The Family that Pays Together: Introducing the Tax Family Concept, with Preliminary Findings

John L. Czajka and Allen L. Schirm, Mathematica Policy Research, Inc.

To assess the potential implications of alternative tax proposals, policy analysts in the Congress and the administration rely heavily on micro-level data generated by the tax system. Using large, statistical samples of tax returns, analysts simulate the workings of a proposed piece of legislation at the micro level and then aggregate these results to determine the impact on total revenue and the incidence of the tax burden.

While the tax return data bases are rich in line item detail, have exceedingly low nonresponse, and contain very accurate responses, they have several important limitations. First, their coverage is limited to the tax filing population based on current filing requirements. As a result, they cannot readily represent the impact of tax legislation that would broaden the filing requirements. Second, they do not include income amounts outside of those currently required to be reported, and they do not include expenses outside of those currently permitted to be deducted. As a result, they cannot readily represent the impact of tax legislation that would expand the definition of taxable income or increase potential deductions. Third, they do not include data on family members outside the basic filing unit (in particular, separately filing spouses, children, and other dependents), making it difficult to estimate the impact of tax legislation upon families — a unit which interests policymakers much more than does the tax filing unit.

To address these limitations, both the Treasury Department's Office of Tax Analysis and Congress's Joint Committee on Taxation employ statistical matching to link Current Population Survey (CPS) records to sample tax records to create data files suitable for tax policy analysis (see, for example, Cilke and Wyscarver 1990). They also impute a number of items that may not be reported on all returns, such as expense items for nonitemizers and nontaxed income components. The enhanced files extend coverage to the entire population, expand income and expenses to include those that might become taxable or deductible, and permit estimation for families as well as filing units. While improving the applicability of the data to tax policy analysis, these enhancements introduce error, which is not measured but may be sizable.

The recent redesign of the Statistics of Income (SOI) sample of individual tax returns — the principal source of microdata for tax policy analysts — addressed the third limitation of the annual tax data. A major feature of the redesign is the collection of data for entire tax families, rather than just filing units. A tax family includes all returns that are linked by dependency. It is hoped that the availability of data on tax families will reduce policymakers' need to rely on simulated or statistically-matched family data for tax policy modeling.

This paper describes the implementation of the tax family concept by the SOI Division of the Internal Revenue Service (IRS) and presents preliminary results on the success achieved in collecting and linking tax family returns. Section 2 discusses the tax family concept, its basis in tax law, and its operationalization in the SOI Division's annual sample of individual tax returns. Section 3 describes the principal research questions addressed in this paper, and Section 4 presents preliminary findings concerning the degree to which the collection of tax family data has extended the base of data available for tax family modeling.

■ The Family Concept

To understand the origins of the tax family concept, it is necessary to understand how the SOI sample is selected. Thus we begin with a description of the sample selection process, after which we discuss how recent tax law changes have affected the ability to assemble data on families. We then describe the introduction of tax family selection and compare the tax family concept with the family concept used in the CPS (and other major household surveys).

SOI Sample Selection

Under the current SOI sample design, the basic sampling unit is the individual tax return. Each tax return processed by the IRS during a given calendar year is assigned to a stratum and then subjected to se-

lection with a probability that varies widely by stratum.

Within each stratum the sample selection procedure utilizes the taxpayer's social security number (SSN). On joint returns (filed by married couples) only the first listed or primary taxpayer's SSN is used for selection. The SSN is transformed, and truncation of the transformed value yields a five-digit pseudo-random number which is compared to a target number for that return's stratum. Returns with five-digit numbers below the target number for their respective strata are selected into the sample.

Tax Law and the Family

Each tax return represents a tax filing unit, as defined by the tax code in the year for which the return was filed. A filing unit consists of those persons whose income is included in the return. Prior to 1989, a filing unit included only one or two persons — in the latter case a husband and wife. A married couple might choose to file jointly (as did the vast majority) or separately. A couple filing separately would become two filing units. (Only in special circumstances did it benefit a couple to file separately, although a couple contemplating divorce might be ruled by considerations other than minimizing their joint total tax.)

Dependents — children or other persons receiving financial support — were required to file their own returns if they had sufficient income to meet the filing requirements. If they had some income but not enough to file (or no reason to file, such as excess withholding, when there was no legal requirement to do so), their income was not included in any filing unit. A couple might claim several dependent children, but the children's income would never be included in the couple's return. To consider the entire family a filing unit when none of the dependents filed returns might be technically correct, but it would be misleading because the family data would be incomplete if the dependents had any income. Furthermore, there was no indication on the tax return as to whether any dependent had in fact filed a return (prior to 1987 there were questions pertaining to each dependent's requirement to file but not the fact of filing). Moreover, before 1988 (for returns with filing years before 1987), parents were not asked to provide the SSNs of their dependents. Thus it would not have been practical to search the tax return data base for the returns of sample members' dependents.

Implementation of the Tax Reform Act of 1986 affected the filing requirements of families in two important ways. First, taxpayers were required to list the SSNs of all dependents aged five and older. (This requirement has since been extended to age one.) Second, filing requirements for dependents were substantially revised, such that a dependent with unearned income as low as $1 was required to file if the combination of earned and unearned income exceeded $500, a value significantly lower than the general filing limit. These two changes made it feasible to search for the returns filed by dependents of sample members and increased the potential catch (far more dependents than previously would be required to file). Beginning with the 1989 tax year, the filing requirements respecting dependents were again modified. For children under age 14, interest and dividend income as high as $5,000 could be reported on the parents' return providing the child had no other income and no withholding. This reduced the potential number of dependent returns and for the first time extended the filing unit to include dependents.

Collection of Tax Family Data

A tax family includes a taxpayer and spouse, if present, plus all dependents claimed by either. The collection of tax family data for the SOI sample began with the 1988 filing year. The sample continues to be a sample of filing units. However, the returns selected by the method described above (except returns whose filers are reported to be dependents of other taxpayers) are supplemented by the identification and collection of all returns filed by their dependents and separately filing spouses, using the SSNs reported on the "parent" returns (a dependent need not be a child, but we find it useful to think of the originally selected return as a parent return). Tax families are not defined for dependents selected into the annual sample because parents' SSNs are not reported on their dependents' returns, ordinarily.

Comparison with the Census Family Concept

There are four principal differences between the tax family concept and the Census Bureau's family definition used in the CPS. First, a tax family can consist of a single individual, whereas the CPS concept (along with most other usages of family) requires two or more persons with some degree of relationship. This distinction is more terminological than indicative of something integral to the tax family concept, however. Second, the tax family does not include coresident family members who are not claimed as dependents. For example, a child or other relative living with the tax family who does not meet the IRS dependency test cannot be included in the tax family. Such a person is included in the CPS family. Third, strictly speaking, the operationalization of a tax family does not include dependents whose incomes fall below the filing thresholds, except (beginning in 1989) where such incomes can be reported on the parents' return. Here, too, a tax family may exclude persons present in the CPS family. Fourth, a tax family may include dependents living in another household whereas the CPS family does not. In this respect, a tax family can be more inclusive than a CPS family. From the standpoint of tax policy, the fourth difference may be the most important. Operationalization of the tax family concept may facilitate the extension of tax policy analysis to consider the implications of tax law for dependency linkages extending across households.

■ Research Questions

This paper addresses two general questions about the implementation of tax family selection. First, how complete is the capture of tax family members? Since the prospective additional returns include both dependents and separately filing spouses, which present different problems and provide different kinds of value, we are interested in the capture rates of both. For dependents, how does the distribution of captured returns compare to the distribution of dependents claimed—both in number and type? For separately filing partners, what proportion of their returns are captured? Second, what are the characteristics of tax families? Specifically, how many are there, how many dependents and dependent filers do they include, and how important are the income contributions of members other than the primary filer? We investigate these questions with data from the 1988 tax year sample — the first year for which tax family data were assembled and the year for which their processing is most complete.

■ Preliminary Statistics

The 1988 Sample

Table 1 provides an overview of tax families as defined from the 1988 annual sample, which included 110,491 returns. From this sample IRS created 106,855 tax families, leaving 3,636 returns that did not define tax families. Fifteen of these returns were filed by separately filing spouses whose partners were also selected into the sample, so they are included in other tax families. The remainder (3,621) were filed by dependents. Tax families are not created for these returns, as discussed earlier.

Table 1. 1988 Annual Sample by Tax Family Headship and Filing Status

Headship and filing status	Number of returns
All returns	110,491
Returns defining tax families	106,855
Married, filing jointly	77,306
Married, filing separately	
Spouse filing	1,833
Spouse not filing	43
Single	22,088
Head of household	5,508
Qualifying widow/er with dependent child	77
Returns not defining tax families	3,636
Dependents	3,621
Separately filing spouses (partner also selected and designated family head)	15

Dependents

To assess the 1988 sample's coverage of tax family dependents, we estimated the total population of dependents and their aggregate adjusted gross income (AGI) based on 31,914 sample records identified and selected as dependents of tax family heads, and we compared these estimates to independent estimates based on the 3,621 dependent returns selected into the annual sample (77 dependents were members of both groups). For the tax family dependents we weighted

the records by the weights assigned to the tax family parent return. In the case of a separately filing couple who claimed dependents, we assigned a weight reflecting the spouses' joint probability of selection. The dependent returns selected into the annual sample were assigned their own sample weights, corresponding roughly to their inverse selection probabilities. The estimates based on the annual sample dependents are considered complete. Therefore, except for sampling error, the estimates based on tax family dependents should be less than or equal to the estimates based upon the annual sample dependents, with the ratio of the former to the latter describing the tax family sample's coverage of the dependent population.

Table 2 summarizes the results. For both the number of dependents and aggregate AGI, the tax family dependents account for 97.8 percent of the independently estimated totals. (We also examined the distributions of returns by AGI class and found them to be very similar as well.) We have not evaluated whether this percentage is significantly different from 100. Even though it may be, we still conclude that the tax family sample's coverage of the dependent population is virtually complete. Errors in the dependent SSNs recorded and transcribed to IRS's master file probably account for most of the shortfall.

Table 2. Sample Coverage of Filing Dependents, 1988

Description	Estimate
Total number of filing dependents (population)	
Tax family dependents	9,792,400
Independent estimate	10,009,900
Percent coverage	97.8%
Aggregate adjusted gross income ($ millions)	
Tax family dependents	$33,089
Independent estimate	$33,821
Percent coverage	97.8%

Table 3 provides sample counts and population estimates of the distribution of the number of dependents claimed by tax family parents and the number of these dependents who filed tax returns, based on our match of dependents to their tax families. Of the 106,855 tax families in the sample, about 55,000 or just over one-half claimed no dependents while approximately 52,000 claimed one or more dependents. Of the 52,000 who claimed one or more dependents, more than 32,000 (over 60 percent) had no apparent filers among their dependents. About 10,600 had one filing dependent and roughly 9,000 had two or more.

Table 3. Tax Families by Number of Dependents Claimed and Number of Dependents Filing

Number of dependents	Number claimed	Number filing
	Sample counts	
None	54,992	87,281
One	18,206	10,577
Two	20,204	6,405
Three	9,299	2,019
Four	2,845	449
Five	855	87
Six or more	454	37
Total tax families	106,855	106,855
	Population estimates	
None	59,619,300	91,719,600
One	17,256,000	5,656,200
Two	14,239,100	1,544,500
Three	5,663,700	258,800
Four	1,685,000	42,000
Five	488,300	6,900
Six or more	277,600	1,000
Total tax families	99,229,000	99,229,000

If we weight these results to develop population estimates, we find that, out of an estimated 99.2 million tax families, 39.6 million (about 40 percent) claimed one or more dependents, with 17.3 million claiming one and 22.3 million claiming two or more. Of the 39.6 million tax families claiming one or more dependents, approximately 7.5 million families (under 20 percent) had one or more filers among their dependents, with 5.7 million having only one filer and 1.8 million having two or more.

That the percentage of filers among claimed dependents is greater for the sample than the estimated population implies that the dependents of higher income families are more likely to file tax returns than are the dependents of lower income taxpayers. (The sample is skewed toward higher income returns.) Table 4 examines differences by AGI class.

While the fraction of tax families claiming dependents varies from about 25 percent at low income lev-

Table 4. Number of Tax Families in Sample, Weighted Mean Families Claiming Dependents, and Weighted Mean Dependents Claimed and Filing, by AGI Class

Adjusted gross income	Sample size	Families claiming dependents	Mean dependents claimed	Mean dependents filing	Percent of dependents filing
< 1	5,902	28%	2.06	0.16	8%
1 - 1,999	1,075	24%	1.64	0.04	2%
2,000 - 4,999	2,318	23%	1.73	0.07	4%
5,000 - 9,999	5,134	26%	1.77	0.08	5%
10,000 - 19,999	9,332	33%	1.78	0.13	7%
20,000 - 29,999	7,194	41%	1.90	0.20	11%
30,000 - 39,999	5,803	51%	1.92	0.23	12%
40,000 - 49,999	4,434	56%	1.91	0.29	15%
50,000 - 74,999	8,756	59%	1.90	0.43	22%
75,000 - 99,999	3,277	57%	1.89	0.60	32%
100,000 - 199,999	7,081	57%	1.95	0.67	34%
200,000 - 499,999	12,678	54%	2.10	0.80	38%
500,000 - 999,999	11,383	51%	2.14	0.94	44%
1,000,000 or more	22,488	46%	2.08	1.02	49%
All returns	106,855	40%	1.86	0.25	13%

els to nearly 60 percent at moderately high levels, the mean number of dependents claimed (among families claiming one or more) is remarkably stable. The mean number of claimed dependents who file tax returns varies widely across the income range, however, with dependents of the highest income parents being 25 times as likely to file as dependents of parents at the lowest positive income level (49 versus 2 percent).

The first two columns of Table 5 compare the average AGI of parents claiming dependents (with the incomes of separately filing partners combined) with the average AGI of filing dependents (assigned to the AGI class of their parents). What we find most surprising in these first two columns is how little the income of the average dependent filer varies with the AGI of the parents. The average dependent with parents' AGI between $100,000 and $199,999 has barely 50 percent more income than the average dependent whose parents are in the lowest postive AGI class. Only as parents' income rises well above $200,000 does their average dependents' income rise proportionately. In part this may reflect the truncation created by the filing requirement. As we saw in Table 4, the percentage of dependents who file is closely related to the parents' income class except at the highest income levels.

How important are the income contributions of dependents? According to the last two columns of Table 5, dependent filers account for only 2.2 percent of the reported AGI of all tax families claiming one or more dependents. This percentage varies some with the AGI of the parents, but the range is not very great. Dependents account for 6.8 percent of the total AGI of tax families in which the parents' income is under $2,000. This percentage drops to 3.3 percent in the next higher income class and virtually levels off through parent incomes up to $100,000. The dependent share then declines to 1.3 percent at incomes above $200,000.

Among families with one or more filing dependents, the income contribution of dependents is naturally much greater. Overall, dependents account for 6.9 percent of the income of these families, and the contribution reaches 69 percent for families with parental income in the lowest positive income class. The contribution of dependents remains above 10 percent for AGI classes up to $20,000, then declines continuously until parents' incomes reach $500,000.

Table 5. Weighted Mean AGI of Parents and Dependent Filers, and Dependent Share of Family AGI, by AGI Class of Family Heads

Adjusted gross income	Mean AGI of parents	Mean AGI of dependent filers	AGI share in families with dependents	AGI share in families with filers
< 1	- 86,961	4,289	--	--
1 - 1,999	1,164	2,326	6.8%	69.0%
2,000 - 4,999	3,566	1,739	3.3%	34.1%
5,000 - 9,999	7,752	2,464	2.5%	25.1%
10,000 - 19,999	15,067	2,764	2.3%	16.8%
20,000 - 29,999	25,006	3,194	2.5%	13.1%
30,000 - 39,999	34,969	3,063	2.0%	9.6%
40,000 - 49,999	44,753	3,319	2.1%	8.6%
50,000 - 74,999	59,830	3,349	2.3%	7.1%
75,000 - 99,999	85,199	3,496	2.4%	5.7%
100,000 - 199,999	132,377	3,749	1.9%	4.2%
200,000 - 499,999	294,212	4,855	1.3%	2.6%
500,000 - 999,999	673,824	8,811	1.2%	2.2%
1,000,000 or more	2,687,025	35,050	1.3%	2.2%
All returns	$37,811	$3,384	2.2%	6.9%

Separately Filing Spouses

Separate filing among married couples is rare because it is seldom advantageous to a couple to file separately. Many couples who file separately are in the process of divorcing, so reasons other than economic advantage account for their use of this filing status. As it turns out, this fact may be relevant to our findings with respect to the coverage of separately filing spouses through the supplemental selection of tax family members.

Table 6 reports by AGI class (of the initially selected return) the number of tax families in which a

husband and wife filed separate returns and the percentage of such families for which the supplemental selection procedure obtained a matched spouse return. Overall, matched spouse returns were found for only 54 percent (990) of the 1,833 tax families with separately filing spouses. This percentage ranged from a low of 19 percent among the 21 sample cases in the lowest positive AGI class to a high of 75 percent among the 71 sample cases with AGI between $100,000 and $199,999. Nearly 30 percent of the sample tax families with separately filing spouses are in the top AGI class, and among this group 58 percent of the spouse returns were matched. Given the small sample sizes in most of the AGI classes, the observed deviations from a match rate of 55 to 60 percent may reflect sampling error more than true differences.

Compounding the differences in match rates by AGI class, match rates within AGI classes were higher, apparently, among returns with higher selection probabilities. Thus when we weight the returns we find that the estimated proportion of the population with matched spouse returns is only 37 percent.

We are aware that some of the missing spouse returns were selected in the supplemental sample but not matched because the missing spouse used a different filing status (frequently head of household). With some additional work we will be able to match these records, but we do not expect to increase the overall match rate by more than a few percentage points. We can also determine how often the first partner failed to report an SSN for the spouse, which would make it impossible to find the spouse's return unless the spouse provided the first partner's SSN.

Table 6. Number of Separately Filing Couples in Sample and Percent with Matched Spouse Return, Unweighted and Weighted, by AGI Class

Adjusted gross income	Sample size	Percent with matched spouse	
		Unweighted	Weighted
< 1	222	44%	22%
1 - 1,999	21	19%	11%
2,000 - 4,999	53	30%	16%
5,000 - 9,999	98	46%	31%
10,000 - 19,999	205	54%	37%
20,000 - 29,999	135	61%	46%
30,000 - 39,999	76	59%	46%
40,000 - 49,999	34	71%	59%
50,000 - 74,999	51	65%	55%
75,000 - 99,999	16	69%	67%
100,000 - 199,999	71	75%	66%
200,000 - 499,999	166	51%	52%
500,000 - 999,999	170	54%	56%
1,000,000 or more	515	58%	58%
All returns	1,833	54%	37%

When spouses provide the information necessary to obtain matches, how do their incomes compare? Since both spouses are subject to selection, we would expect that, on average, each would contribute the same amount. Table 7 reports by AGI class of the first partner (the annual sample member) the weighted mean AGI for that partner's return, the weighted mean AGI of the matched spouse's return, and the ratio of the average matched spouse's AGI to the average first partner's AGI. Over all returns the two means are nearly identical, with the spouse AGI being 95 percent of the first partner's AGI. Across AGI classes defined by the first partner's income there is an inverse relationship between the two incomes, with the matched spouse's income ranging from nearly 12 times the first partner's income in the lowest AGI class to about 1/12 the first partner's income in the highest income class.

Table 7. Weighted Mean AGI of Separately Filing Sample Members and Matched Spouses, by AGI Class of Sample Members

Adjusted gross income	Mean AGI of sample members	Mean AGI of matched spouses	Ratio: spouses to sample members
< 1	- 84,898	- 4,805	--
1 - 1,999	1,118	13,091	11.71
2,000 - 4,999	3,322	11,744	3.54
5,000 - 9,999	7,352	13,887	1.89
10,000 - 19,999	14,754	16,441	1.11
20,000 - 29,999	24,704	23,952	0.97
30,000 - 39,999	34,106	26,946	0.79
40,000 - 49,999	43,739	30,281	0.69
50,000 - 74,999	56,389	38,273	0.68
75,000 - 99,999	82,464	54,379	0.66
100,000 - 199,999	133,954	44,065	0.33
200,000 - 499,999	295,451	58,910	0.20
500,000 - 999,999	715,537	110,304	0.15
1,000,000 or more	3,362,789	253,650	0.08
All returns	$22,858	$21,770	0.95

■ Future Research

The most important question for future research is: Where are the missing returns for separately filing spouses? We mentioned two strategies for investigating this problem: (1) searching the supplemental sample for partner returns with filing statuses other than married filing separately and (2) determining how often the first partner's return lacked an SSN for the separately filing spouse. Problems related to missing SSNs should diminish in the future as IRS strengthens its quality control procedures. On the other hand, if the problem lies in IRS's specifications for the supplemental sample selection and matching, the match rates will not be improved without corrective action, and

appropriate revisions need to be implemented as soon as possible to minimize the potential lost returns.

A broader question raised by this research is the following: Can tax family data replace statistically matched CPS data as the source of family data for tax policy modeling? Complete replacement of the CPS seems doubtful because dependents in families with low income may have incomes that are below the filing limits but nevertheless important to the family. We saw indirect evidence of this in Table 5. Nevertheless, our preliminary findings suggest that the role of CPS data might be reduced to one of supplementing the data from matched spouses and dependents. Limiting the CPS role is particularly desirable at the high end of the income distribution, where the CPS sample is exceedingly thin in comparison to the SOI sample.

Finally, our findings suggest how the results of tax family matches can be used to evaluate the statistical matches of CPS data to IRS data. For example, for couples filing separately, the tax family data provide evidence of the relationship between separately filing partners' incomes. If statistical matching does not reproduce these relationships, the matching algorithms should be modified.

■ Acknowledgments

This research was performed under contract to the SOI Division of the IRS. We are grateful for this support and want to express our thanks in particular to Fritz Scheuren for his encouragement of our work. We also gratefully acknowledge Daisy Ewell of MPR for her very capable programming and research support.

■ Reference

Cilke, James and Wyscarver, Roy A. (1990). *The Treasury Individual Income Tax Simulation Model.* Washington, DC: Department of the Tresury, Office of Tax Analysis. ■