Crossing the Bar: Predicting Wealth from Income and Estate Tax Records

Lisa Schreiber Rosenmerkel and Jenny Wahl

Draft 10-17-2010

Clarifying the connections between income and wealth is essential for ascertaining individual economic status and establishing informed policy. For the very rich, realized income may reveal little about true well-being because tax liability can drive decisions about form and timing. Fortunately, estate tax records provide a superb additional source of information.

We have constructed a unique data set that links together several years of U.S. Individual Income Tax Returns (Forms 1040) for persons who died between 1996 and 2002, as well as the U.S. Estate and Generation–Skipping Transfer Tax Return (Form 706) when the decedent's estate size exceeded the filing threshold.[1] The included individuals were members of a panel representing the cohort of tax families (primary and secondary filers and their dependents) that filed Form 1040 in Tax Year 1987.[2] We use the data to do three things: (1) predict the probability of filing a Form 706 from Form 1040 information,[3] (2) estimate individual wealth from Form-1040 data via a Heckman two-step approach that corrects for selection bias, and (3) outline an approach for predicting the amount of total gross estate that will ultimately appear on Forms 706.

[1] Jacobson et al. (2007) describes the tax treatment of estates. Total gross estate includes the decedent's assets plus the relevant share of jointly owned and community property. Property over which the decedent had a general power of appointment, most life insurance proceeds, and certain inter vivos transfers are also included.

[2] A tax family is defined as the taxpayer, spouse, and all dependents as claimed on the Form 1040. Married couples who elect to file separately are treated as two distinct tax families. Only the partner whose return was selected into the sample was included in the panel. As a result, the tax family differs significantly from the more common "household" measure used by many national surveys (Czajka and Schirm 1993).

[3] An alternative is predicting the probability of surpassing a given wealth threshold. The threshold for filing estate tax returns changed over time – $600,000 for 1996-97, $625,000 for 1998, $650,000 for 1999, $675,000 for 2000-01, and $1 million for 2002. We report results for both the probability of filing and of exceeding the largest threshold of $1 million (in real dollars).

I. A Behavioral Model: Linking Theory to Estimation

A simple model of individual choice might look something like this

$$\text{Max } U = f(C, L, G) \tag{1}$$

subject to

$$C + G = wH + N \tag{2}$$

and

$$T = H + L, \tag{3}$$

where U is lifetime utility, C lifetime value of consumption, L lifetime leisure hours, G lifetime value of gifts bestowed (including bequests), H lifetime work hours, w hourly wage rate, N lifetime nonwage income (including gifts received and inheritances), and T length of life. We can express the change in wealth at time t (ΔW_t) and wealth at time t (W_t) as follows:[4]

$$\Delta W_t = w_t H_t + N_t - (C_t + G_t) \tag{4}$$

$$W_t = \Delta W_t + \Delta W_{t-1} + \Delta W_{t-2} + \ldots \tag{5}$$

In other words, the change in wealth in period t equals income in period t less consumption plus gifts bestowed at time t. Wealth at time t is the sum of net additions to wealth up to that point (including any unspent inheritances and gifts received). Equations (4) and (5) are basically accounting identities; the amounts people choose to add to wealth and realize as income in any period are of course functions of tax treatment.

Modeling what occurs the last period of life (T) brings home the messiness of uncertainty: a person may not exhaust all resources because he or she mispredicts the date of death. The decedent-to-be may also wish to influence the behavior of potential

[4] ΔW_t can be negative and, in theory, so could W_t. Real-world policy regarding bankruptcy and debt, particularly at the end of an individual's life, could add more constraints.

heirs.[5] This process is complicated by any desire to minimize overall tax liability, which can depend on policy regarding capital gains, gifts, and estates as well as income. Consequently, W_T is likely the outcome of a complex interaction of constraints and preferences about own consumption, own leisure, gift-giving, bequests, tax avoidance, and mistakes. Nonetheless, equations (2), (4), and (5) make clear that wealth at any point in time, including the time of death, is related to prior income:

$$W_T = f(w_T H_T, w_{T-1} H_{T-1}, \ldots, N_T, N_{T-1}, \ldots). \quad (6)$$

Transforming equation (6) to something that could be estimated using tax data requires us to face several issues: (1) complete lifetime income information for a reasonably sized sample of individuals simply is not available, in tax data or anywhere else, (2) tax considerations could imply different allocation and timing of income for people at different points in the life cycle and in different wealth categories, (3) end-of-life wealth amounts are available from tax data only for decedents whose estates surpass a certain threshold size, and (4) the income-wealth relationship for the living population could vary from that associated with a group of decedents. These are significant, but not insurmountable, issues.

To address the first two, we include demographic and portfolio data from a given year's income tax return for each individual.[6] Variables indicating age, gender, and filing

[5] For a discussion of the strategic bequest motive, see for example Bernheim et al. (1985).

[6] We have experimented with using income information from a fixed number of years before death, as well as with income information from several years. See Johnson et al. (2009) for information about income patterns for several years prior to death. Adjusted gross income (AGI) is highly correlated across years for decedents whose estates did not file a 706 return, but less so for estate-tax decedents. What is more, the relationship of gross estate to AGI of whatever year also depends upon macroeconomic conditions. As a consequence, the choice of year for AGI observation could indeed matter for estimation.

Yet our research is motivated in part by the desire to predict the probability of filing an estate tax return (or exceeding a given wealth threshold) from given income tax information but unknown date of death. We decided that this objective is best served by using a cross-section of income-tax returns rather than one that requires knowledge of death dates.

status act as proxies for life-cycle differences.[7] Breaking income into its components – salary, dividends, tax-exempt interest, and the like – allows us to evaluate coefficients and assess implied rates of return on different asset types. This may help us ascertain whether variations in tax treatment affect portfolio composition across different wealth categories, at least indirectly. Equation (6) thus becomes something like this:

$$W_T = \alpha + \beta_i Y_i + \gamma_j D_j + \varepsilon, \tag{7}$$

where Y_i refers to the ith component of income and D_j refers to the jth demographic trait.

James Heckman's seminal research provides a method for us to overcome the third issue: selection bias associated with truncated data.[8] This procedure first estimates a probit model of the form

$$X = a + b_i y_i + c_j d_j + e, \tag{8}$$

where X=1 when an estate tax return is filed (or a given wealth threshold is exceeded) and 0 otherwise, and vectors y and d include relevant income-tax and demographic information.[9] The predicted values from the probit regression are then used to construct an inverse Mills' ratio λ to correct for selection bias in equation (6):

$$W_T = \bar{\alpha} + \bar{\beta}_i Y_i + \bar{\gamma}_j D_j + \phi\lambda + \varepsilon. \tag{9}$$

The second experiment – using multiple years of income-tax data – is part of our continuing work. Here, we have to grapple with the issue of using different numbers of years of information for decedents (depending on the year of death) versus using a given number of years for everyone, the latter of which implies again that we know the year of death. Because many components of income are highly correlated across years, we hope that inferences from a single year of income information will hold if we instead use multiple years.

[7] The work presented here distinguishes people married and filing jointly in 1993 from those who were single, married filing separately, widowed, or separated. We have done some preliminary work that uses filing status information from multiple years but have not yet settled on the best way to incorporate this into the analysis. Estate wealth pertains to the decedent, but income reported on a 1040 could include spousal wages, non-labor income from jointly owned assets, and the like. We have experimented with different ways to cope with this – assigning half the income in the case of joint filers, for example, and analyzing long-married and long-single persons separately. The clearest way to present our current work is simply to include gender and filing status as of 1993 in our regression analysis.

[8] Heckman (1979).

[9] We use lower-case letters for the explanatory vectors because this step must naturally include at least one identifying variable not included in the second step.

The resulting coefficients on Y_i and D_j are unbiased, after correcting standard errors for heteroskedasticity.

What remains to be addressed is the fit of the model for the general population. Our data require us to include only people known to have died, because only they report wealth information that we can observe via Form 706. One way to cope with this is simply to assume that our data are representative as to mortality rates. Then we could gross up estimated wealth for the decedent population using the inverse probabilities of death for particular age groups to obtain wealth for the living population. Because unweighted sample sizes are fairly small, however, this may not be appropriate for age groups in which mortality rates are quite low.[10] What is more, anticipation of death could encourage some individuals – the very old, for instance – to adjust spending patterns to reflect the decreased uncertainty about the end of life.

Besides estimating equation (9) for all decedents in our sample, we therefore also estimate it for two subsamples: decedents who very likely would have died between 1996 and 2002 simply because humans have limited lifespans, and decedents who had a relatively low *ex ante* probability of death. That is, we focus on groups of very old and fairly young decedents to ascertain whether results using these data differ appreciably from those generated by the full sample.

[10] The distribution of AGI for the living population (obtained at http://www.irs.gov/pub/irs-soi/06in05tr.xls) is fairly similar to that for our decedent population, although the age distribution is undoubtedly different. This is not terribly surprising, as the sample is intended to represent the income-tax-filing population, which hopefully includes representativeness in terms of mortality rates. Even so, we cannot be sure that "representativeness" extends to the income-wealth relationship for the entire population, in part because relatively few young people die and even fewer leave large estates.

II. Data

The data used here consist of 8,942 individuals who filed a Form 1040 in 1993 and died between 1996 and 2002; 4,226 of these decedents also had wealth that exceeded the Form 706 filing threshold.[11] By weighting the sample to reflect the population,[12] we find that the number with a Form 706 constitutes about 8.8 percent of decedents.[13]

We obtained the original income-tax data from panel data collected by the Statistics of Income (SOI) Division of the Internal Revenue Service. Each year, SOI obtains a stratified sample of income-tax returns. Several years ago, SOI incorporated into its annual cross-sectional sample a panel component that represents the cohort of tax families filing Form 1040 for the 1987 Tax Year.[14] These data are called the 1987 Family Panel and consisted initially of 89,755 returns.

Starting in 1994, SOI began to include in its annual sample of estate tax returns all Forms 706 for 1987 Family Panel members who died in that year. Between 1994 and

[11] The correlation between AGI and estate wealth is largest for AGI reported 4 years before death for the full sample, 5-6 years for the old subsample, and 3 years for the young subsample. Consequently, we did not want to use income information from the earliest filing years – 1992 is 10 (or possibly 11) years before the date of death for our latest-dying individuals, for example. And Form 1040 filed in the year of death typically reports activity for only part of a year. We therefore did not want to use Forms 1040 filed for Tax Year 1996 or later. What is more, some decedents did not have to file a Form 1040 in the year before death because much of their income went toward deductible medical expenses. This left us with two good possibilities for income-tax filing years: 1993 and 1994. We report results using Form-1040 information from Tax Year 1993 here; results using Forms 1040 from Tax Year 1994 are not substantially different and are available from the authors.

[12] Choudry (2001a, 2001b, 2001c), Czajka and Schirm (1991) and Schirm and Czajka (1992) offer additional information about the weights used in the sample. The weights are the inverse of the probability of initial selection into the 1987 Family Panel. They therefore do not easily translate into something with precise meaning for our sample. Although not optimal for our purposes, the weights still indicate something reasonable about the number of individual filers that a given observation represents. We therefore report all results using these weights, as the unweighted sample is far from representative.

[13] Jacobson et al. (2007) report that fewer than 2 percent of the estates of adult decedents filed a Form 706 from 1916 to 2004, although the figure grew considerably in the 1990s, which may help explain the agitation for reform and ratcheting up of the estate tax filing threshold after 1997.

[14] For additional description, see Schirm and Czajka (1992), Nunns et al. (2008), and Johnson and Schreiber (2006).

2003, SOI gathered 5,557 estate tax returns for 1987 Family Panel members.[15] Over the last two years, we have also identified 1987 Family Panel members who died between 1996 and 2002 but whose estates were not required to file a Form 706.

We have constructed two datasets, one based on income tax returns and the other on individual decedents. Each observation in the return-based data set consists of information collected from a given Form 1040 (plus additional data from the corresponding Form 706 where available) filed between 1988 and the year of death for a Family Panel member who died between 1996 and 2002. Each observation in the individual-based data set encompasses information on all the Forms 1040 (and Form 706 where present) for a given Family Panel decedent. Table 1 summarizes the number of observations (unweighted) in each data set.

TABLE 1: Number of observations in two relevant datasets

	Estate ≥ Form 706 Filing Threshold	**Estate < Form 706 Filing Threshold**	**Total**
1040-return-based data			
(1987-year of death)	55,160	60,061	115,221
Individual-based data			
death year 1996	522	557	1,079
death year 1997	575	564	1,139
death year 1998	541	618	1,159
death year 1999	642	647	1,289
death year 2000	636	700	1,336
death year 2001	668	773	1,441
death year 2002	645	873	1,518
Total 1996-2002	4,229	4,732	8,961

[15] See Johnson and Schreiber (2006) and Johnson et al. (2009) for additional discussion of the Family Panel Decedent Dataset.

The decedents in the individual-based data set ranged in age from 14 to 99.4 years in 1993.[16] By inspecting the distribution of death ages, we find that 10 percent of decedents died by age 52.2, 25 percent by age 65.1, 90 percent by age 89.5, 95 percent by age 92.5, and 99 percent by age 97.4.

A rough method of obtaining a subsample of Family Panel members who very likely would have died between 1996 and 2002 is to put a lower bound on the age of a person filing a Form 1040 in 1993. For our old "likely-to-have-died" sample, we chose a cutoff age of 82.8 for Filing Year 1993. If these individuals filed the 1993 Form 1040 in timely fashion, they would have reached age 92.5 (or the 95th percentile) by the end of Filing Year 2002. This unweighted subsample consists of 956 individuals. For our young "unlikely-to-have-died" sample, we include individuals reporting an age of 55.4 or younger for Filing Year 1993. This unweighted subsample includes 1,509 individuals.

III. Variable Choice and Regression Results for Individuals

The selection issue revisited

Two of our analytical tasks are determining how best to predict from detailed Form-1040 information the likelihood of a later Form-706 filing and to model the relationship of income to wealth for decedents.[17] The first is likely a prerequisite for the second, because unobservable factors affecting the probability of a decedent's surpassing the Form-706 filing threshold could reasonably affect the size of the estate as well. Call

[16] Because income can fall short of the Form 1040 filing threshold, not all decedents filed an income tax return in every year. Of the 8,961 individuals who died between 1996 and 2002, for example, only 8,942 filed a Form 1040 in 1993.

[17] We cannot include in our analysis persons who died in the period 1996-2002 but did not file a Form 1040 in Tax Year 1993 – those omitted may include elderly or retired persons whose 1993 income falls below the zero-bracket amount but who still may have assets. We speculate that omitted persons are unlikely to fall in the upper part of the wealth distribution, however, which will probably be the focus of any analysis using these data.

the individuals whose estates file an estate tax return "F706 decedents" and those whose estates fall short of the estate-tax filing threshold "non-F706 decedents." Once we have a method of determining how an eventual decedent "selects into" F706 status, we can use this to correct for selection bias in a regression that has wealth information only for F706 decedents, as outlined in equations (8) and (9) in section I. The resulting unbiased coefficients can in turn help us predict wealth for non-F706 decedents.[18]

Variation in tax treatment for income earned from different sources means that certain types of assets may appear more attractive to wealthier taxpayers. Assets that generate tax-exempt income or unrealized capital gains (or realizable capital losses) might particularly appeal to richer individuals. Thus, both the presence of a Form 1040 item and its size may help predict the probability of filing a Form 706 and the size of total gross estate.

Descriptive information

Charts 1 and 2 display information about the presence and average size (in $2001) of various Form-1040 items (for Filing Year 1993) for F706 and non-F706 decedents.[19] Virtually all members of both groups report adjusted gross income (AGI). But while nearly half of F706 decedents report tax-exempt interest, for example, fewer than 10 percent of non-F706 decedents do.[20] Mean real AGI is over $128,000 for F706

[18] The techniques we currently have available may not be as refined as we would like. Quantile regression, for example, would allow us to estimate the median (or other quantiles); it is more robust than OLS regression when outliers are present. See for example Hao and Naiman (2007). We hope to extend our analysis using quantile regression analysis once we have the requisite computing software.
[19] Because our Form 706 information comes from returns filed in different tax years, we converted all dollar amounts to dollars of a given year (2001).
[20] For the old subsample, just over 60 percent of F706 decedents report tax-exempt interest income in 1993, whereas only about 13 percent of non-F706 decedents report this item of income. For the young subsample, the figures are 22 and 1.5 percent, respectively.

decedents, about 5 times the mean real AGI for non-F706 decedents.[21] The numbers next to the y-axis labels indicate the ratio of the relevant figures for F706 and non-F706 decedents. For example, 6.64 times as many F706 decedents report tax-exempt income as non-F706 decedents. But the average amount of tax-exempt income by F706 decedents is over 33 times that reported by non-F706 decedents.

CHART 1: Presence of Form-1040 items, by Form-706 filing status

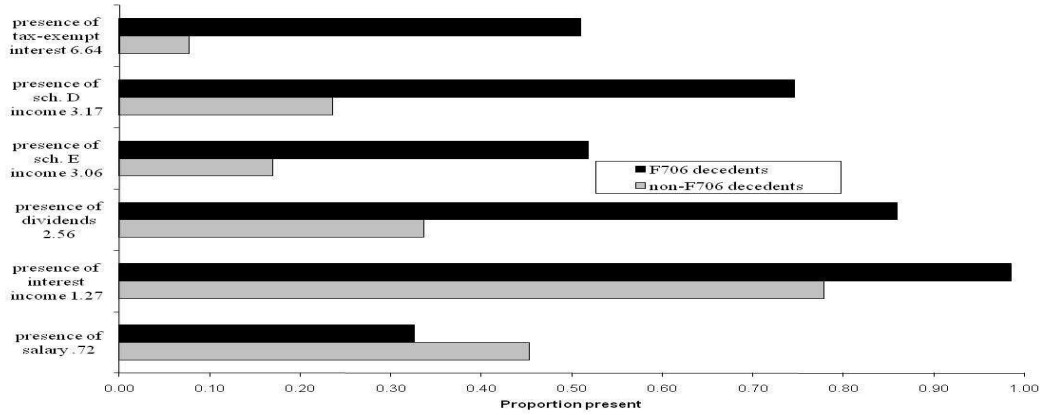
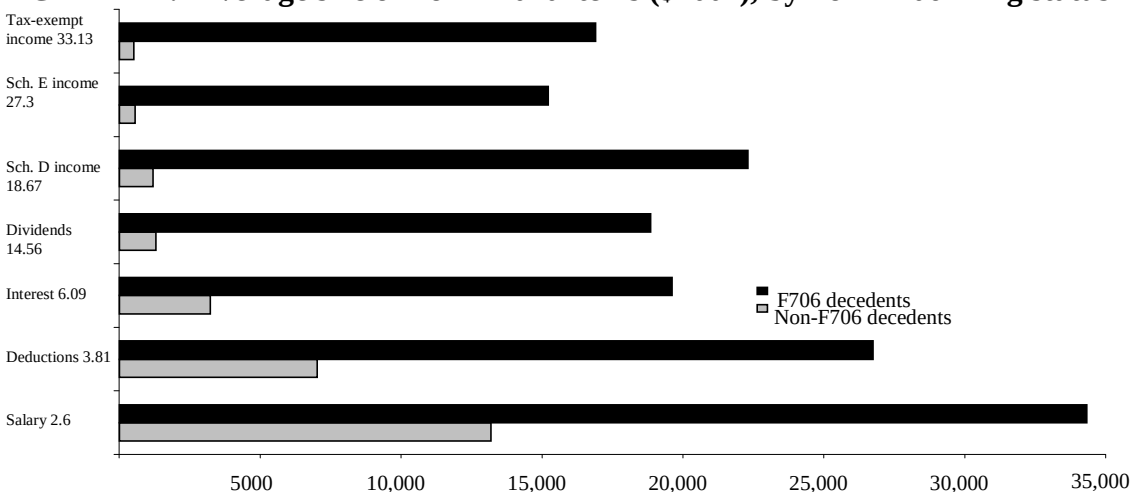


CHART 2: Average size of Form-1040 items ($2001), by Form-706 filing status



[21] Mean real AGI in 1993 for elderly F706 decedents is 4.24 times that for non-F706 decedents; the figure is 5.82 for the young subsample. We do not show AGI on chart 3 because it is so large relative to the other items, making the chart difficult to read.

We have also examined other moments of the distribution, particularly variance and skewness, for several variables. In sum, F706 decedents exhibit higher mean values for many variables (particularly tax-exempt interest and dividend income) as well as greater variability.

Chart 3 indicates the averages of ratios of particular line items to AGI for F706 and non-F706 decedents for the full sample. Note particularly the differences for deductions and for tax-exempt income. The average deductions-to-AGI ratio for non-F706 decedents is 0.795; it is only 0.242 for F706 decedents. The respective figures for the tax-exempt-income/AGI ratios are 0.016 and 0.204.

CHART 3: Averages, ratio of Form-1040 items to AGI, by Form-706 filing status

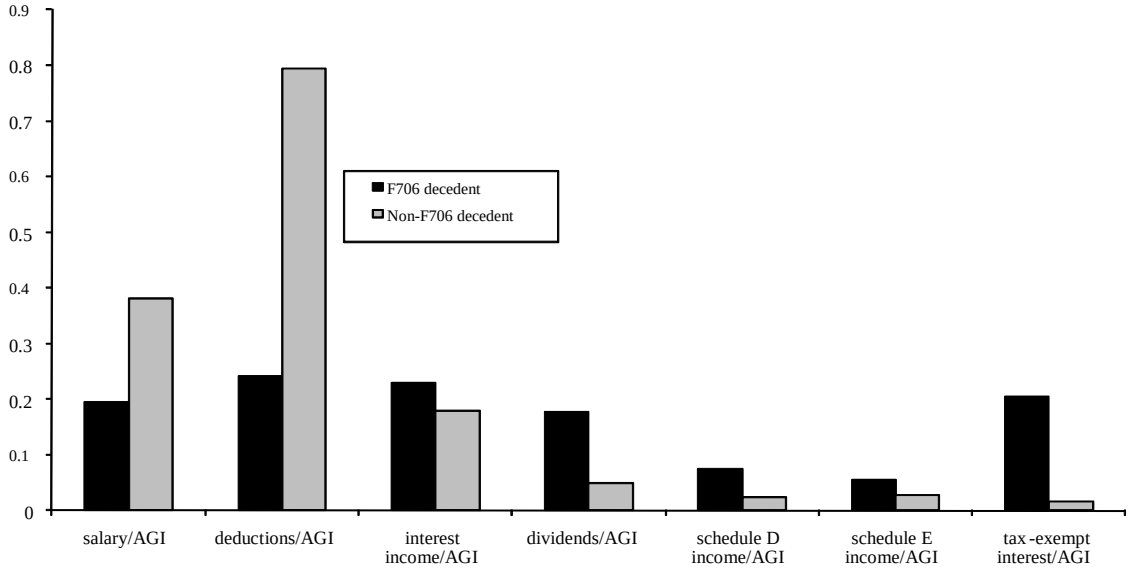


Chart 4 shows the F706 decedent/non-F706 decedent ratios of the proportions shown in Chart 3 for the full sample and the two subsamples for Forms 1040 that reported positive AGI. The F706 decedent/non-F706 decedent ratio for deductions/AGI for the full sample equals 0.30 (0.253/0.856), for instance, and the full-sample ratio for tax-exempt income/AGI equals 13.1 (0.21/0.016). The difference between F706 and non-F706 decedents is especially notable for the young subsample with respect to dividends, capital gains and losses as reported on Schedule D, and tax-exempt interest income. Young F706 decedents with positive AGI in Filing Year 1993 exhibit nearly 16

times the average dividend/AGI ratio as their non-F706 counterparts, for instance, whereas the same figure is only 1.87 for the old subsample and 3.45 for the full sample.

CHART 4: Ratios, F706 decedents to non-F706 decedents (AGI >0)

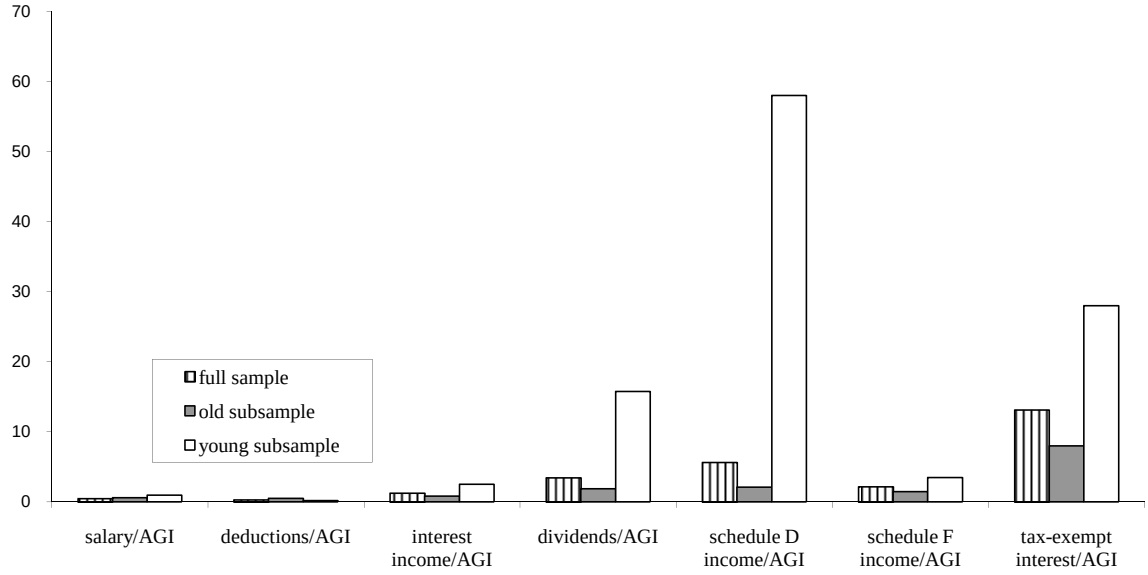


Table 2 gives the percent married filing jointly in 1993 for various demographic groups. The smallest numbers are for females in the old subsample. This reflects the relatively longer lifespan for females and the small average gap in ages for married couples in the U.S.[22] Interestingly, the percentages are all fairly close for men in the two subsamples, although older unmarried men are probably more often widowed whereas younger unmarried men may be more likely to be divorced or never married, relative to the full sample. Another intriguing pattern is the gender difference between F706 and non-F706 decedents in the young subsample. The percent for male F706 decedents is only 8 percentage points larger than for non-F706 decedents, but the corresponding figure is 27 percentage points for females. Although we decline to speculate, we find it

[22] For data on U.S. lifespans, see Shrestha (2006). The average spousal age gap in the U.S. has fallen from about 5 years in 1900 to about 2 years in 2000. Rolf and Ferrie (2008).

fascinating that such a large proportion of relatively young wealthy female decedents were married.[23]

TABLE 2: Percent married filing jointly (Form 1040) in 1993, by gender and Form-706 filing status

	Non-F706 decedents	F706 decedents
Full Sample		
Male	73	77
Female	51	38
Old Subsample		
Male	59	60
Female	21	10
Young Subsample		
Male	59	67
Female	61	88

Step 1: Predicting the probability of filing an estate tax return

The first step of the Heckman method calls for a probit model designed to predict the probability of filing an estate tax return. Given the information presented above, we include the following as independent variables: age of the filer, size of AGI, presence of tax-exempt income, presence of dividend income, gender ("male"=1 for males, 0 otherwise), filing status ("married"=1 for married persons living with spouses in 1993, 0 otherwise), and an interactive variable ("male"*"married") to account for potentially different effects of marital status upon men and women. Interpreting the coefficients on a

[23] Henry James's novel *The Wings of the Dove* springs to mind.

probit regression can be challenging, so we instead present the odds ratios from a similar logit regression in Table 3.[24]

TABLE 3: Odds ratios, logit regressions for probability of estate exceeding the Form-706 filing threshold

	Males		Females		Overall		
	Single	**Married**	**Single**	**Married**	**Full**	**Old**	**Young**
Filing age	0.98	1.00	1.04	1.00	1.01	0.96	1.06
AGI	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Presence of tax-exempt income	2.87	4.33	3.41	4.15	3.96	3.45	3.72
Presence of dividend income	23.76	3.39	3.91	7.19	5.05	2.76	8.94
Male					0.71	0.51	6.88
Married					0.35	0.18	1.26
Male*married					1.87	3.40	0.39
Percent concordant	85.5	86.7	89.8	86.4	86.6	88.1	85.7
N	1120	4280	1348	2194	8942	956	1509

Table 3 shows two things: we can use a parsimonious set of regressors to predict accurately the probability of filing an estate tax return,[25] and we observe different odds ratios depending upon gender and filing status (and, to some extent, age group), particularly for the variable indicating the presence of dividend income. For example, single males with dividend income are more than 20 times as likely to generate an estate tax return as their counterparts with no dividend income, *ceteris paribus*. The corresponding figure is far lower for all other groups. In the wealth estimation analysis, we aggregate the data for purposes of estimating the selection bias variable (λ). To

[24] All coefficients are significant. We report here the weighted results using actual AGI as a regressor. The odds ratio indicates that a person with $x of AGI is just as likely to file a Form 706 as a person with $(x+1), all else constant – an extra dollar of AGI simply is not that influential. But, if we use ln AGI as a regressor, the odds ratio ranges from 2.08 to 2.87 – that is, an extra 1 percent of AGI does matter in predicting the likelihood of Form-706 filing.

[25] We experimented with different regressors but found that several Form 1040 items are correlated with other Form 1040 items. We also had to determine the placement of variables into the two steps of the regression. Ultimately, we decided that the presence of certain items on the Form 1040 (dividends and tax-exempt interest) seemed more important for predicting the likelihood of filing but the size of various items more plausibly belonged in the wealth regression.

evaluate policies regarding the likelihood of exceeding the Form-706 filing threshold, however, these results suggest that analysts might want to conduct separate studies by gender and filing status.[26]

Step 2: Predicting wealth

Table 4 shows coefficients from the wealth regression for the full, old, and young samples.[27] The first two columns both pertain to the full sample; the first column includes a λ constructed from the probability of exceeding the Form-706 filing threshold and the second a λ created from the probability of exceeding the maximum filing

TABLE 4: Wealth Regression

	Full sample (x706)*	Full sample (xthresh)**	Old subsample*	Young subsample*
Intercept	ns	3,657,754	ns	11,945,626
Dividends	69.52	69.06	35.31	18.45
Interest	ns	ns	22.62	ns
Tax-exempt interest	7.09	6.00	19.46	ns
Deductions	24.70	24.78	12.08	24.13
Age at filing	ns	-51,129	ns	ns
Sch. D income	ns	ns	-4.19	-1.40
Sch. E income	4.01	3.84	2.86	4.80
Male	ns	ns	1,719,209	-7,859,775
Married	-1,525,068	-1,375,439	ns	-11,527,675
Male*Married	ns	ns	ns	8,792,592
λ	1,192,116	510,842	ns	ns
Adjusted R-squared	0.482	0.481	0.735	0.276

Note: "ns" means not statistically significant.

*λ is constructed from a probit model that uses probability of exceeding the current-year filing threshold as the dependent variable.

**λ is constructed from a probit model that uses probability of exceeding the maximum filing threshold ($1 million in 2002) as the dependent variable.

[26] Recall, however, that "married" indicates filing status (married filing jointly) in Tax Year 1993. "Single" filers include not only long-single persons but also recently widowed individuals, whose income and wealth patterns might more closely resemble those for "married" persons. In future research we hope to refine our marital-status indicator.

[27] We use weights in the regression, but we report the underlying unweighted number of observations.

threshold for gross estate ($1 million in 2002). Table 5 offers means and standard deviations for the variables in the wealth regression.

TABLE 5: Means and Standard Deviations for Wealth Regression Variables

	full sample		old subsample		young subsample	
	mean	**S.D**	**mean**	**S.D.**	**mean**	**S.D.**
gross estate	2,332,377	211,061,905	2,381,685	149,374,390	3,525,805	103,103,314
dividends	18,841	1,607,562	24,497	2,126,588	7,689	1,131,482
interest	19,611	1,292,951	25,621	1,927,411	8,724	940,353
tax-ex.int.	16,896	1,124,990	22,396	1,282,347	4,081	398,406
deductions	26,742	1,777,739	25,360	2,283,583	39,247	1,319,098
age at filing	71.47	151.18	86.46	41.5	48.81	70.51
sch. D inc.	22,301	4,216,457	16,328	2,696,479	32,567	6,938,998
sch. E inc.	15,212	3,681,648	10,142	5,368,819	29,036	4,206,000
male	0.53	6.19	0.34	7.54	0.76	5.06
married	0.59	6.11	0.27	7.09	0.72	5.36
male*married	0.41	6.1	0.2	6.41	0.51	5.95
λ	1.4	7.27	1.14	8.52	1.41	8.41
N (unweighted)	4,226		542		533	

Note: All dollar amounts are reported in constant (2001) dollars.

In the full-sample regressions, note particularly the large coefficients on dividend income. They indicate that $1 in additional dividend income yields nearly $70 of estate wealth, or an estimated rate of return on dividend-bearing assets of only about 1.4 percent.[28] Compare this to an estimated rate of return (column 1) on tax-exempt assets of 14.1 percent and on assets yielding Schedule-E income of 24.9 percent. Of course, these assets do not truly generate these rates of return; the coefficients correspond to realized returns.[29] What we find for dividends is in line with previous research; the comparatively

[28] The value of a consol equals its coupon divided by its rate of return. A crude way of estimating the rate of return on the underlying asset generating dividends, then, is to act as if an asset worth $69.52 generates $1 of dividends in perpetuity, thus implying a rate of return equal to 1/69.52 = 1.4 percent.

[29] Note that the implicit yield on tax-exempt interest income for the old subsample – about 5.1 percent – is reasonably close to posted yields during the time period in question. The yields on state and local bonds

low realized return suggests that people may choose investments in part to time income for tax purposes.[30] The next section explores this possibility in greater detail.

Note as well the difference in coefficients on dividend income for the full, old, and young samples. The implicit rate of return is greatest for the young subsample, followed by that for the old subsample. Although we cannot definitely say why, these results suggest that these decedents chose investments that yielded relatively more immediate cash, perhaps to pay off mortgages and child-rearing expenses (younger decedents) and medical bills (older decedents).[31]

The coefficients on Schedule-D income, Schedule-E income, and deductions are also worth discussing. Schedule D reports capital gains and losses. What the negative coefficients for the old and young subsamples may imply is that wealthier people, especially those close to death and those in prime working years, may take more advantage of the timing of capital losses.[32] What is more, the step-up in basis at death for assets with accrued capital gains means that the elderly may tend to avoid realizing gains, thus saving their heirs future capital gains taxes.[33] The large implicit yield on Schedule-E income (income from rental real estate, royalties, and partnerships) is consistent with our knowledge of the increasing importance of limited partnerships over the period 1989-2004 (Jacobson et al. 2007). That the yield is especially large for the old subsample may

---

from 1993 to 2002 ranged from a low of 4.85 in 2002 to a high of 5.95 in 1999. http://www.federalreserve.gov/releases/h15/data.htm.

[30] Johnson et al. (2009).

[31] The latter is also suggested in Johnson et al. (2009).

[32] The step-up in basis at death for assets with accrued capital gains does not work in reverse – accrued capital losses have no benefit for heirs. Consequently, wealthier people – especially those who anticipate leaving a large estate fairly soon– could find it especially advantageous to realize accumulated capital losses during their lives.

[33] This may especially be true for decedents whose spouses inherit the bulk of the estate. Because spousal bequests are fully deductible, accrued capital gains do not generate a tax burden via the estate tax for these heirs.

indicate that some of these decedents were holding onto assets that enabled them to retain control over a family business. The positive coefficient on deductions indicates that wealthier people take more deductions – not too surprising, as richer individuals are more aware of deduction possibilities, deductions include items likely to be larger for wealthier taxpayers (including other taxes, mortgage interest, and charitable contributions), and deductions in this regression may act partly as a proxy for AGI.[34] The relatively smaller coefficient for the old subsample could reflect a weaker connection between housing and terminal wealth, perhaps due to downsizing or mortgage payoffs among the elderly.

The significance of the coefficient on λ in the full-sample regressions tells us that selection bias is indeed an issue. Obtaining unbiased coefficients on the independent variables in equation (7) requires us to use the Heckman two-step method, as outlined in equations (8) and (9). That the coefficient on λ is positive indicates, not surprisingly, that unobserved factors positively associated with estate wealth are also positively associated with the probability of filing a Form 706.

Interestingly, the coefficient on λ is not significant for the old and the young subsamples. The argument we put forth to explain the coefficients on dividends could apply here as well: realized income and underlying wealth more closely match for people who have significant current out-of-pocket expenses – those with children at home, mortgages to pay, or large medical bills. Observable factors thus do a good job at predicting both the probability of exceeding the Form-706 filing threshold and the size of wealth for the relatively old and the relatively young. For the elderly, anticipation of death may also alter income realization patterns. Knowing that you can't take it with

[34] The original data included an amount for itemized deductions for itemizers. We assigned the standard-deduction amount to non-itemizers.

you, and knowing that you don't have much more time to enjoy your wealth, could mean that realized AGI (and thus current spending) more closely mirrors underlying wealth for older persons. For the young, we note that the coefficients on "male" are large (in absolute value) in both stages as compared to the same coefficients for the full sample. This observed trait may perform especially well in partitioning the data so as to reduce selection bias.

The differing results for the full, old, and young samples suggest that constructing the wealth distribution for the living population from that for the dead could be complicated. Accounting for the differences in the degree of uncertainty about impending death could be part of this process. Our future research will grapple with the modeling of the income-wealth relationship for living adults at the extremes of the age distribution.

An aside on income timing and wealth

Previous research suggests that wealthier people time the receipt of income so as to minimize tax liability, pointing to the lower realized rates of return on various income categories associated with higher-wealth individuals.[35] Our work reinforces those findings.

Table 6 offers regression results for F706 decedents from different wealth percentiles. Column 1 includes only F706 decedents whose gross estate fell in the top 90 percent, column 2 includes only those with gross estate in the top 50 percent, and so

[35] See Johnson et al. (2009) and Steuerle (1985).

forth.[36] The generally increasing size of the coefficients from left to right across the table for dividends, tax-exempt interest, and Schedule-E income suggests that, the wealthier the individual, the lower the realized rate of return on the assets generating these sorts of income.

TABLE 6: Wealth Regressions by Percentile of Wealth

	Top 90%	Top 50%	Top 10%	Top 5%	Top 1%
Intercept	4,740,924	7,092,792	19,082,851	30,882,135	83,423,691
Dividends	69.33	69.62	71.49	72.63	77.80
Interest	5.44	5.45	ns	ns	ns
Tax-ex. interest	6.24	6.53	7.80	7.96	ns
Deductions	24.48	24.53	24.23	23.76	22.99
Age at filing	-59,763	-90,182	-239,437	-375,011	-1,074,325
Sch. D income	ns	ns	ns	ns	ns
Sch. E income	3.52	3.57	3.64	3.54	3.98
Male	ns	ns	5,365,229	ns	ns
Married	-1,405,757	-2,176,530	-5,808,024	-10,279,937	-22,606,035
Male*married	ns	ns	ns	ns	ns
Adj. R squared	0.481	0.480	0.480	0.480	0.490
N (unweighted)	4,126	3,664	2,350	1,737	1,043

Note: These regressions include only observations for which a Form 706 was present, so no λ variable appears.

IV. Practical Considerations: Estimating Wealth Reported on Forms 706

Thus far, the analysis has focused on wealth estimation across the entire spectrum. But policy makers might be interested in a different question: can our models help predict the amount of wealth that will be reported on Forms 706? This estimate in turn will indicate something about the amount of wealth that will be transferred – often across

[36] Because these regressions include only those known to have filed an estate tax return, they do not have a λ variable. Note that these regressions, like the wealth regressions above, include weights.

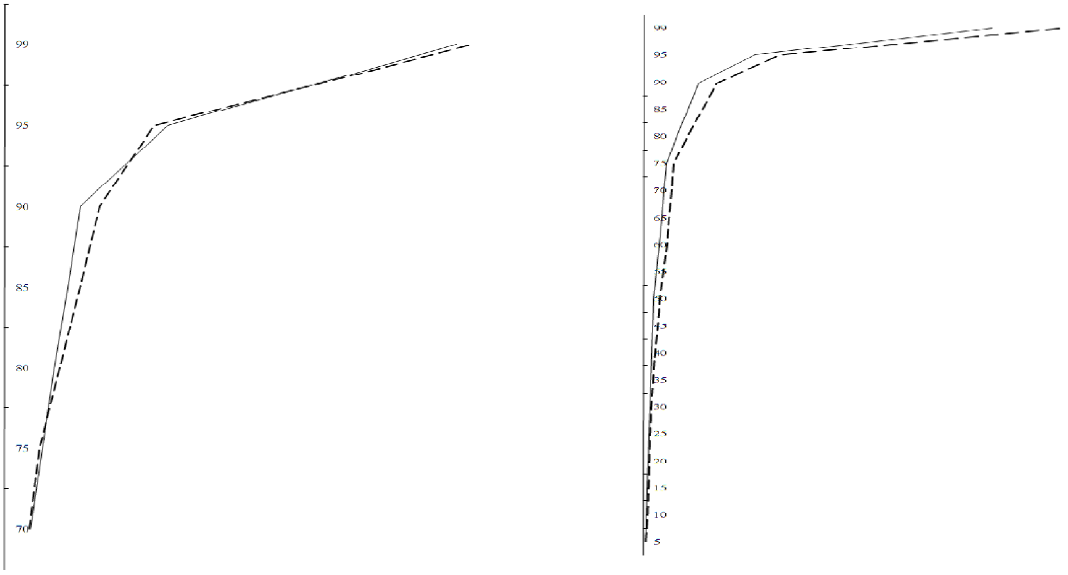
generations -- during a certain period of time, as well as the size of the expected estate tax base.

Because we use estate tax returns from several different years, and because the filing threshold increased over the period, we conduct this part of the analysis using information about whether a particular decedent left at least a certain amount of gross estate rather than whether his or her estate filed a Form 706. We have only a finite period for which we obtained estate tax returns; thus, we do not capture wealth information for those who died after 2002. Effectively, what we estimate is wealth associated with those who filed a Form 1040 in 1993 and who died between 3 and 9 years later with gross estate of at least $1 million in real dollars.

We proceed the same way in this section as we did in the previous one – first predicting the probability of exceeding a given wealth threshold and then estimating the amount of wealth for each individual. To see how well our model predicts the total amount of gross estate reported in the period 1996-2002, we construct cumulative distribution functions (cdfs) of actual and predicted wealth. Because of the truncation of the actual wealth distribution, however, a straight comparison of values may not be particularly informative. Recall that the first step of the model generates a predicted probability. People either leave an estate exceeding a given wealth threshold or they don't; one way to cope with the truncation issue is to look only at actual and predicted values of wealth for people whose predicted probability exceeds a certain size. A larger cutoff means excluding more people whose estates actually do surpass the threshold; a smaller cutoff means including more people whose estates actually will fall short of the threshold – in other words, a classic type I-type II error tradeoff.

In each graph in Chart 5, the predicted gross estate is denoted by the dotted line and the actual gross estate is the solid line. The left graph indicates the cdfs for predicted and actual wealth of all persons whose predicted probability of filing exceeded 0.1. The right graph includes only cdfs of those for whom the predicted probability

CHART 5: Cumulative density functions, actual and predicted wealth



Predicted probability ≥ 0.1 Predicted probability ≥ 0.5

exceeds 0.5. Because we have no information on actual wealth for persons whose estates did not file a Form 706, we only report the percentiles for which we can make a meaningful comparison.[37]

The choice made for the threshold predicted probability depends on how much of the distribution one wishes to model. Suppose only the very top of the wealth

[37] For example, only the top 30 percent of persons whose predicted probability of filing is at least 0.1 were actually required to file a Form 706. The higher the threshold for predicted probability, the greater percentage of persons who were required to file – which explains why the vertical axis extends farther down for the right-most graph. Note that we only map up to the 99th percentile – actual wealth for the very top wealth-holders far exceeds predicted wealth, so the cdfs intersect again, between the 99th and 100th percentiles. We do not show this, however, because doing so would compress the graphs so much at the lower percentiles that no distinction between the cdfs would be visible.

distribution – say the top 5 percent -- is of interest. Choosing a low threshold makes sense – this will capture virtually all of the actual F706 decedents; the captured non-F706 decedents are likely to generate predicted wealth that falls below the top of the distribution. As the left-most graph shows, imposing a threshold of 0.1 for the predicted filing probability will yield wealth that is slightly overpredicted at the 90th percentile, slightly underpredicted at the 95th percentile, and slightly overpredicted at the 99th percentile.[38]

V. Conclusions

The research presented here suggests that linked estate and income tax records offer a promising data source for investigating a variety of important economic and policy issues. These include predicting whether an individual's terminal wealth will exceed a given threshold, imputing wealth from income and demographic information (particularly for high-wealth taxpayers), determining the degree of income realization across different wealth classes and age groups, modeling non-compliance, understanding unintended consequences of the estate tax, and estimating the potential tax base associated with an estate tax.

Our work reveals that we can accurately predict the probability of a decedent's estate filing a Form 706 from a relatively small set of Form-1040 information. We find that adjusting for this selection issue is important if we wish to estimate wealth from data on income. We also show that portfolios differ significantly across wealth classes, and

[38] Technical considerations made including the horizontal axis on the graph difficult. Here are the numbers for actual (predicted) gross estate for the various percentiles for the left-most graph: 90th percentile $2,993,263 ($2,372,494), 95th percentile $4,695,133 ($5,127,255), 99th percentile $14,645,974 ($14,164,627).

that people with greater wealth tend to have smaller realized yields on their assets. This strongly suggests that income data underestimate the differences in true well-being across individuals, and that wealth (or imputed wealth) – particularly for high-wealth people – presents a particularly useful alternative. The information provided on linked Tax Forms 706 and 1040 undeniably provide a singular source of data for mapping the connections between income and wealth.

SOURCES

Bernheim, Andre Shleifer, and Lawrence Summers. (1985) The Strategic Bequest Motive. *Journal of Political Economy* 93(6), 1045-76.

Choudhry, G. Hussain. (11 June 2001) Development of Family Panel Weights for Tax Years 1994-96. Unpublished internal document prepared by Westat for Internal Revenue Service, Statistics of Income Division.

_____. (13 June 2001) Development of Family Panel Weights for Tax Years 1987-1993. Unpublished internal document prepared by Westat for Internal Revenue Service, Statistics of Income Division.

_____. (13 June 2001) Evaluation of Family Panel Weights for the Tax Year 1987. Unpublished internal document prepared by Westat for Internal Revenue Service, Statistics of Income Division.

Czajka, John, and Alan Schirm. (1991) Cross-Sectional Weighting of Combined Panel and Cross-Section Observations, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC.

Czajka, John L and Bonnye Walker (1989) Combining Panel and Cross-Sectional Selection in an Annual Sample of Tax Returns. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Hao, Lingxin, and Daniel Naiman. (2007) *Quantile Regression*. Sage.

Heckman, James. (1979) Sample Selection Bias as a Specification Error. *Econometrica* (47), 153–61.

Jacobson, Darien, Brian Raub, and Barry Johnson. (Summer 2007) The Estate Tax: Ninety Years and Counting. *Statistics of Income Bulletin*, Internal Revenue Service.

Johnson, Barry, Kevin Moore, and Lisa Schreiber. (2009) The Income-Wealth Paradox: Connections between Realized Income and Wealth Among America's Aging Top Wealth-Holders. 2009 IRS Research Conference.

Johnson, Barry, and Lisa Schreiber. (2006) Creativity and Compromise. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Nunns, James, Deena Ackerman, James Cilke, Julie-Anne Cronin, Janet Holtzblatt, Gillian Hunter, Emily Lin, and Janet McCubbin. (2008) Treasury's Panel Model

for Tax Analysis. *OTA Technical Working Paper 3*. U.S. Department of Treasury, Office of Tax Analysis. http://www.treas.gov/offices/tax-policy/library/OTAtech03.pdf.

Rolf, Karen, and Joseph Ferrie. (2008) The May-December Relationship Since 1850: Age Homogamy in the U.S. Working Paper. http://paa2008.princeton.edu/download.aspx?submissionId=80695.

Schirm, Alan, and John Czajka. (1992) Weighting a Panel of Individual Tax Returns for Cross-Sectional Estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Shrestha, Laura. (2006) Life Expectancy in the United States. CRS Report for Congress, http://aging.senate.gov/crs/aging1.pdf.

Steuerle, Eugene. (1985) Wealth, Realized Income, and the Measure of Well-Being. In Martin David and Timothy Smeeding, eds., *Horizontal Equity, Uncertainty, and Economic Well-Being*. National Bureau of Economic Research.