ANALYZING LONGITUDINAL DATA LINKAGES IN A PANEL OF INDIVIDUAL TAX RETURNS

Diane Steffick, Internal Revenue Service

KEY WORDS: Tax Statistics, Administrative Record Research, Exact Matching

In recent years, the Statistics of Income Division (SOI) of the Internal Revenue Service (IRS) has redesigned its sample of tax returns filed by individuals to provide more useful data for tax policy analysis. In addition to an enhanced cross-sectional sample [1], two components of this redesign are particularly important --the implementation of a longitudinal panel of returns embedded within the cross-section sample and the construction of tax family units. Longitudinal data are important to enhance tax modeling because many tax policy issues relate to the distributional consequences of a proposal, i.e., will a specific policy increase the welfare of low income, middle class, or wealthy taxpayers [2]? To see the effects of a change, it is necessary to have data on the same units. Tax family units are important because the family, rather than the individual, is the relevant economic unit for tax policy analysis, since family members generally pool their income for a common level of economic well-being. The use of the family unit also makes the SOI data more comparable to other data sets based on the household unit.

This paper focuses on the recent review of the linking of panel and tax family returns. Section 1 provides background on the panel and tax family concept, Section 2 reports the preliminary results of the review project, Section 3 details the methodology used to review these records, Section 4 describes examples of tax family behavior, and Section 5 presents some conclusions from the review and future plans.

1. BACKGROUND

The redesign of the individual sample grew out of the implementation of Total Quality Management principles in the Statistics of Income Division. The quality planning process included consultation with the major customers and suppliers of IRS statistical data from the initial stages through to the current review. The changes that resulted, including the panel and tax families, stemmed directly from their stated needs.

The individual panel was designated from the 1987 Individual SOI cross-sectional sample [3]. The 1987 SOI sample generally contained returns filed for the calendar year 1987 tax period, but also included prior year return filers, taxpayers who were filing their calendar year 1986 or earlier returns during the filing season for 1987 returns. Non-dependent returns (those belonging to taxpayers who indicated they were not being claimed as a dependent on another return, including prior year return filers) were designated for the panel by subsampling the cross-section sample strata, selecting a return based on a transformation of the Social Security Number (SSN) of the primary taxpayer (Harte, 1986). Returns were also designated for the panel if they had been selected for the SOI sample because they were part of the Social Security Administration's Continuous Work History Study [4]. In addition to the primary taxpayer on the tax return, all family members of that taxpayer -- spouses and dependents -- claimed on that return were selected as panel members. The 1987 Tax Year was the first year taxpayers were required to report the SSNs of their dependents. For various legitimate reasons, some taxpayers were not able to report the SSNs of their dependents in 1987. Therefore, all dependents for which exemptions were claimed in 1987, even if their SSNs were not listed in 1987, are considered members of the panel.

The tax family unit is defined as the non-dependent taxpayers on a tax return (the primary taxpayer and spouse) and the dependents claimed by those non-dependent taxpayers on their return, for a specific year. Any individual tax return filed by a member of this set of taxpayers is linked to the returns of the other members of the set. Therefore, this concept of the family is determined administratively by tax law and taxpayer reporting behavior; there is no outside contact with the taxpayer to determine the members of the household or family. Although families are constructed for all returns in the SOI sample, both panel and cross-sectio:ı, this paper will focus on tax families for the individual panel only.

1.1 The Linking Process

Each year, starting with the 1988 SOI sample, panel returns are linked based on a simple SSN match. If the SSN on a return is the same as a panel member's SSN, the return is linked to the panel unit as a panel return.

Similarly, the family links are based solely on the SSN. If an SSN on a return matches the SSN of any member of that tax family, the return is linked as a family return.

However, we suspected that there could be a substantial amount of error in the reporting of SSNs by taxpayers; especially for dependents, since 1987 was the first year their SSNs were required. When discussing the panel, specific SSNs are referred to by their placement on the tax return. The first SSN listed on a return is referred to as the primary taxpayer's SSN or primary SSN. The second SSN listed on the return is the spouse's or secondary taxpayer's SSN. Dependent SSNs are those listed on the tax return in the area for claiming dependent exemptions. Preliminary research had estimated the error rate to be .5 percent for spouse SSNs and .1 for primary SSNs (Czajka and Schirm, 1992a). Dependents were expected to have even higher error rates than secondary SSNs. This likelihood of error implied that returns of unrelated taxpayers, the correct users of the SSNs in error, could be selected and linked to panel families. Since these returns do not represent panel members, a review was necessary to remove them from the data.

2. RESULTS OF THE REVIEW

The Panel Review Project began in November, 1991, with the manual review of panel and family links. The review process is not yet complete, so the results presented in this section are only preliminary indications, based on incomplete data.

As stated previously, all types of SSNs are used to link panel members and family members, and an error in any of these SSNs could result in the linking of the incorrect tax return to the panel and tax family. Correcting SSNs was, therefore, a main goal of the review project.

Figure 1 shows the percentages of primary, spouse, and dependent SSNs that were found to be in error or to be questionable. "Errors" are SSNs that contained mistakes that we could and did correct during the review. "Questionable" SSNs are ones where we did not have enough information to determine the correct SSN, but we have indications that these SSNs are not correct.

Figure 1.--Results of Review of SSNs

Type of SSN	Percent in error	Percent questionable
Primary	0.03	0.15
Spouse	1.1	2.4
Dependent	1.7	1.8

As expected, dependent SSNs had the highest percentage of error, 1.7%, although not much higher than spouses, 1.1%. Spouses had the highest percentage of SSNs in the "questionable" category, 2.4%. The total of error and questionable SSNs is about 3.5% each for both spouses and dependents. The low percentage of error and questionable SSNs for the primaries, .03% and .15%, reflects the amount of IRS administrative processing in place when our data were collected. Primary SSNs were put through more rigorous processing than spouse or dependent SSNs and many transposition and keypunch errors were eliminated. These results also undercount the actual number of primary SSNs in error. Returns are selected for the panel by both the primary and secondary SSN. Therefore, if the primary is incorrect and the spouse is correct, the return is still part of the data file and the primary SSN can be corrected. However, if a panel member is single, and files with an incorrect SSN, we will not be able to select that return and, therefore, cannot correct the SSN in our review. These errors in primary SSNs lead to attrition in the panel, a topic which has not been examined yet. The overall error rate for all SSNs combined was 0.8%.

Figure 2.--SSN Errors by Number of Incorrect Digits

Type of SSN	Types of errors (%)		
	1-2 Digits	3-4 Digits	5-9 Digits
Primary	33.7	1.3	65.0
Spouse	76.7	2.5	20.8
Dependent	81.4	4.3	14.3

Figure 2 breaks out the known SSN errors by the number of incorrect digits. For primary SSNs the largest percentage of errors had five or more incorrect digits, what we considered a completely wrong SSN. However, spouses and dependents had the highest percentage of their errors with 1 or 2 incorrect digits, most likely transposition and keypunch errors. Much of this difference, again, is explained by the difference in IRS administrative processing for primaries and other SSNs.

3. THE PANEL REVIEW PROJECT

Longitudinal review of the panel returns did not begin until three years of data had been collected, which amounted to 330,956 tax returns. Before beginning such an extensive project, SOI found it desirable to gain the input of the end users of these data--the customers. A main concern of our customers with regard to the review and editing of the panel links was the need for flexibility. Since the collection of longitudinal data for tax policy analysis on such a large scale is unprecedented, it is likely that future analysis and use of the panel will dictate that previous decisions be amended. Our customers were also given the final say in the treatment of "odd" cases that were discovered during review, so that they would be linked according to their needs for policy analysis.

3.1 Objectives of the Review

There were several objectives for this initial review of the panel links. The overriding goal of the project was to examine suspicious linkings and determine the validity of the links. False links would be coded for exclusion from the data setand valid links would be coded to indicate that they had been verified.

A concurrent goal was the cementing of the original panel selection. This included determining units that were out of scope due to taxpayer reporting errors and processing errors in the base year. An example of a common error that affected panel selection is the dependent status indicator field. Dependent taxpayers are supposed to check off a box on their tax return, indicating that they are claimed as a dependent on another person's return. However, this requirement was first introduced in Tax Year 1987, our base year, and errors occurred. Consequently, some dependent taxpayers were mistakenly selected into the panel under the assumption that they were non-dependents and, therefore, are out of the scope of the panel sample.

Another issue in cementing the panel selection was the determination of the correct dependent panel members. In order to identify all true panel member dependents, those claimed by panel members in 1987, every dependent SSN that was listed on a panel return in 1987, 1988, or 1989 was designated a panel member. Part of the review separated the true panel dependents from the transitory, non-panel dependents, who became part of the panel member's family in later years.

One important topic that was not the focus of our panel review was the correction of the tax data items--the "money amounts." These data had been processed through the SOI data editing system as the returns were originally selected in 1987, 1988, and 1989. The money fields were tested and corrected during this editing. The panel review was only focused on determining that the return linkings and identifying characteristics of the panel members (such as SSNs) were correct. This allowed us to preserve the original taxpayer reporting behavior, rather than altering the data to mimic compliance with the current tax law. We also implemented coding to indicate when incorrect identifying information may have prevented the correct panel member's return from being selected and linked.

3.2 Computer Review

Since this first panel review was to examine over 330,000 returns, developing computer screenings to reduce the amount of manual review was critical. Consistency over the three years was the most important criterion for our computer review. If certain key characteristics were the same for all three years, we could confidently designate those panel links as correct links. The key characteristics were marital status, number of dependents claimed, social security number, dependency status, and name controls. The name controls involved a comparison of the name control from the tax return to the name control from the Social Security Administration (SSA) for that specific SSN. If a panel member was single in all three years, claimed no dependents, used the same SSN, was never claimed as someone else's dependent, and the name controls matched, the links between the returns were considered valid. Similar comparisons were also done for head of household and married filing joint statuses, as well as all three filing statuses that claimed two or less dependents. For joint returns, the secondary SSN was checked for consistency in all three years, and when dependents were claimed, their SSNs were checked for consistency.

We also delineated certain conditions which we believed strongly indicated an incorrect linking. These conditions overrode the "clean link" conditions described above. Examples of such conditions are as follows:

- Non-dependent returns that were claimed as dependents;
- Prior year return filers for the base year;
- Dependent Social Security name controls that did not match their parents; and
- Dependent returns whose zip codes didn't match their parents.

As the manual review progressed, it was determined that several of the conditions listed above were too general and further computer review would be beneficial. The addition of a check of dependent SSA name controls against both parents' SSA name controls allowed us to confidently designate these returns as being linked correctly. As a result, a total of nearly 54% of the panel return links were designated clean through computer review. Figure 3 illustrates the volume of returns for each stage of this iterative process.

Figure 3.--Computer Review Results

Results of each stage of computer review	Number of returns
Total returns	330,956
Initial computer review	118,565
Subtotal	212,391
Prior year computer review	1,517
Subtotal	210,874
Additional computer review	57,721
Returns for manual review	153,153

3.3 Manual Review

To determine the validity of links through a manual review, several additional pieces of information were available for the editors. Identification information such as full name and address was important, as well as the marital status, dependent status, and types of exemptions claimed. Most important in the actual review, as mentioned before, was information obtained from the Social Security Administration files provided to the Internal Revenue Service -- a four-digit name control, the date of birth, and the date of death, all of which corresponded to the Social Security Administration records for a particular SSN. Income amounts, such as wages and salary income, adjusted gross income, profit from a sole proprietorship, and capital gains or losses, are helpful in determining the legitimacy of family links. Combinations of items allowed further analysis of the links. If an editor was examining the link between a dependent and his or her parents, she would match the name of the dependent on the parents' return to the name on the linked return. She could also compare the ages of the parents to the dependent, based on the type of exemption claimed. If an exemption for a dependent parent was claimed, the dependent should be older than the taxpayers; if a child, younger.

When editors examine the links between the records, they compare SSNs between years, noting changes in martial status and other situations that were not captured by our computer screening. For example, if a couple files jointly and the wife keeps her maiden name, the name control in the IRS records from the tax return may not match the name control obtained from Social Security records for the wife's SSN. When the editors are able to see the complete names of these taxpayers, as well as the complete names of the dependents, they are able to determine that the SSN for the spouse is correct. Social Security Administration information is important to determine which taxpayer is using the correct SSN (the one originally assigned to them by Social Security) when two or more taxpayers are using the same SSN. The editors will then correct the SSN of the panel member and assign appropriate codes to designate the incorrect links that resulted from the SSN error.

Sometimes the data available from the SOI database are not descriptive enough to make a determination of the correct SSN for a panel member. In such cases, we turn to other IRS computer systems which contain more information on taxpayers than the SOI data, which were limited to 1987, 1988 and 1989. We search these systems for the correct SSN of the taxpayer and are able to use information about the taxpayer prior to 1987 and after 1989. If we still cannot find the true SSN for the panel member, we code the existing SSN to serve as a flag for the possibility of erroneous selections in future years.

Flexibility, mentioned earlier as a main concern of our customers, was purposefully built into the coding for the manual review process. A status code was created to designate what action had been taken on a specific return --was the link modified, the record determined to be clean, etc. Reason codes were developed to indicate why a certain action had been taken. Then, if a change is made in the reasoning behind the modification of fields on the record, the affected records would be easy to identify and alter.

4. TAX FAMILY BEHAVIOR

The vast majority of the tax families in the panel exhibit tax filing behavior that is consistent with expectations. Married couples file jointly and claim the same children each year. Single people file and claim no children. We even see the evidence of common life events: single taxpayers marrying, joint couples divorcing, single parent families, and the birth of new children. However, we did see some unusual behavior that peaked our interest, because it was not easily explained. Since this survey is based on administrative records alone, we don't have the input of the respondents to clarify their filing patterns. We can only speculate on the possible reasoning behind the behavior.

4.1 Romance in the Panel

We began to suspect romance in our panel when we had indications that some married couples from joint returns stemmed from two separate tax families. Could two single panel members have met and married? In some instances this turned out to be the case. We determined this by looking at the two original families and the return of the newly married panel members, and we were able to verify that the newlyweds had been dependents in two separate panel families. However, many of these cases turned out to be the result of SSN errors, where a spouse, who was originally selected with his own correct SSN in 1987, lists an incorrect SSN in 1988, which happens to belong to another panel member. The chances that a panel member would use an incorrect SSN that correctly belonged to another panel member are increased by the inclusion of two sets of SSNs belonging to the Continuous Work History Study (CWHS). These SSNs are selected based on the identical last four digits. Therefore, if any of the first five digits of a correct CWHS SSN are transposed, the new SSN is likely to belong to another panel member.

4.2 The Bigamists

We have what can be called "tax bigamists" in our panel. There have been a few instances of the same man or woman (same full name and SSN) filing two joint tax returns with two different spouses for the same tax year. One particular case had one wife on the east coast and one on the west coast. Although this could be an instance of bigamy, it also could be an honest error made as a result of confusion over the timing of a divorce and remarriage. The remarried spouse considers himself married to the new person for most of the tax year, while the ex-spouse believes she was married to him for most of the tax year.

4.3 Who Claimed These Kids?

Another family behavior mystery we had to unravel was the appearance of children's tax returns which indicated that they were claimed as dependents on another return, and the corresponding parents' tax return on which no exemptions for dependents were claimed. By looking at all three years of data together, we noticed that this seemed to have an alternate year pattern. One year the parents would claim these children as dependents, the next year they wouldn't, and the following year they would claim them again. However, the children would indicate on their returns that they were being claimed as dependents all three years.

The most plausible explanation for this type of behavior is an alternate year custody arrangement that resulted from an earlier divorce decree. The arrangement must call for the mother to claim the tax exemptions one year and the father the next. However, since the divorce took place before our 1987 base year, we do not have the other parent's return to verify our speculation.

4.4 The Problems with Dependent SSNs

The Tax Reform Act of 1986 for the first time required SSNs to be reported for dependents age five and older. The age requirement has been lowered through the subsequent years; first, to age two and older and, finally, to age one and older. Before this, many parents waited until their children were old enough to begin to work in after-school or summer jobs (and begin to have deductions for Social Security) to obtain SSNs for their children. While they were waiting to obtain SSNs for their dependents, some parents filed their tax returns, reporting their own SSNs as the SSNs of their dependents. This behavior sometimes led to some unusual families being created in our data. In one case, a taxpayer used the SSN of an ex-spouse as one of their dependent SSNs. In following years, the ex-spouse and his family were pulled into the original panel family. To further complicate matters, the original panel couple and the ex-spouse couple were both claiming the same two children as dependents.

4.5 Child, Spouse or Parent? Relationships Between Taxpayers

Sometimes the relationship between the members of the tax families is a bit cloudy. The same person is claimed in one year as one type of family member and the next year as another. One example is the phenomenon of claiming a person as a child one year and filing jointly with that same person as the spouse the next year. Many times there is a large age difference between the primary taxpayer and the child/spouse; the spouse is young enough to be the primary taxpayer's child. Exemptions for dependents are classified into several categories in the SOI data: exemptions for children at home, children away from home, dependent parents, and other dependents. Rather than classify this spouse as an "other dependent" that first year, the taxpayer indicated that it was an exemption for a child. Since it seems unlikely that a taxpayer would marry his own child, we assume that he made a mistake in classifying that dependent as a child. However, this still leaves us with the puzzling combination of a primary taxpayer married to his own former dependent.

Another case, similar to the dependent spouse above, involved claiming a person as a spouse one year and a dependent parent the next. Again, the age difference makes it plausible that the spouse could be the parent of the taxpayer. Did this taxpayer have a startling revelation one year and find out that the person he married was actually his mother? Probably not. It is more likely that he misinterpreted the instructions for filing his return, but we can never be sure.

The concept of creating a tax family seems rather straightforward, when dealing with one static moment in time. However, when you begin to track these families over time, the lines between distinct families begin to blur. In another case, a single panel member claiming two dependents was selected in 1987. One of these dependents turns out to be the brother of the panel member. In 1988, the panel member files, claiming two dependents, but claims a new child, instead of the brother; this is one tax family. The brother also files in 1988 and claims a dependent parent; this is a second tax family. In 1989, we again have two families, but the dependent parent has switched to the original panel member's family and the brother claims no dependents. We will track the original panel member, the one child in 1987, and the brother, since they were all identified in the base year. Although we have only one blood family here, we have two different tax families each year.

5. CONCLUSIONS

Our review of the first three years of panel data has provided some useful insights forthe future. We now know what data items are most helpful in review, and can begin to incorporate this knowledge into more comprehensive computer screenings. The identification of incorrect SSNs will help prevent extraneous returns from being selected for the panel in future years. Although panel data for the next few years will need to be reviewed in a similar fashion, eventually, the longitudinal review of the panel will be incorporated into the current on-line editing process, which is used to edit the tax data items. Questionable links will then be resolved before they become part of the final data, thus speeding up the delivery of the panel to our customers.

Plans to release panel data to the public are in the early thinking stages now. By their nature, panel data provide a greater risk of disclosing confidential information and a greater benefit from identifying specific taxpayers. A method must be designed to ensure that this disclosure doesn't happen, by discouraging people from attempting to identify individuals, before data can be made available for public use.

NOTES

[1] See Hostetter et al. (1990) for a description of the sample design for the enriched cross-section, and Schirm and Czajka (1991) for an evaluation of this new design.

[2] See Hubbard et al. (1992) for an example of the Treasury's use of existing SOI data.

[3] For a more detailed description of the panel selection process, see Czajka and Schirm (1992a). For weighting issues seeCzajka and Schirm (1990) and Czajka and Schirm (1992b).

[4] The Social Security Administration's Continuous Work History Study is selected by the last four digits of the SSN. The SOI Individual sample was designed to have a built-in overlap with the CWHS. In the case of the panel, two specific sets of last four digits coincide with the CWHS.

REFERENCES

CZAJKA, JOHN L. and SCHIRM, ALLEN L. (1992a). "Enhancing the Representativeness of a Longitudinal Sample of Individual Tax Returns: Weighting and Sample Supplementation." *Proceedings of the 1992 Annual Research Conference*, U.S. Bureau of the Census.

CZAJKA, JOHN L. and SCHIRM, ALLEN L. (1992b). "Model-Based Alternatives to Design-Based Weights in a Panel of Individual Tax Returns." *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

CZAJKA, JOHN L. and SCHIRM, ALLEN L. (1990). "Overlapping Membership in Annual Samples of Individual Tax Returns." *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

HARTE, JAMES M. (1986). "Some Mathematical Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS." *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

HOSTETTER, SUSAN, CZAJKA, JOHN L., SCHIRM, ALLEN L., and O'CONOR, KAREN (1990). "Choosing the Appropriate Income Classifier for Economic Tax Modeling." *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

HUBBARD, R. GLENN, NUNNS, JAMES R. and RANDOLPH, WILLIAM C. (1992). "Treasury Report on Income Mobility." *Tax Notes*, vol. 55, no. 9, June 1, 1992.

SCHIRM, ALLEN L. and CZAJKA, JOHN L. (1991). "Alternative Designs for a Cross-Sectional Sample of Individual Tax Returns: The Old and the New." *Proceedings of the Section on Survey Research Methods*, American Statistical Association.