would appear to be infinite, which as discussed above would push the optimal tax rate (ignoring the additional impact of the externality) toward zero. But in a model with stochastic ability draws that may break the chain of future bequests (because bequests cannot be negative), the distinction from the warm-glow approach lessens, as Piketty and Saez show. Yet another potential motivation, the accumulation of wealth, resembles the warm-glow approach, in that the donor's well-being relates to the size of wealth not consumed, but if this well-being really doesn't depend on the size of the bequest received, then wealth transfer taxes are, as in the case of accidental bequests, not distortionary with respect to the bequest decision.

The actual motivation for bequests naturally differs across individuals (depending, for example, on whether they have children), and there is no reason to expect that any individual's bequests would reflect only one of the several motivations. Kopczuk's *Handbook* chapter surveys the evidence, based on different approaches to determining bequest motives. An additional point that may be important here is that some observed patterns of bequests and *inter vivos* wealth transfers require further explanation, not being consistent with any of the foregoing motives. In particular, as discussed in Poterba's *Handbook* chapter, even individuals with large accumulations of wealth, who would benefit from transferring assets during their lifetimes (because the tax treatment of *inter vivos* gifts is more favorable than that of estates), appear to take too little advantage of such opportunities.

Given the factors that influence estate tax design, notably the nature and strength of the bequest motive, the intergenerational correlation of abilities (and preferences for bequests), the underlying distribution of abilities, and of course the other tax instruments available to government, what can one say about the shape of the optimal estate tax? One approach, which side-steps the need to determine the nature of bequest motives, is to adopt a strategy based on "sufficient statistics" for the optimal estate tax schedule, deriving a formula based on observable elasticities (including that of bequests), ability distribution characteristics and social welfare weights. Using this approach, Piketty and Saez estimate optimal inheritance taxes as a function of the size of inheritance received, using parameters based on France and the United States. For both countries, they estimate marginal tax rates that are substantial throughout most of the inheritance distribution (around 50 percent for the United States; higher for France) but drop sharply and become negative within the top bequest quintile (where the externality of wealth transfers outweighs other factors). While this is an interesting finding, one should keep in mind that it comes from a model focusing on intergenerational transfers, with labor income taxes the only other tax instrument, and with no capital income taxes or lifetime intertemporal decisions. While one might cite other arguments for not using lifetime capital income taxes, the fact that such taxes exist certainly affects one's conclusions about the optimal estate tax. Although they are not perfect substitutes, one would expect higher capital income taxes to translate into lower estate taxes, although the analysis is complicated by the fact that capital gains taxes are avoided at death, a fact often used to justify the estate tax as a backstop. Also, in more complex models, there are many margins of taxpayer response that are relevant to the design of the estate tax within the broader tax system.