Creating Household Data From Individual Income Tax Returns

Peter Sailer and Michael Weber, Internal Revenue Service

The following paper presents some preliminary analysis of a file of tax return data, which, to every extent possible, has been reassembled to approximate family units. All Statistics of Income reports since 1916 have used "the Tax Return" as the unit of measurement. We have now created an alternative that should be more recognizable to analysts accustomed to using the family or household data produced by other agencies, such as the Census Bureau.

Plans for such a file were first announced at the 1991 meetings of the American Statistical Association (Hostetter and O'Conor, 1991). At the 1992 meetings, some of the problems in matching returns of family members were discussed (Steffick, 1992). At the 1993 meetings, low-match rates for certain types of returns were noted, and a suggestion made that further research be undertaken to verify our selection criteria (Czajka and Schirm, 1993). This paper will pick up the narrative at that point and discuss the selection program, as well as any weaknesses found in that program. Then, we will show the steps that were taken to assemble data for the "tax families." A separate section will be devoted to the problems involved in assembling married persons filing separately into tax families. And, finally, comparisons of tax return family groupings to those shown by the Census Bureau will be presented.

■ Selection of the Family File

The selection process for the 1993 Family Cross-Section File was, in theory, very simple. The starting point was the regular, annual cross-sectional sample pulled for the production of the report, *Statistics of Income, Individual Income Tax Returns*--a highly stratified sample of approximately 105,000 returns. (See Hostetter et al., 1990, for a further discussion of the Statistics of Income sample design.) From it, we removed all dependent returns--in other words, those where the filer checked a box that said, "If your parent (or someone else) can claim you as a dependent, check here."

At this point, we had three types of non-dependent returns: returns representing one unmarried taxpayer; joint returns representing two married taxpayers; and separate returns representing one half of a married couple. For each of these entities, we took all the SSN's listed for the dependents they claimed and looked for matches on the IRS's Individual Master File of all individual income tax returns. All returns found in this match were transcribed and added to the sample. In addition, for returns of married persons filing separately, we took not only the dependent SSN's, but also the secondary SSN's and looked for matches. The secondary SSN on a married filing separately return should identify the primary SSN on another married filing separately return. And, assuming everybody used their SSN's correctly, these two taxpayers should be married to one another. Extensive checks of duplicate filings with the same SSN, as well as those SSN's (primary, secondary, or dependent) that did not match to the name controls generated by the Social Security Administration, weeded out most of the false matches.

Finally, the research suggested by Czajka and Shirm was undertaken, some problems with the matching programs were detected, and missing records were retrieved from the Internal Revenue Service's historical files.

■ Assembling the Family Data

In effect, each non-dependent owner of a primary SSN in our cross-sectional sample became the head of a tax family. By getting the tax returns of their family members, we were able to add together the tax and most of the income of the tax family. The word "most" is used advisedly: many dependents with small amounts of income may not be required to file tax returns. However, to the extent that the unreported income was shown on information documents (1099's and W-2's filed by payers of income, such as banks, brokerage houses, and employers), we obtained that information by matching to these documents. A further enhancement was to obtain age and gender data for all members of each tax family. IRS gets these data from the Social Security Administration for the purpose of testing compliance with various age-specific provisions of the Tax Law, as well as for research purposes.

Up to this point, no weighting issues have arisen. Joint and single returns retain their weights, and the data on the matched dependent returns are given the same weights as those on the parents' returns. Dependent returns no longer count as frequencies when numbers of units are counted, thus eliminating the double counting of these individuals usually associated with tax return data (once as dependents, once as taxpayers). Married persons filing separately, however, do have to be arranged and weighted in a totally new manner. Basically, the returns of each spouse must be brought together into one unit, and a weight assigned that represents the probability of either spouse being selected for the sample. At this point, when this file is used to produce data, the unit of measurement can be the "tax family," not the "tax return."

Assembling Families from Married Filing Separately Returns

As indicated in the previous section, we attempted to match up married filing separately returns into groups of two in order to form tax families. However, we soon found this is not always possible or even desirable. As is shown in Table 1, about 960,000 married filing separately returns do not match to another married filing separately return. In part, this could simply be a matter of procrastination--one spouse did not get around to filing a 1993 return until after December 31, 1994. To the extent this is true, some family units will be incomplete. But there are other reasons for not finding matching married filing separately returns, and these cases do not constitute incomplete data, since they involve couples who are legally married but not, in fact, living together. This is certainly the case when a sampled married filing separately return matches to a head of household return. Legally, two individuals who file this way should not have been living together for the last six months of the tax year, and one of them (the spouse using the head of household filing status) should have been providing a home for at least one child during those six months. In all these cases, we classified the married filing separately return and its dependents as a separate tax family from the head of household return and its dependents.

Table 1. Married Filing Separately (MFS) Returns (Numbers in thousands)

Total	2,369
No match to another MFS return:	960
Match, but different addresses	378
Match, same address	1,031 *

* These returns represent 516,000 families

There were also instances where we achieved a match of two married filing separately returns but decided not to combine them into the same tax family. This occurred when the two spouses filed from different addresses. Our sample yields a weighted estimate of 378,000 of these couples. In a legal tax sense, they represent one "tax family"--they are required to use the same form of deduction, and most of their deduction limitations and amounts are split evenly between their two returns. However, under the Census concept of family, they would represent two distinct families. So, for the purpose of this paper, we used the first nine digits of the taxpayer return address to keep separately those married filing separate taxpayers who lived at different addresses from each other. (The reason we used only the first nine digits of the return address is that our research showed that after the ninth digit, things like use or nonuse of apartment numbers or abbreviations made otherwise identical addresses appear to be different.)

This leaves us with a little over a million married filing separately taxpayers who actually lived together, and whom we combined into 516,000 families, along with their dependents. They meet all the criteria of the Census "married couple households."

Why Do Married Taxpayers File Separately?

If the million or so married persons filing separately whom we are combining into families do indeed constitute "married couple households," why are they filing separately? The tax return instructions state that married couples will usually owe less in taxes if they file jointly, and all are legally entitled to do so. However, it should be noted that if both spouses have similar levels of income, so that they both fall into the same taxable income class in the marginal tax rate tables, they will not end up paying any more taxes filing separately than if they had filed jointly. This is because the size classes in the married filing jointly marginal tax rate table are exactly twice the size of the size classes in these tax rate tables, so these couples will fall into the same tax rate class whether they file jointly or separately.

1993 Tax Rate Schedules

Married filing separately

$1 through $18,450	15%
$18,451 through $44,575	28
$44,576 through $70,000	31
$70,001 through $125,000	36
$125,001 and above	39.6

Married filing jointly

$1 through $36,900	15%
$36,901 through $89,150	28
$89,151 through $140,000	31
$140,001 through $250,000	36
$250,001 and above	39.6

As it turns out, 69.5 percent of the married filing separately couples who lived at the same address consisted of two taxpayers who were in the same tax bracket, so we know they lost nothing by filing this way. When we look at the married filing separately couples who did not share the same address, only 47 percent were in the same marginal tax class.

So, we know a majority of these couples did not lose by filing separately. But did they gain? They did to the extent that they had some of the deduction items, which are reduced by a percentage of adjusted gross income, most importantly, medical deductions and miscellaneous deductions (including employee business expenses). Medical expenses are reduced by 7.5 percent of adjusted gross income, and miscellaneous deductions by 2 percent of adjusted gross income. So, the less income you have, the bigger the deduction. Not surprisingly, a high proportion of married filing separately returns have medical and miscellaneous deductions: 10.1 percent have medical deductions (as opposed to 4.8 percent for all returns), 19.4 percent have miscellaneous deductions (6.8 percent for all returns).

■ Tax Families versus Census Families

Column 1 of Table 2 shows data from the March 31, 1994, Current Population Survey. Households are distributed by those categories, which, at least in theory, can be replicated by the new SOI Family File. (Such categories as one-person households, institutionalized individuals, children living with other than parents, and unrelated individuals living in households were omitted because no similar category could be constructed from the SOI Family File.) Column 2 shows comparable data from the SOI Family File, with family groupings placed into the various household categories based on marital status and presence or absence of exemptions for dependent children living at home. Overall, our file of income tax returns covers nearly 233 million individuals, representing 89.4 percent of a population of 261 million counted by the Census Bureau. The fact that the file covers only 22 million of the 28 million couples without children (about 78 percent) is not, on the face of it, too alarming, since this category includes a lot of elderly people who have more lenient filing requirements. What is alarming is that we show 1.9 million more households of married couples with children than does the Census Bureau, and all other categories involving children are

Table 2. Households classified by household type, Census and SOI Family File (in thousands)

Type of household	Census	SOI Family File
Total population	260,651	232,920
Married couples living together, total	53,171	48,814
Without children	28,113	21,861
With children	25,058	26,953
Children in those households	48,084	51,551
Other householders with children	8,961	13,839
Children in those households	18,591	21,085
All other individuals	78,673	48,817

similarly overstated on the SOI side.

There is, however, a logical reason why our children (and, therefore, our families with children) are so high when compared to data from the March 1994 Current Population Survey. The Census side includes only children under 18, whereas the SOI side includes all dependent children living at home who are being claimed as exemptions. For tax purposes, parents who want to keep putting up a child for the rest of their lives can keep claiming that child as an exemption. For children under the age of 19, and for full-time students under the age of 24, it does not even matter how much income the children have, as long as the parents are supplying the majority of the support. Luckily, as mentioned earlier, we have year of birth information for all taxpayers and dependents for whom we have a valid SSN.

Table 3 shows data from the SOI Family File adjusted to count as children at home only those dependent children who were under the age of 18. This adjustment caused the number of married couples without children to rise to 25 million, or about 88 percent of the Census figure. Married couples with children at home have come down to 24 million, or 96 percent of the Census figure. As a matter of fact, all categories save one are below the Census figure, and in what appears to be a reasonable range of the CPS data. The one remaining problem is the number of households, other than those of married couples, showing dependent children living at home. This phenomenon has been bedeviling Census and IRS for some time. IRS data consistently show more unmarried heads of households than appear in Census reports. As Table 3 shows, it is the number of households, not the number of children in these households, that appears overstated in the SOI data.

Table 3. Households classified by household type, Census and SOI Family File modified to include only children under 18 in each household (in thousands)

Type of household	Census	SOI Family File (modified)
Total population	260,651	232,920
Married couples living together, total	53,171	48,814
Without children	28,113	24,843
With children	25,058	23,971
Children in those households	48,084	44,748
Other householders with children, total	8,961	12,984
Children in those households	18,591	16,449
Male householders with children	1,314	4,990
Children in those households	2,257	5,829
Female householders with children	7,647	7,994
Children in those households	16,334	10,620
All other individuals	78,673	61,111

Table 3 further classifies unmarried heads of households by sex. It shows that the overstatement on the IRS side is largely among males. It could be that couples living together without being legally married are more likely to declare themselves married when answering a household survey than when filling in a tax return, where a misstatement would have legal consequences (and, if both are working, could actually be disadvantageous from the point of view of total tax bill). On the other hand, it is also possible that men who have joint custody of their children with their ex-wives are counting themselves as heads of households, even though the children have not lived with them for the requisite "more than half" of the year. The fact that the number of exemptions for children at home in these households is not overstated on the SOI side, only the number of households, may be significant. It is easy to imagine a scenario where the exemptions for children are divided between ex-husband and ex-wife in such a way that both manage to claim head-of-household status; for Census purposes, however, only the ex-wife may claim to be the actual custodial parent.

■ Conclusion

It would appear that, with a little bit of "tweaking" of the data, we have succeeded fairly well in replicating some of the Census household categories. Married couples with children, married couples without children (except for the aged), and households with children and a female head are all well represented in our data base. More research is needed on the overstatement of male-headed households with children. Unfortunately, it is doubtful that much can be done to further categorize what is now labeled as "all other individuals." But the categories that have been established should be invaluable in studying the many "family" issues currently under discussion.

■ References

Czajka, John L. and Schirm, Allen L. (1993), "The Family That Pays Together: Introducing the Tax Family Concept, with Preliminary Findings," *Proceedings, 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, Section on Survey Research Methods*, American Statistical Association.

Hostetter, Susan and O'Conor, Karen (1991), "Satisfying the Need of Income Policy Modelers While Preserving the Reliability of Descriptive Statistics," *Proceedings, Section on Survey Research Methods*, American Statistical Association.

Internal Revenue Service, *Statistics of Income--1993, Individual Income Tax Returns*, Publication 1304 (Rev. 3-96).

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