LINKING INDIVIDUALS IN A CAPITAL GAINS PANEL FOR TAX POLICY ANALYSIS

Susan C. Hostetter, Internal Revenue Service
Statistics of Income Division, IRS, PO Box 2608, Washington, DC 20013-2608

Key Words: Exact Match, Panel Linking

Historically, the Internal Revenue Service's Statistics of Income (SOI) Division produced data on Sales of Capital Assets (SOCA) every fourth year to gain insight on taxpayer gains and losses by capital asset type. Because these cross-sectional data have limitations for tax policy analysis, SOI established a panel of 12,980 taxpayers with and without reported capital gains to provide more complete data for our customers to observe the capital gains transactions for the same individuals over time. For details describing why IRS developed the panel, the history of capital gains taxes, previous capital gains studies, goals for the panel, and a description of the panel size in relation to the annual cross-sectional sample as well as the periodic capital gains study conducted for 1985 -- the base year of the SOCA panel -- see Hostetter, 1993.

The cost of a capital gains study or panel is considerable because, for each transaction, we capture the date and price for each purchase and sale. We also classify each transaction by type. It is not unusual for large (high income) returns to have a great many transactions in a single year. However, our customers -- Treasury's Office of Tax Analysis and Congress' Joint Committee on Taxation -- were so impressed with the benefits they found in using the longitudinal aspects of the SOI 1981-based pilot panel that they were convinced that the additional costs of capturing capital gains transactions and the -- not inconsequential -- costs of managing a panel would be well worth the value for policy analysis (Holik et al, 1989).

ACCURATE PANEL LINKS

"Fit for Use"

The SOCA panel records include, not only all the detail on each transaction, but also the full range of income and tax data associated with the SOI annual cross-sectional sample. So, the SOCA panel is actually used for many purposes. Although SOI has years of experience in capturing and editing taxpayer data, it has, as most statistical agencies have, little expertise in reviewing and editing the longitudinal aspects of linked records. If the longitudinal characteristics are a unique and important aspect of the SOCA panel, it follows that accurate panel links across years are crucial to determining the "fit for use" qualifications to meet our customers' needs.

This paper will focus on a review of the methods used to ensure that the longitudinal characteristics of the data were, indeed, fit for use. First, it describes how we identified individuals and outlines criteria used to identify potential linking error for manual review. Then it mentions some problems and issues calling for customer input, and discusses how some were resolved. Next results are presented, some of which have implications for the quality of IRS Master File social security numbers (SSN's) and the longitudinal quality of the SOCA Panel. And, of course, some recommendations for future IRS efforts are provided.

Defining Base Year Panel Units

Full background on the development of the SOI SOCA panel is detailed in Hostetter, 1993. Before the manual review of longitudinal linking could begin, we established initial panel units and assigned panel identifiers to each individual and each return. To each return record we associated selected income and tax information. Also available to the reviewer was computer-assigned coded information and correction fields to support our "clean-up" operation. We began as follows:

- Each base year 1985 SOCA tax return was initially assigned a panel number -- the Panel ID. For this identifier we used the basic SOI control number for tax returns in all individual samples.
- The Panel ID was then associated with both the primary and secondary SSN on the return. It was assigned to **individuals.**
- Each individual was given a Taxpayer Code -- "1" was assigned to all primary taxpayers and "2" was assigned to all secondary taxpayers in the base year. Although we will be linking individuals across years, it is important to remember that the data for these individuals come from tax returns and that filing patterns vary and fluctuate considerably.
- We then assigned a **return** Panel ID, so that if, in later years, our panel members marry panel members from other units, we can retain each individual Panel ID and still assign a panel identifier to the return, since all data are stored by return.

The 1985 base year was then established as our initial, pre-correction panel.

MANUAL REVIEW

Setting up the File

With the initial base-year panel in place, we then constructed the multi-year panel. The panel units were computer linked across all seven years for which we had data -- 1985-1991 -- using the Panel ID. We developed a

record for each tax return, for each year, consisting of the one or two original members from the 1985 base year and any taxpayers they married and included on their tax returns on subsequent years. Then, to examine the accuracy of the linkages, we conducted a major manual review. For each record we provided the following information for manual review:

- **Panel ID, SSN, Taxpayer Code,** which identifies whether the individual was the primary or secondary taxpayer in the base year.
- **Filing Status** (married or single), **City and State**
- **Income and Tax Information**, such as adjusted gross income, wages, interest, dividends, several capital gains items, pensions, itemized deductions, and tax liability. (These data were helpful in review for a consistent tax profile, which, at least in some cases, helps to verify the correct panel member.)
- **Test Failure Indicator,** which identifies problems for reviewers.
- **Return Name Control** (the first four letters of the taxpayer's last name, taken from the return.)
- **Name Control and Date of Birth** (information from the Social Security Administration that they associated with the SSN.)

Agreement on these critical matching keys was considered important to assure us that the appropriate individuals/returns were linked. Where disagreements occurred, we conducted a manual review. Generally, we performed manual review on panel units where any return for any year had a name control match failure, was a potential duplicate (two returns covering the same tax period with the same SSN), or returns that had two different panel ID's representing two panel units on the same return. When a potential error was identified, all returns for all years for the panel unit were reviewed together.

The name control match (**comparing the name control on the tax return to the Social Security name control**) was our most reliable tool for discovering error, and it was extremely accurate for reviewing primary SSN's. However, it was less reliable for secondary SSN's. On joint returns, over 90 percent of primary SSN's are men, and women's name controls may not match that of their husband even when the SSN is correct. Mostly, this is because women either don't change their name when they marry or fail to notify Social Security of the change. Fortunately, we did have all the name controls used by an individual available during review. By using other information we could almost always determine incorrect SSN's.

Although about half of all returns received by IRS are filed jointly, about 80 percent of the returns in SOI samples are from joint filers. This was very helpful in correcting and retaining panel members SSN's. If one of the SSN's on the joint return matched, but the other did not, we could retain the unit in the panel because of the individual with the correct SSN, so long as the couple did not split (i.e., continued to file jointly).

Who's In and Who's Out -

Initially we started with two types of individuals, and these, as well as groups we defined later, were identified by their taxpayer code. We started with:

- **Panel Members** - the individuals we had coded 1 or 2 from the 1985 return,

and

- **Visitors** - taxpayers who, in later years, married our panel members and appeared on their returns. They had no Panel ID and were given a taxpayer code of "3."

During one of our meetings with Treasury staff, we discussed cases involving people entering the Panel because of problems with their SSN's or those who enter with their correct SSN because a Panel member mistakenly used it in the base year. We developed new categories to cover these problem cases and dealt with them as follows:

- **Volunteers** - Treasury said they would like to retain returns for people who were drawn into the panel, beginning in 1986, with **their** correct SSN because a panel member had used it erroneously in 1985. Because these SSN's were on the selection file, they were picked up each year, so for many we have data from 1986 through 1991.

  These returns will have a weight of 0, an altered Panel ID, and will not be considered part of the panel. Their taxpayer codes 4, 5, and 6, correspond to the 1, 2, and 3 for panel members and visitors.
- **Intruders** - Of course there were invalid returns caused by taxpayers incorrectly using our panel member's SSN's. These returns had their Panel ID removed and were deleted from the panel file.

CORRECTING THE FILE

To correct the file we developed a review plan that included interaction with Treasury staff. We began with specifications for developing initial panel links and tests to identify panel units for manual review. During the process of meeting with Treasury and changing our methods, these specifications were updated to reflect the most current decisions. We prepared written review procedures, and amended them with notes from meetings with Treasury. Our manual review process was confined to a few SOI staff, so the actual production was closely monitored and controlled. This enabled us to adapt to new situations as they occurred.

Treasury staff was explicit in saying they did not want any taxpayer-reported data changed -- they only wanted the

individual identifiers corrected to provide accurate panel links. They also noted specifically that, when SSN's are corrected, the original value should also be retained. Further, Treasury staff said they did not want duplicate returns for the same tax period deleted if they were filed by panel members -- they only wanted them identified.

SOI file corrections were typed into a Word Perfect document, using a simple, single-line format. The first two fields -- the return ID and the Tax Year -- were preprinted on the review document and identified the precise return for correction. The only fields we corrected for the primary taxpayer and/or the secondary taxpayer were the following:

- the **Panel ID**;
- the **SSN**; and
- the **taxpayer code**, which identified characteristics for panel members, visitors, and intruders.

The last field was the action code, which indicated:

- a **modification**;
- a **deletion**;
- that a primary or secondary **panel member** used a **wrong SSN** for all years, and we didn't know the correct SSN; or
- that a wrong panel SSN caused us to **lose all or part of the panel unit** -- primary, secondary, or both.

Table I.--Characteristics of Panel File

Initial Number Panel Units	**12,980**
Units Lost Due to Incorrect/Missing SSN's	**94**
Number Volunteer Units	**112**
Number Intruders, 1986-1991	**736**
Number Unit Mergers	**1**
Percent Panel Returns with Visitors -1985	**0%**
-1991	**10%**
Number Returns, 1985-1991	**93,363**
Number Records with Manual Corrections	**3,274**

THE SOCA PANEL FILE

Counts of Taxpayer Groups

Table I summarizes the SOCA Panel profile after the manual review. In 1985 we started with 12,980 panel returns or panel units. We lost 94 of those, including some with missing SSN's, because we had an incorrect SSN and one or both filers disappeared before the correct SSN was entered to the annual selection file.

We had 112 of the volunteer units, where our panel members borrowed another taxpayer's SSN in 1985. The correct SSN for the volunteer was on the selection file, so we continued to select them. Although these units are not part of the panel, they can be used in many instances where users are making unweighted panel comparisons across years.

There were 736 intruders, beginning in 1986. They were there because they inadvertently used a panel members SSN, and usually only for one or two years. All were deleted.

We had one unit merger or marriage: two panel members, from two different panel units in 1985, married each other. This is an issue for weight adjustment.

Another issue for reweighting is that in 1985, by definition, all individuals were panel members, but by 1991, ten percent of the returns had a visitor. These taxpayers married our panel members, brought their own income and tax characteristics, and must be included in estimates because they are on the panel returns. Over time their presence creates difficulties for weighting. (See Czajka, 1994, for a discussion of the effects of the data anomalies on weighting and estimation.)

Finally, there were 93,363 returns in the seven year panel, of which 3,274 were corrected.

Identification of Incorrect SSN's, 1985

The characteristics that most affected our definition of the base year panel were the missing or incorrect SSN's. Table II summarizes those we identified in the manual review. Using the name control match as a basic indicator, we show 59 primary SSN's and 856 secondaries whose name controls didn't match in the base year.

There were 102 returns with a joint filing status that had no secondary SSN -- that is, they filed as married but didn't report an SSN for a spouse. Generally, we view the taxpayer's reporting of filing status as more accurate than their reporting of secondary SSN's. We were able to insert the secondary SSN in 58 cases, based on data reported in later years.

We also had the opposite problem -- 17 returns with a filing status claiming that they were single, but with a secondary SSN reported. In all but four of these cases the 1985 primary taxpayers were determined to be single, and the secondary SSN's were deleted. Our manual review showed evidence that the four we retained were married in 1985, and simply omitted reporting the secondary SSN.

For the primaries, we corrected ten, and these were almost certainly on joint returns. We lost 19 panel units, and these were mostly single filers, who we had no way to identify, plus the two units where both the primary and

Table II.-- SOCA Panel: Summary of Incorrect SSN's, for 1985 Base Year

	Primary	Secondary
Number Name Control Nonmatches	59	856
Number Missing Secondary SSN's		102
Action Taken		
Corrected SSN's	10	140
Missing SSN's Inserted		60
Units Lost, Incorrect SSN	19	44
Units Lost, Both SSN's Incorrect	2	2
Units Lost, Missing SSN's		29
Incorrect SSN, Unable to Correct	2	46
Total	**33**	**321**
Adjusted Percent SSN's Wrong	**0.25%**	**3.0%**

secondary SSN's on the return were incorrect. The remaining SSN's were not incorrect -- they may have had a transcription error in the name control, either at IRS or at Social Security. However, neither error is common.

Finally, two units had an incorrect primary SSN that was used consistently for seven years and provided no information for correction. However, because these were joint returns that stayed intact for the seven years and where the secondary SSN was correct, there were no loss of data. Presumably, SOI staff will be able to access additional IRS accounts to obtain correct SSN's for such cases, and will then include these SSN's on the Panel selection file. We did not consider such units a loss during this review. So, in all, 33 primary SSN's were considered incorrect. This extremely low error rate -- one fourth of one percent --is no doubt due to the fact that IRS processes about 117 million returns a year and is continually concerned about the error rate of primary SSN's.

We also corrected 200 secondary SSN's and lost 75 units (two of these were included in the primary unit loss). Of the remaining 73 units, twenty-nine units were lost because of missing secondary SSN's and 44 because of incorrect ones. We had 46 units with continuing incorrect SSN's for seven years.

One interesting anomaly occurred with 13 1985 returns with missing secondary SSN's. Our manual review led us to change the filing status on six of these to single, and seven were left with no secondary SSN and a married filing status. Of these seven, three were joint returns for all years, so no data were missing. Four were filed from other countries where spouses may not have U.S. SSN's, or were filing with a separated status, but claiming the spouse as a dependent (rarely used). The last of these filed a 1986 return indicating the spouse died, and for the remaining returns filed single.

With 321 secondary SSN's confirmed incorrect, leaving 642 suspicious SSN's that were considered correct during our manual review, that yielded an overall error rate of three percent.

Error Rates By Income Class

Table III shows the percent of incorrect SSN's by adjusted gross income class, both weighted and unweighted, for the 1985 SOCA Panel. The interesting feature for the primary SSN's is that any significant error rate is in the low, and mostly positive, income classes.

The secondary SSN's have higher error rates, as we would expect. IRS does not automatically perform editing and correction functions for secondary SSN's to the same extent that it does for primaries. Our incorrect rates include missing SSN's. Overall, for all returns filed with IRS, about half the returns don't have secondary SSN's, and most returns that do are in the higher income classes. So, there are not that many low income returns with secondaries, but this is where the higher error rates are -- in the low, positive income classes.

Characteristics by Income Class

Table IV shows some of the panel characteristics shown in Tables I and II by income class. Unlike the previous tables, Table IV covers characteristics for all seven years. In the first row of the table -- "Percent with No Error" -- you can see a small but steady improvement in the quality of SSN's as income rises.

The percent of missing or incorrect secondary SSN's is more consistent across income classes in Table IV because the data cover all seven years, and because the income classes in Table IV do not separately identify the negative

Table III.--SOCA Panel: Percent Incorrect SSN's, by Income Class, in the Base Year, 1985

Income Class (Dollars)	Percent Incorrect Primary SSN's		Percent Incorrect Secondary SSN's	
	Unweighted	Weighted	Unweighted	Weighted
Less than -$49,999	.2	1.6	5.8	3.0
-49,999 to 0	1.2	.0	7.2	5.3
1 to 4,999	**1.9**	**3.1**	**13.0**	**20.5**
5,000 to 9,999	.7	1.0	5.8	7.0
10,000 to 24,999	.4	.4	3.4	3.3
25,000 to 49,999	.1	.1	1.8	1.1
50,000 to 99,999	.1	.0	1.8	1.2
100,000 to 199,999	.2	.1	2.9	2.7
200,000 to 499,999	.1	.0	3.4	2.5
500,000 to 999,999	.2	.2	2.2	1.2
1,000,000 or more	.1	.1	2.9	3.1
Total	**.2**	**.8**	**3.1**	**3.4**

and low positive income classes where so much of the variability occurred. The percents of volunteers, where panel members reported incorrect SSN's with random selection, are generally representative of the population of returns. Hence, there are none shown above $250,000 because there are so few returns in that stratum in the population. On the other hand, the percent of intruders, where nonpanel taxpayers used SSN's belonging to SOCA Panel members, reflects the tendency for low income taxpayers to make more errors.

Table IV.--SOCA Panel: Percent with Selected Characteristics by Income Class

Panel Return Characteristic	Percent Returns by Income Class (in Thousands of Dollars)				
	Under $30	$30 to 60	$60 to 100	$100 to 250	$250 or more
No Error	95.2	95.4	97.0	97.7	98.3
Incorrect Primary SSN	.1	.0	.0	.0	.0
Incorrect Secondary SSN	.8	.5	.6	.9	.9
Missing Secondary SSN	1.5	.5	.4	.7	.6
Volunteers	1.1	2.0	1.2	.3	.0
Intruders	1.6	1.7	.8	.3	.1

RECOMMENDATIONS FOR FUTURE SOI PANELS

In conclusion, the manual review -- thought costly -- helped us improve the quality of the panel linkages and re-assured us that the resulting panel is fit for use. Our experience in reviewing and linking both the SOCA Panel and the Individual Family Panel has given us insights to some of the problems we, in SOI, face as our efforts to develop longitudinal time series data continue. The following are recommendations that should improve the value and timeliness of future panels:

- Make sure that the initial panel covers all demographic characteristics and all tax data requested by Treasury and Congress for their policy analysis.
- Design a panel that is broad enough to more adequately (than the SOCA Panel) represent the population over time. At the same time, design a method for panel replacement, to maintain the representative nature of the data, while continuing to provide the year-to-year analysis of change. (For further discussion of collaborative sample design efforts, see Hostetter and O'Conor, 1991.)
- Begin early, after the first year of processing, to review SSN's that fail name control matches by comparing them to other IRS entity information. Review for possible duplicates, which could be intruders. New improvements in IRS-controlled data accessibility should provide this as a new opportunity.
- Stay in touch with the **customer.**
- Share information about incorrect SSN's and filing patterns with other areas of IRS. Offer to assist them in developing methods to improve, even more, the quality of SSN's, particularly for secondary taxpayers and dependents (who are included in other SOI panels). (Hostetter, 1992)
- Review Panel SSN's that disappear after the first year against the Social Security Date of Death File to see if they died. Also, review such losses to other IRS tax information files to see if the taxpayer dropped below the tax filing requirement.
- Do these types of review annually, after the first year, especially for panel units with change or potential error. Monitor all unclear panel units annually.

If these kinds of suggestions are implemented, we can expect even better results from future evaluations of SOI panel data.

ACKNOWLEDGMENTS

The author is grateful to Ruth Glover, Jefferi Bass, and Darlene Reynolds for their valuable assistance in data preparation for the SOCA Panel. She is also appreciative of the editorial assistance provided by Wendy Alvey.

REFERENCES

Czajka, J. (1994) Income Stratification in Panel Surveys: Issues in Design and Estimation, **American Statistical Association 1994 Proceedings of the Section on Survey Research Methods.**

Holik, D.; Hostetter, S.; and Labate, J. (1989) The 1985 Sales of Capital Assets Study, **American Statistical Association 1989 Proceedings of the Section on Survey Research Methods.**

Hostetter, S. (1993) Membership in a Linked Panel of Individual Tax Returns: Review and Results, **American Statistical Association 1993 Proceedings of the Section on Survey Research Methods.**

Hostetter, S. (1992) Exploring Nonsampling Error in a Longitudinal Survey of Individual Taxpayers, **Symposium 92: Design and Analysis of Longitudinal Surveys,** Ottawa, Ontario, Canada.

Hostetter, S. and O'Conor, K. (1991) Satisfying the Need of Income Policy Modelers While Preserving the Reliability of Descriptive Statistics, **Statistics of Income: Turning Administrative Systems Into Information Systems -- 1991-1992,** Internal Revenue Service.