# MANAGING MULTIPLE USES OF PANELS

**Susan Hostetter, Internal Revenue Service and Joint Committee on Taxation**

KEY WORDS: Longitudinal, Income

## INTRODUCTION

The Internal Revenue Service (IRS) has developed an important new Panel of 90,000 tax families. In 1987 Treasury's Office of Tax Analysis asked Statistics of Income (SOI) at IRS to begin a major redesign of its Individual tax return sample, to improve it for more accurate modelling of the effects of tax policy recommendations (Hostetter and O'Conor, 1992; Czajka and Schirm, 1992). SOI was asked to do three things:

- **To design and implement a Tax Family Unit**, so that Treasury, and Congress's Joint Committee on Taxation (Joint Tax) could model the effect of tax law changes on family economic units (Nelson, 1986);
- **To redesign the stratification of the sample selection**, to strengthen the sample of income components of importance to tax policy, and to obtain better coverage of certain demographic groups (Hostetter et al., 1990); and
- **To design and implement a panel of individual tax returns**, to measure the effect of tax policy on individual taxpayer behavior over time, as opposed to measuring change in aggregates. This has particular importance for tax policy involving capital gains.

Previous papers at these meetings have described our efforts with regard to the first two objectives. This paper focuses on the last of these -- the design and implementation of a panel of tax returns. The paper will begin with some background information on important milestone panels. Then, it will describe the SOI Panel and its components. The next section will focus on panel management issues and the paper concludes with a discussion of our plans for the future.

## BACKGROUND

### Historical Panels

There have been many panels designed and selected over the years -- some of the earliest research and analysis was presented in the 1970's, and the 1980's brought a rash of panel work. In the 1990's we want to learn from this experience to improve our current panel designs. Following are brief descriptions of three early panels which served as the basis for some of our current panel work.

The **Continuous Work History Sample (CWHS)**, initiated in the late 1930's by the Social Security Administration, was the first major longitudinal database developed. It represented samples of the ending digits of social security numbers (SSN's). Using the Longitudinal Employee-Employer Data sample of CWHS records, covering years 1957 to 1969, Nancy and Richard Ruggles firmly established the extent of difference in estimates of change over time by comparing longitudinal to cross-sectional data for the same years (Ruggles and Ruggles, 1974). Additionally, their study showed that the longitudinal data were valuable for isolating the behavioral change of groups by age, sex, and race. For example, the cohort born in 1911 and entering the workforce in 1930, at the beginning of the Depression, earned less income throughout their earning years covered by the study compared to the cohorts two years earlier or later. These kinds of valuable insights firmly established the need for more and better panel data.

The **Michigan Panel Study of Income Dynamics**, begun in 1968, creates a linked parent-child file with panel features for both generations and is useful for studying intergenerational earnings and for income analysis (Duncan et al., 1984). The intergenerational aspect of it was useful to us at IRS when we began following tax families.

The **Survey of Income and Program Participation (SIPP),** evolved slowly, with several pilot studies during the 1970's, and by 1983 it was initiated by the Census Bureau as a regular national survey (Kasprzyk and McMillen, 1987; Kasprzyk and Frankel, 1985). It is notable for it's coverage of transfer payments; as a replacement panel (with households remaining in the panel for 2 1/2 years); and for its public use file, which was later improved to be more user friendly (David 1989; David et al. 1988).

In the 1970's the SIPP panel was identified as a candidate for integration of survey and administrative records, with administrative sources used in interim years between a cyclical survey (Scheuren, 1975). Realizing the value of combined survey and administrative data -- in terms of wealth of information, cost savings and respondent burden reduction -- this integrated aspect of the SIPP has been carried over to the SOI panel effort (Scheuren, 1985; Kilss and Scheuren, 1980).

Drawing on what we had learned from earlier panels, Treasury and SOI began the **1981 Sales of Capital Assets Study.** This eventually became a transitional milestone panel. The panel was 13,000 returns and included IRS Master File data covering Tax Years 1979 to 1988. Treasury staff found that, in spite of its small size and lesser data quality, for a number of policy issues, where change at the individual level was important, the Capital Gains Panel was more useful for policy analysis than the much larger SOI cross-sectional data. It was that success which led to Treasury's insistence on creating a new broader SOI panel.

### Limitations of Panels

Clearly, the panels discussed here offer a considerable source of knowledge, particularly concerning life cycles of income and wealth or individual behavior caused by tax policy changes or economic changes. Nevertheless, we did not enter this venture with rose-colored glasses. We realized there is a cost and some limitations to these gains. For example:

- **Tax law changes** are good for reviewing the before and after picture, but major changes cause a break in continuity of data.
- Large quantities of **resources** are necessary to initiate and maintain a panel, mainly to keep track of all the people. Keeping knowledgeable and trained staff devoted to a project for many years is difficult.
- The **weighting issues** for such a series of panels are very complex. Such work will frequently require contractor experts. Generally, both design-based and model-based weights will be needed (Schirm and Czajka, 1992).
- There is a lot of **expense** to creating and continuing a large panel, quite necessary if it's to run a long time, or if it's to provide information on diverse characteristics.

All of these concerns were taken into consideration before and during the SOI panel design phase.

## The Individual Program Panel

### Description of Basic Panel

Bearing these limitations in mind, let me introduce SOI's current Panel work. **The SOI Panel** is the basic Individual Program Panel of 90,000 tax family units. It was initiated for Tax Year 1987 by selecting 90,000 nondependent returns (we call these parent returns) and returns for any dependents claimed on the parent returns. A file with all SSN's reported as primary, secondary, or dependent on the 90,000 returns was created. This file defines the actual "card-carrying" members of the panel. It is a little confusing, but important to remember that this is a **panel of individuals reported on tax returns.** With expansion due to changes in family structure, SOI has about 135,000 returns in the Panel for 1990, but they still represent and will be weighted as 90,000 Panel units.

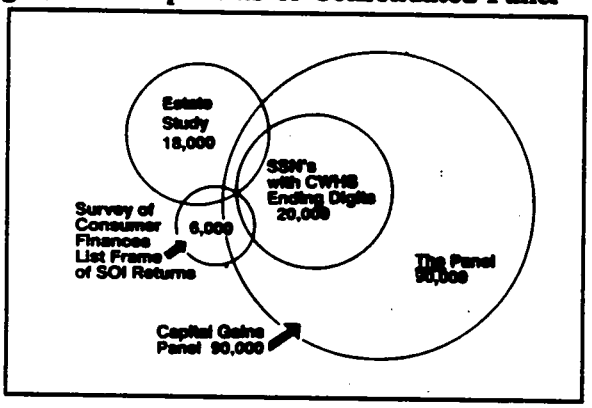
### Description of the Consolidated Panel

The basic Panel, just described, however, is only the starting point. Our full plan is to develop a Consolidated Panel, which overlaps with the basic Panel, in order to benefit multiple uses for policy analysis. In other words, by overlapping several

sets of longitudinal data on individual taxpayers, we can provide richer data for each of the specific studies, and at a lower cost. (See Figure 1.) This section describes four other major panels that will be included in the SOI Panel design, management, and processing. For simplicity, the other panels will be referred to by their specific survey name or abbreviation:

- A CWHS Sample of 20,000 SSN's is included for selection in the Panel. IRS has included the same sample of 20,000 CWHS SSN's for selection in its cross-sectional sample since 1979. As part of the Panel, the returns selected in this sample will also have their dependents selected and linked, to make tax families. This group provides a direct overlap with the cross-sectional sample, for use in comparative time series studies.

**Figure 1: Components of Consolidated Panel**




- The Survey of Consumer Finance (SCF) is a triennial household survey conducted by the Federal Reserve Board, in cooperation with SOI. It is used by the Federal Reserve, Congress, Treasury, and researchers to study a broad range of financial characteristics of households (Kennickell and Woodburn, 1992). Selected individual tax data are used to develop and administer a list sample for the Survey. The list sample supplements a traditional area sample to provide better estimates of skewed household financial characteristics, such as assets and wealth. For Tax Year 1992, about 6,000 returns, originally used in the development of the 1989 SCF, will be followed in the SOI Panel. Future surveys will use Panel tax families as a surrogate for households. Severe limitations govern the use of these datasets, restricting the linkage of tax data and survey data for survey respondents.

- Based on knowledge gained in a pilot study of 18,000 decedents and beneficiaries listed on 1989 estate returns, SOI will initiate a new 1993 Estate Collation Study, with partial overlap of beneficiaries and their tax families, who will be followed for an indefinite period. The study will be particularly valuable for research and analysis of life cycle relationships of income and wealth (Johnson and Woodburn, 1992).

- Beginning with 1993, SOI will incorporate Capital Gains data for the entire Panel of 90,000 units. The Panel units and, within those, the family units will remain; the difference is in the additional data -- the asset codes and transactions information for all capital gains or losses reported on the Panel returns. Because the capital gains work is more centralized and time consuming, there will be some processing changes necessary to ensure that the basic Program data will be delivered on time, while continuing to develop the capital gains data.

Obviously, constructing a Consolidated Panel along the lines just outlined is no easy task, especially when starting with administrative data designed for other purposes. How do we pull this Panel of over 300,000 individuals together? How do we identify each of the panels that each individual belongs to -- the Basic Panel and the "Guest Panels" we just described? (Remember the overlap of those membership circles.) How do we assign each individual to the correct Panel unit in each Panel to which he belongs? And, how do we assign each individual to the correct tax family unit? It's very difficult! The next section of the paper will focus on management issues -- first on the human behavior that affects the filing and

reporting of taxpayers and, therefore, our ability to construct accurate tax family units through selection and record linkage; then, on the development of the Masterfile and Control files that will be necessary to manage and track the membership of individuals in families, panel units, and panel surveys during the life of the Panel.

## PANEL MANAGEMENT

### "Family Matching Sins"

Creating tax families, in itself, is not always straight forward. Using the SSN, name, and dependent status as taxpayers report them, however, we can usually identify "family members" from tax returns. Creating and maintaining tax families over a number of years, though, is complicated by the fact that taxpayers commit what we have dubbed "family matching sins." Seven of the most "deadly" bear mention:

- **Marrying** – Change in status requires identification and review, since spouses must be added to the Panel tax family. The worst case is marrying another Panel member, which, additionally, requires weighting adjustments.
- **Divorcing** – This creates two families instead of one in the same Panel unit. Since both individuals were in the original Panel, both must remain in the current Panel. Their new filing status - single - creates two separate families.
- **Remarrying** – This brings in a nonmember (a "**visitor**") to the Panel, and it might bring in dependent visitors that arrived with the new spouse. Since at least one individual from each of the newly created families is a card-carrying Panel member, both must remain in the sample.
- **Claiming Dependents** – Claiming either children or parents, who are not Panel members, creates more visitors. Only dependents claimed on the originally selected returns are card-carrying members, and all others claimed in subsequent years are visitors.
- **Divorcing a Visitor** – When this occurs the visitor must be deleted from active selection, and subsequently any dependent visitors that are not claimed by our Panel member must also be deleted. Even though they are not Panel members, visitors are "selected" only as long as they are claimed on a member's return.
- **Sharing Your SSN** – If taxpayers share with many friends or just with their "cousin," we have to review all returns with duplicate Panel SSN's to make sure we keep the correct return and eliminate all wrong returns.
- **Reporting the Wrong SSN** – This is difficult, whether it is our Panel member or someone using a member's SSN. It sometimes requires extensive review.

All of these taxpayer behaviors complicate the ability to match and keep track of tax families over time. Yet, in most cases, they are legitimate taxpayer responses (not response error). Add to that the potential for IRS processing error – which is actually quite low, particularly for the primary SSN on a return (Steffick, 1992) – and we certainly have our work cut out for us. Therefore, we began, early on, to think a about Panel management.

### Designing a Master File and Control Files

Managing change and retaining categories for this Consolidated Panel has been a real challenge. One of the things we've designed and developed is a **Master File** – a master list of all SSN's selected for any of the panels – correct or incorrect, with all identifying characteristics. Its primary uses are to:

- Develop the annual **SSN file** used to select all Panel returns;
- Develop and control the Panel unit cleanup process;
- Keep a **historical record** of all Panel classification activity; and
- Provide the bulk of information necessary to **develop weights** for the Panel file.

We have also developed two **Control Files** -- one representing all Panel member **individuals** and one representing all **returns** filed with Panel members on them. Both files will have a 1987 base year module and will have separate modules for each processing year.

One important use of the Control File will be to handle returns that are filed one or more years late -- what we call prior-year returns since they cover a prior Tax Year. For instance, if a couple had an overseas assignment, came home at the end of the three-year assignment, and filed tax returns for 1990, 1991, and 1992 in April of 1993, while their two dependent college children dutifully filed returns on time for each year, the modeler using the three data files would be able to link each of these returns within the correct Tax Year using information on the Control File.

## FUTURE PLANS

Five years after initiating the Panel we have edited over 330,000 returns, including manual review of over 150,000. This time span is key to our correction and improvement timetable, because you can't review the relationships and activities of panel members intelligently until you have at least three years of data. Therefore, you can't even link the returns into families until this cleanup work is completed. The review covered **only** the Panel characteristics, to develop "clean" Panel Tax Families for the first three years. Within this year, SOI will redeliver data files for 1988 and 1989 to both Treasury and Joint Tax, with cleaned panel data and with tax families linked. Using information obtained in this first round of review, SOI will link and edit returns for 1990 and 1991.

For 1992 we will use new cleanup methods, based on predicting individual tax family behavior from the five-year panel database. This cleanup will be at the tail-end of the service center production work, which covers a detailed computer-assisted manual review of potential error conditions. Continuing to upgrade this cleanup operation, SOI will develop and implement front-end cleanup for 1993. Again, this operation will be based on continually updated intelligence for each Panel unit.

The first step to improving the Panel is to say we've only begun. As long as the Panel continues, it will be in a state of development, change, and improvement. Clearly, the Panel development will be an iterative process. It will also be a data driven process, with the designer and user forming a team. Continual conversation and sharing with our primary customers will be a requirement for success in this endeavor, as their feedback is necessary to help us identify and improve the process. The process that will evolve will look something like this:

- We **benchmark** -- to see what other people have done and what they're doing now to manage and improve their panels.

- We **connect our analysis** by using the results. We use analytical results from the Panel data **and** from the additional Panel management data from the master and control files to continually feed improvements into the design methods and process. Particularly, our analysis should include a study of the causal affects of nonresponse and their impact on weighting issues (Scheuren, 1989).

- And, then, we try to **gain wisdom.** -- The Panel is part of a system (a process) to gather information, analyze it, question it, and to get better information to continually change the system. In other words, we need to continually remind ourselves that we want to run this Panel in a research lab.

It is our hope that by applying these Total Quality Management principles (Juran, 1988) to the Panel's management, the database will continue to grow and meet the changing needs of our principal users.

## ACKNOWLEDGMENTS

The author thanks Wendy Alvey, Beth Kilss, and Clémentine Brittain for their editorial assistance and visual aids for the preparation and presentation of this paper.

## REFERENCES


Czajka, John L. and Schirm, Allen L. (1992), "Enhancing the Representativeness of a Longitudinal Sample of Individual Tax Returns: Weighting and Sample Supplementation," Proceedings of the 1992 Annual Research Conference. Bureau of the Census.

David, Martin H. (1989), "Managing Panel Data for Scientific Analysis: The Role Relational Database Management Systems," The American Statistical Association International Symposium on Panel Surveys. John Wiley & Sons.

David, Martin H., Robbin, Alice, and Flory, T.S. (1988), "Access to Data: Handling the 1984 SIPP," 1988 Proceedings of the Statistical Computing Section. American Statistical Association.

Duncan, Greg, Juster, F.Thomas, and Morgan, J.N.(1984), " The Role of Panel Studies in a World of Scarce Resources," The Collection and Analysis of Economic and Consumer Behavior Data (S. Sudman and M.A. Spaeth, eds.), Bureau of Economic and Business Research, Champaign, IL.

Hostetter, Susan and O'Conor, Karen (1991), "Satisfying the Need of Income Policy Modelers While Preserving the Reliability of Descriptive Statistics," Statistics of Income and Related Administrative Record Research: 1992. Internal Revenue Service.

Hostetter, Susan et al. (1990), "Choosing the Appropriate Income Classifier for Economic Tax Modeling," 1990 Proceedings of the Section on Survey Research Methods. American Statistical Association.

Johnson, Barry and Woodburn, Louise (1992), "The Underlying Methodology of the Estate Multiplier Technique: Recent Improvements and Estimates for 1989," Paper presented at the 1992 Joint Statistical Meetings, Boston, MA.

Juran, Joseph M. (1988), Juran on Planning Quality. New York, NY, The Free Press, 1988.

Kasprzyk, Daniel and Frankel, Delma (eds.) (1985), Survey of Income and Program Participation and Related Longitudinal Surveys:1984. Bureau of the Census.

Kasprzyk, Daniel and McMillen, D.B. (1987), "SIPP: Characteristics of the 1984 Panel," 1987 Proceedings of the Social Statistics Section. American Statistical Association.

Kennickell, Arthur B. and Woodburn, Louise (1992), "Methodological Issues in the Estimation of Household Net Worth: Results from the 1989 Survey of Consumer Finances," 1992 Proceedings of the Survey Research Section. American Statistical Association.

Kilss, Beth and Scheuren, Frederick (1980), "Goals and Plans for a Linked Administrative Statistical Sample," 1980 Proceedings of the Section on Survey Research Methods. American Statistical Association.

Nelson, Susan C. (1986), "Family Economic Income and Other Income Concepts Used in Analyzing Tax Reform," Compendium of Tax Research, 1986. Department of Treasury, Office of Tax Analysis.

Ruggles, Nancy D. and Ruggles, Richard (1974), "The Anatomy of Earnings Behavior," The Distribution of Economic Wellbeing. (F. Thomas Juster, ed.), Cambridge, MA, Ballinger.

Scheuren, Frederick (1989), "Nonresponse Adjustments: Discussion," Panel Surveys. John Wiley & Sons.

Scheuren, Frederick (1985), "Methodological Issues in Linkage of Multiple Data Bases," Record Linkage Techniques - 1985. Internal Revenue Service.

Scheuren, Frederick (1975), "ORS Management of the HEW Income Security Survey -- Some Administrative Issues," Working Paper, Office of Research and Statistics, Social Security Administration.

Schirm, Allen L. and Czajka, John L. (1992), "Model-Based Alternatives to Design-Based Weighting in a Panel of Individual Tax Returns," 1992 Proceedings of the Section on Survey Research Methods. American Statistical Association.

Steffick, Diane (1992), "Analyzing Longitudinal Data Linkages in a Panel of Individual Tax Returns," 1992 Proceedings of the Section on Social Statistics. American Statistical Association.