Weighting Panel Data for Longitudinal Analysis

John L. Czajka and Larry M. Radbill
Mathematica Policy Research, Inc.

This paper addresses problems of weighting associated with the longitudinal analysis of multi-person units. The paper focuses on a specific type of unit, the tax filing unit, represented by the tax return as a unit of observation. We identify alternative approaches to weighting and analyzing longitudinal tax return data and present examples illustrating the application of two rather different methodologies to a research question in the area of tax policy.

First we discuss some of the problems that arise in viewing tax filing units in a longitudinal context. The next section outlines alternative strategies that have been suggested for analyzing multi-person units over time. Then, we describe an IRS panel database that we use to illustrate the application of two of these strategies. The section which follows poses a research question for longitudinal tax return data and describes the two approaches that we employ to answer this question. Finally, we present and discuss our empirical findings.

■ Longitudinal Tax Units

Longitudinal data are used to track the actual experiences of specific "actors" over time. Commonly studied actors include countries, sub-national regions, firms, households, families, and persons. In the area of tax analysis, one actor of interest is the tax filing unit.

A tax filing unit is that collection of persons (tax filers) who either choose or are required to report their income and pay their taxes as a single entity. For our purposes, a filing unit may consist of a single filer (who is not married), a married couple filing a joint return, or a married person filing a return separate from his or her spouse. Ordinarily, filing unit data cannot be disaggregated by person.

Analyzing tax units at a single point in time presents no unusual conceptual difficulties. Such units are the tax policy equivalents of families or households. Potential difficulties arise as soon as we move to a longitudinal context: tax units can change composition over time, and those changes are often accompanied by changes in tax filing status. This happens when:

- ❑ Married persons divorce
- ❑ Married persons who have filed joint returns choose to file separately
- ❑ Married persons who have filed separately in the past choose to file jointly
- ❑ One spouse of a married couple dies
- ❑ Single persons marry.

Each of these changes in tax unit composition has implications for the construction of a longitudinal database and for the analysis of taxpayer experiences over time.

■ Strategies for Analyzing Multi-Person Units Over Time

Alternative strategies suggested for the construction and analysis of longitudinal household and family data are relevant to tax filing units:

- ❑ Limit analyses to those units which do not show any change in composition
- ❑ Define new units with every change in composition and weight them by duration of existence in the panel
- ❑ When units combine or separate, select one partner at random to represent the prior or post-history
- ❑ Analyze person-level data, with weights inversely proportional to unit size
- ❑ Analyze person-level data, treating unit characteristics as contextual variables.

Each of these approaches has certain strengths and weaknesses, which vary with the problem being investigated.

The first approach introduces potentially substantial sample selection biases: those units which have stable composition over the duration of the panel are also likely to have more stable income and tax experiences than units which change in composition. Since most of the interest in longitudinal data stems from an interest in change, this seems a largely unacceptable strategy.

The second approach, suggested as a means of counting families in poverty, retains information about units with changing composition as well as those with stable composition. Because of the time-weighting that has been advocated by some (Citro, Hernandez, and Herriot, 1986), the unit of analysis becomes the tax-unit-year equivalent. Those tax units that have stable composition for longer periods of time represent more of the aggregate tax-unit experience over the duration of the panel and so receive greater weight. This strategy may sacrifice some potentially important information. While all tax units with changing composition are retained, the ability to link tax units and study those changes may be limited. Even when information allowing such links is retained, there may be no unambiguous way to generate longitudinal tax-units per se. This is especially clear in the case where a single tax unit splits into two units: in these cases, the "parent" tax unit has two distinct "futures."

The third solution solves the problem of understanding longitudinal tax units when two combine into one or when one splits into two. There are three tax units to work with: the single combined unit -- unit A -- and the two "child" (or donor units) -- units I and II. To construct an unambiguous time-line for a tax unit, either I or II is chosen to represent the post (or pre-) change experience of the combined tax unit and linked to A. It is important that the choice between units I and II be made at random: those filers whose SSNs occupy the primary position in the return for unit A are likely to be systematically different from those whose SSNs occupy the secondary position. This approach retains information about tax units with changing composition in a way that allows the direct analysis of change. The only apparent cost of this approach is the loss of information (or at least the failure to use information) about the child (or donor) tax unit not chosen when forming the longitudinal tax unit.

The fourth approach addresses this shortcoming, retaining all available information about the full longitudinal experience of tax units that change composition over time. While the unit of ***analysis*** is the tax filer, appropriately scaling the case weights allows the unit of ***measurement*** to be the tax unit.

The approach can best be understood by first considering the cross-sectional analog, wherein tax unit composition can be regarded, readily, as fixed. In this case, each sampled tax unit represents a single tax (and income) "experience" for that tax year. If a tax unit was sampled from all tax returns with probability p, its tax experience represents those of (1/p) tax units in the population. If the tax return was for a married couple filing jointly, the experience of each of the two filers on the return represents that of (1/2)*(1/p) tax units in the population. (Note, though, that each filer represents (1/p) filers in the population.) The fact that the two filers on the joint return were sampled as a unit and have identical tax (and income) characteristics means only that, in the cross-sectional context, the second filer provides no information about the population of tax units that was not already learned from the first filer. When constructing a database for analysis we could create two records for each joint return, one representing each of the two filers, and assign each of the two records a weight of (1/2)*(1/p). Doing this would introduce no biases into any parameter estimates based on the data. Because the second filer provides no new information about the sampled tax unit (or, by implication, the population of tax units), however, there is no reason to create the second record.

Little changes in the longitudinal context. The tax (and income) experience of a tax unit sampled with probability p continues to represent (1/p) tax units in the population. If a sampled return belonged to a married couple filing jointly, the tax experiences of each of the two filers still represent those of (1/2)*(1/p) tax units in the original population. However, in cases where tax units split into two units (due to divorce or a decision to file separately), the two filers have different tax experiences over time. In this case the second filer provides information about the population of tax units ***not*** already learned from the first filer. Because the second filer provides new information about the population there is good reason to create separate records for each of the two filers from the original joint return and assign each the correct tax unit weight of (1/2)*(1/p).

Creating two records from a single (joint) tax return does complicate the computation of standard errors. Doing this turns what began as a simple stratified sample into a stratified cluster sample, where clusters are defined as pairs of primary and secondary filers on joint returns. Standard error estimation must take account of the lack of independence of the two observations within each cluster. As long as the two filers continue to file as a single joint unit there is complete lack of independence: the two filers really do provide only a single sample observation. When the two file separately (either as a married couple or after divorce), each is a separate sample observation representing different (populations of) tax units, but the two observations are still (at least partly) correlated with each other. Conventional techniques for standard error estimation in stratified cluster samples should apply directly to this case.

The only difference between strategies four and five lies in the definition of the unit of ***measurement:*** the tax ***unit*** versus the tax ***filer***. The fifth strategy is identical in construction to the fourth except in the choice of weights. While the fourth approach assigns ***tax unit*** weights of (1/2)*(1/p) to each member of a jointly filed tax unit, the fifth approach assigns ***tax filer*** weights of (1/p) to each member of a jointly filed tax unit. This change in weighting forces a change in the interpretation of resulting analyses: the fourth approach allows for the measurement of ***tax units*** while the fifth approach provides measures for the individual ***tax filer***. The choice between these last two designs is determined by the specific analytic or modeling task at hand.

■ The Sales of Capital Assets Panel

The 1985 Sales of Capital Assets (SOCA) Panel is a subsample of the 1985 Statistics of Income (SOI) sample of individual tax returns. The cross-sectional sample in 1985 included 121,418 returns. From these returns, a stratified probability sample of 12,980 was selected to form the SOCA Panel. All primary and secondary taxpayers listed on these returns were designated as members of the 1985 SOCA Panel and identified by their social security numbers (SSNs). In each subsequent processing year, every tax return that contained a SOCA SSN in either the primary or secondary position was captured for the panel. Along with the data items abstracted for the annual SOI sample, the IRS collected supplemental data on individual transactions associated with the sale of capital assets.

■ Measuring the Concentration of Capital Assets Sales Over Time

A question posed by staff of the SOI Division provides the example on which this paper focuses: To what extent are sales of capital assets (as reported to the IRS) concentrated among the same set of filers, year after year?

We elected to operationalize this question with reference to a fixed, prior year -- specifically, 1985, the base year of the SOCA Panel. We then posed two questions:

❑ What proportion of filers/returns with capital transactions in a given, later year also had transactions in 1985;

- ❑ What proportions of the total gains and losses in a given year are attributable to units that reported transactions in 1985?

To evaluate these questions requires only aggregate tabulations, calculated contingent on a prior year binary variable -- specifically, the presence or absence of a transaction in 1985.

We considered two approaches to the longitudinal analysis of tax filing units: limiting the analysis to units with fixed composition over time, and treating filing unit characteristics as attributes of panel individuals, which enables us to conduct our analysis on individual filers rather than filing units. Policy analysts in the Treasury Department have used the former approach in some previous analyses of tax return panel data.

Data preparation for an analysis limited to units with fixed composition involves first defining fixed composition and then constructing a longitudinal record for each filing unit that includes all observations for that unit up to but not including the year that composition changes. Obviously, the first step in the application of this methodology involves critical choices. These include defining operationally what constitutes a change in composition and determining how missing observations (for which composition cannot be observed) are to be treated.

Longitudinal analysis of filing units with stable composition is straightforward. The base year weight applies to a filing unit's entire, stable history. In a given year the sample of filing units represents the population of units with fixed composition through that year -- or possibly a later year, if one set of weights has been defined to serve analyses of different durations.

Treating filing unit characteristics as attributes of panel individuals involves creating a longitudinal record for each ***filer***, where a joint return yields two filers, and assigning the weight of the base year return to each filer's entire history. In a given year the sample of filers weights up to the population of survivors of the base year filers. New filers (that is, those who did not file in 1985) who marry 1985 filers are not counted.

To avoid double counting when tabulating returns or dollars for a given year, it is necessary to multiply the filer's base weight by the filer's share of the unit size in that year. Commonly, this share is either .5 or 1.0, consistent with unit sizes of two or one, but these fractions could vary. There may be reason to give a primary filer a greater share of the filing unit's characteristics, for example.

To produce the estimates reported for the second method in this paper, we employed an alternative operational scheme, which did not require that we create person-level records. This option was available because of the simplicity of the research questions that we were addressing. First, we attached the 1985 gains status, at the person level, to each subsequent return on which a given filer appeared. This enabled us to retain the cross-sectional design of the database, consisting of one record per return per filing period. Next, we weighted each joint return after 1985 by the "equal person method" (Kalton and Brick, 1994). This involved assigning the 1985 base weight to each filer, with nonpanel filers receiving weights of zero, and then averaging the two weights to obtain unit weights.

To define the 1985 gains status at the return level (for joint returns), we used the panel member's 1985 status. If a joint return included two panel members who did not file jointly in 1985 and who had different gains statuses in 1985, we based the assignment on the primary filer. The few instances in which this situation arose made our decision to use this versus another strategy inconsequential.

An advantage of treating unit characteristics as attributes of individuals, however this is operationalized, is that this approach uses all of the data. Furthermore, the manner of discounting some of the information (with fractional weighting) is consistent with the widely used, equal person method of weighting panel data for cross-sectional estimation. Thus, our estimates of the shares of transactions attributable to persons with prior transactions are based on all of the gains and losses reported by the survivors of the 1985 filing population, rather than just a nonrepresentative subset.

■ Empirical Findings

Change in Unit Composition Over Time

Table 1 displays the percentage of base year SOCA Panel filing units with unchanged composition as of each tax filing year, 1985 through 1991. The unweighted percentage declines by about four points per year, reaching 74.5 percent in 1991. In other words, 25.5 percent of the base year filing units changed composition or stopped filing (see below) in the six years following 1985. Weighting accentuates the changes in composition. For the population represented by the SOCA Panel, we estimate that only 58.7 percent of the base year filing units continued to file through 1991 with no change in composition. The implication is that the high weight or lower income filing units were more likely to experience a change in composition than were the units with low weight (high income).

Table 1.--Percentage of 1985 Tax Filing Units with Unchanged Composition, by Filing Year: 1985-91

Filing Year	Unweighted	Weighted
1985	100.0%	100.0%
1986	97.7	88.0
1987	90.1	80.5
1988	86.4	74.5
1989	82.7	69.0
1990	79.1	64.9
1991	74.5	58.7

Table 2 disaggregates the base year filing units by their base year filing status and displays for each filing status the weighted percentage with unchanged composition by year. The status "single," which includes persons filing as head of household as well as those with no dependents, encompassed 52.0 million filing units in 1985, compared to 48.6 million for married couples filing joint returns. An additional 1.0 million filing units consisted of married persons filing separately from their spouses.

Table 2.--Percentage of 1985 Tax Filing Units with Unchanged Composition, by Filing Year and 1985 Filing Status (Weighted)

Filing Year	Single	Married Filing a Joint Return	Married Filing Separate Returns
1985	100.0%	100.0%	100.0%
1986	82.0	95.5	38.9
1987	72.8	90.2	8.5
1988	64.7	86.5	8.5
1989	56.4	83.8	8.3
1990	51.9	80.1	6.0
1991	44.3	75.2	5.6

The proportion of filing units maintaining stable composition over time varies widely by 1985 filing status. Only 44.3 percent of the single filing units versus 75.2 percent of married, joint filing units continued to file with the same status through 1991. Among married persons filing separately in 1985, only 38.9 percent filed the same way a year later. This proportion dropped to 8.5 percent by 1987, then declined gradually to 5.6 percent by 1991. Clearly, married filing separately was an exceedingly transitory status for all but a few of the persons who filed in that manner in 1985.

Some of the decline in units with stable composition over time is due to exits from the filing population. Units that leave the filing population -- and therefore the sample -- are not available for longitudinal analysis. It is appropriate, therefore, to include them in the count of stable units only for as long as they are present in the sample. Table 3 reports the percentages of filers (as opposed to filing units) who filed in subsequent filing years. There is much less differentiation across base year filing statuses than was evident in Table 2. Of those who filed single in 1985, 82.9 percent filed in 1991. Of those who filed joint returns with their spouses in 1985, 88.7 percent filed in 1991, while 73.2 percent of those who filed separately from their spouses in 1985 filed in 1991.

Table 3.--Percentage of 1985 Filers Still Filing in Subsequent Years, by Filing Year and 1985 Filing Status (Weighted)

Filing Year	Single	Married Filing a Joint Return	Married Filing Separate Returns
1985	100.0%	100.0%	100.0%
1986	93.4	97.8	92.7
1987	89.8	95.7	80.2
1988	89.1	94.1	86.0
1989	86.6	93.0	92.7
1990	86.7	91.4	80.5
1991	82.9	88.7	73.2

A comparison of Tables 1 and 3 reveals that a strategy of limiting longitudinal analysis to units with stable composition for the entire duration of the panel would discard over 40 percent of the 1985 filing units, whereas all but about 15 percent of the filers in the 1985 filing units filed in 1991 and, therefore, are represented in the SOCA database in that year. Seemingly, a longitudinal analysis strategy that could utilize more of the panel sample would better represent the experience of the filing population over time.

Concentration of Capital Assets Sales

Table 4 displays the results of our estimation of the fraction of capital transactions attributable to filing units that reported transactions in 1985. We utilize two measures of capital assets sales -- net capital gains and net capital losses -- and we describe concentration in terms of the percentage of returns and the percentage of dollars attributable to persons who reported sales in 1985. We separate net gains and losses because of the possibility, owing to the carryover provisions for capital losses, that taxpayers reporting net losses in a given year might be more likely to have reported transactions in an earlier year. The table presents three sets of alternative estimates, based, in turn, on the filing unit attributes of individual filers, filing units with stable composition through the filing year, and filing units with stable composition through 1991.

The method that makes the fullest use of the SOCA data generates the following findings, reported in the top panel of Table 4. Nearly two-thirds of the filing units that reported gains or losses in 1986 (and filed in 1985) also reported sales in 1985. These filing units accounted for 86 percent of the total dollar value of the gains and losses reported by the survivors of the 1985 filing population. In 1987 the proportion of returns with gains that also reported sales in 1985 drops to 57 percent while the corresponding proportion among returns with net losses falls to 55 percent. Both fractions rise and fall over the remaining years through 1991, but neither percentage is ever appreciably lower than the 1987 number. The dollar share attributable to filers reporting capital assets sales in 1985 shows a gradual decline for net gains, with the exception of a sharply nonmonotonic drop in 1990. For net losses, there is a sharp decline from an 86 percent share in 1986 to 65 percent in 1987. This level is maintained through 1988, after which the share of losses attributable to filers with 1985 sales stabilizes at around 60 percent. Finally, contrary to expectation, persons reporting capital losses in a given year do not show a greater likelihood of having reported sales in 1985.

Turning our attention to the estimates based on filing units with stable composition through the

Table 4.--Percent of Capital Transactions Attributable to Filing Units with 1985 Transactions, by Filing Year, Based on Alternative Uses of SOCA Panel Data

Filing Year	Net Capital Gain		Net Capital Loss	
	Returns	Dollars	Returns	Dollars
	Estimates Based on Filing Unit Attributes of Individual Filers			
1985	100.0%	100.0%	100.0%	100.0%
1986	65.0	86.1	65.9	86.3
1987	57.3	79.4	55.2	65.1
1988	62.1	79.9	65.3	66.8
1989	56.4	71.7	60.8	60.1
1990	59.1	57.5	56.6	59.8
1991	55.6	68.8	57.6	61.4
	Estimates Based on Filing Units with Stable Composition through Filing Year			
1985	100.0%	100.0%	100.0%	100.0%
1986	66.9	86.4	66.0	87.3
1987	59.6	79.4	55.7	64.8
1988	65.8	81.2	68.0	68.9
1989	59.3	73.7	65.9	66.0
1990	62.3	70.3	60.9	64.0
1991	59.4	63.9	62.2	65.6
	Estimates Based on Filing Units with Stable Composition through 1991			
1985	100.0%	100.0%	100.0%	100.0%
1986	67.3	86.6	72.9	91.3
1987	61.2	77.9	65.2	71.5
1988	66.9	86.7	71.0	73.8
1989	58.6	72.6	66.0	66.6
1990	61.2	68.8	60.7	63.8
1991	59.4	63.9	62.2	65.6

filing year or through 1991 (panels two and three, respectively), we find only small deviations from the findings reported in the top panel. Neither of the alternative methods based on stable filing units exhibits the sharp decline in 1990 in the percentage of capital gains attributable to filers with sales in 1985. More generally, for net capital losses, whether we count dollars or returns, the method based on units with stable composition through 1991 yields a higher estimate of the percentage of activity attributable to filers with sales in 1985. For net capital gains, we see the same pattern for returns, but for total dollars there are two years in which the estimated proportion attributable to filers with 1985 sales is actually lower than the estimate based on the filing unit attributes of individual filers.

Not surprisingly, the estimates based on filing units with stable composition through each filing year generally lie between those reported in panels one and three (in 1991, of course, the results reported in panels two and three are identical). The exceptions do not fall into any pattern.

Basing estimates of the concentration of capital gains activity solely on filing units with stable composition over time does indeed, yield evidence of greater concentration than we observe with a methodology that utilizes a broader sample of filers. The differences are not large enough to be important, however (although we suspect that an appropriate statistical test would find them to be significant overall).

Perhaps units that experience change in composition tend to have few capital transactions during those years. If they have no capital assets sales, then including or excluding them from an estimate of the characteristics of persons with gains is of no consequence. From the dollar estimates that underlie Table 4 we determined that filing units with stable composition accounted for 72 to 90 percent of the gains reported in a given year by the survivors of the 1985 filing population. The behavior of filers who experience changes in unit composition is not inconsequential, then, but clearly filers with stable composition account for a disproportionate share of gains activity. This attenuates potential differences in the estimates of the concentration of capital gains activity over time based on the alternative methodologies examined here.

■ References

Citro, Constance F.; Hernandez, Donald J.; and Herriot, Roger A. (1986). Longitudinal Household Concepts in SIPP: Preliminary Results, *Survey of Income and Program Participation Working Paper No. 8611*, Washington, DC: Bureau of the Census.

Kalton, Graham and Brick, J. Michael (1994). Weighting Schemes for Cross-Sectional Analyses of Household Panel Surveys, *Proceedings of the Section on Survey Research Methods Section, American Statistical Association.* ■