NUMBER 7

FEDERAL INCOME TAXES, SOCIAL SECURITY TAXES, AND THE U.S. DISTRIBUTION OF INCOME, 1972

Daniel B. Radner

Division of Economic and Long-Range Studies

APRIL 1979
Social Security Administration
Office of Policy
Office of Research and Statistics

Working papers from the Office of Research and Statistics are preliminary materials circulated for review and comment. These releases have not been cleared for publication and should not be quoted without permission of the author. The views expressed are the author's and do not necessarily represent the position of the Office of Research and Statistics, the Office of Policy, the Social Security Administration, or the Department of Health, Education, and Welfare.

Summary

This paper reports on estimates of Federal income tax and social security tax liabilities of family units in 1972 and summarizes the methods used to make the estimates. Distributions of income both before and after subtracting those liabilities are shown. Several microdata files were combined using both "exact" and "statistical" matching of individual observations in the process of making these estimates. The 73.2 million family units had a mean total family income of $11,135, paid a mean amount of $1,234 in income tax (11.1 percent of income) and a mean amount of $321 in social security tax (2.9 percent of income), excluding the tax paid by employers. At the lower end of the income distribution, mean social security tax exceeded mean income tax in most income classes, while mean income tax was substantially higher in the middle and upper income ranges. For all family units, mean total family income after income and social security taxes was $9,581. The distribution of after-tax income by quintiles showed a seven percent greater share for the bottom quintile and a four percent smaller share for the top quintile when compared with the ranking by size of before-tax income. Income tax exceeded social security tax for about 60 percent of all family units, social security tax was higher for more than 20 percent of the units, and more than 15 percent of the units paid neither tax.

FEDERAL INCOME TAXES, SOCIAL SECURITY TAXES, AND THE U.S. DISTRIBUTION OF INCOME, 1972*

Introduction

This paper reports on one part of on-going research on the role of social security in the tax-transfer system being conducted by the Division of Economic and Long-Range Studies of the Office of Research and Statistics (ORS) of the Social Security Administration. In the aspect of the research reported on here, the emphasis is upon Federal individual income tax liabilities and social security tax liabilities of family units at different income levels in 1972.[1/][2/]

Federal individual income tax returns must be filed by all persons (or husband-wife couples) whose income is above a specified level.[3/] Marginal rates in 1972 varied from 14 percent to 70 percent. Not all persons required to file had tax liability; e.g., large deductions or tax credits could have resulted in zero liability.

Social security tax is paid by all employees in covered occupations (employers in most cases pay an amount equal to the amount paid by the employee) and by the self-employed whose net income is above $400.[4/] Most occupations are covered--civilian government employees are the major uncovered group. In 1972, for employees, the tax rate was 5.2 percent of covered wage and salary income up to a maximum of $9,000 (a maximum tax of $468).[5/] For the self-employed, the rate was 7.5 percent of net self-employment income up to $9,000 (a maximum tax of $675).[6/] In this paper, the employer share of the tax is excluded. Here we are interested in a relatively simple question--the impact incidence of income and social security taxes in relation to the cash income of the family units. That is, how much did each family unit pay directly, in cash.

For many purposes (e.g., analysis of proposals to integrate the income and social security taxes), data on who pays these taxes are needed on a family unit basis. However, such data are not regularly available for either tax. Income tax data are available on a tax unit (person or husband-wife couple) basis, and many family units are excluded because they did not file returns. Social security earnings and tax data are available on a person basis, and persons not working or working in jobs not covered by social security are excluded.[7/]

In recent years, several estimates of Federal individual income tax liability and social security tax liability by size of family unit income have been made, usually as part of larger studies of total tax burdens or of tax burdens and transfer benefits.[8/] While Federal individual income taxes are usually shown separately in such studies, social security taxes often are included as part of a larger item (e.g., all payroll taxes). While earlier estimates used aggregate income class data in the construction of the estimates, some of the more recent work has used microdata, thus allowing much greater flexibility. This paper presents estimates made using a new microdata base constructed at ORS.

In Section I of this paper, the data inputs used to construct the data are described. In Section II, statistical matching, the principal technique used to construct the data base, is discussed. Section III contains definitions and a brief discussion of the quality of the data, and the estimates are presented in Section IV. A summary of the statistical match carried out appears in Appendix A, and a more theoretical discussion of statistical matching appears in Appendix B.

I. The Data Inputs

The basic data source used to make the estimates presented in this paper was the 1972 ORS Statistical Match File. This file was constructed by matching several different data files, using both "exact" and "statistical" matching. In an exact match, information for the same person from two or more files is brought together using personal identifying information (e.g., name, address, Social Security Number).[9/] In a statistical match, the information brought together from the different files ordinarily is not for the same person, but is for similar persons; the match is made on the basis of similar characteristics, rather than personal identifying information.

The initial data file used in the construction of the Statistical Match File was what is called the 1973 Exact Match or "EM" file. The construction of the EM was carried out jointly by the Social Security Administration (SSA) and the Bureau of the Census, with the assistance of the Internal Revenue Service. The EM consists of an exact match of three sets of data.[10/] The first set was the March 1973 Current Population Survey (CPS), a survey of roughly 50,000 households conducted by the Bureau of the Census. Among other items, the CPS provided data on family composition and cash income of persons, although the income data suffered from some deficiencies (as noted in Section III). In exact matches, SSA earnings and demographic data and a limited amount of Federal individual income tax return information were matched to the CPS sample.[11/] Using those SSA earnings data, it was possible to construct relatively reliable estimates of social security tax liability for family units in the CPS. Both the SSA earnings data and the tax return data provided information which greatly increased the reliability of the statistical match which followed.

Unfortunately, the limited income tax return information contained in the EM did not include income tax liability and included very little income detail which could be used to construct estimates of income more reliable than those contained in the CPS. In order to remedy these deficiencies, detailed Federal individual income tax return information, including income amounts and amounts of tax liability, was matched in using a statistical match. A sample of roughly 95,000 Federal individual income tax returns was chosen by subsampling the Internal Revenue Service 1972 Statistics of Income (SOI) sample.[12/] SSA earnings and limited demographic information were matched to the SOI subsample using an exact match. These data were added primarily to improve the quality of the statistical match which was to be performed. This file, which will be called the "Augmentation File," or "AF," was then matched to the EM using a statistical match. The data in the two input files to the statistical match are summarized below:

Exact Match File	Augmentation File
1. Current Population Survey data (demographic, work experience, income, family composition)	
2. Social Security Administration data (earnings, demographic)	1. Social Security Administration data (earnings, demographic)
3. Internal Revenue Service data (limited income)	2. Internal Revenue Service data (detailed income, tax)

II. Statistical Matching

Because statistical matching is not a well-known technique, it will be useful to describe it briefly here. Statistical matching is a relatively new technique which has developed as a result of increased access to computers and the increased availability of computer microdata files. The principal use of the technique in economics has been to combine, at the single observation level, data from two different samples, each of which contains some data items which are absent from the other file. The most common application has been to combine data from a household survey with data from income tax returns. Usually the match is between a "base" file, which remains unchanged in the match, and a second file which is matched to the base file. Observations from the second file may be chosen with or without replacement.

Early statistical matches in economics were performed at the Bureau of Economic Analysis of the U.S. Department of Commerce in connection with estimates of the size distribution of family personal income[13] and the Brookings Institution in connection with analysis of the tax system.[14] More recent matching work has been done at Statistics Canada,[15] Yale University,[16] the Office of Tax Analysis of the U.S. Treasury Department,[17] and Brookings,[18] in addition to the work described in this paper.[19]

A statistical match can be viewed as an approximation of an exact match. The exact match cannot be performed because very few (or none) of the same people appear in both files, because information which identifies individuals is absent, or for other reasons (e.g., cost). In the statistical match used in this work, both files were samples, with very few persons appearing in both files.

In this match, for each unit in the base file (the EM), the second file (the AF) was searched for the observation which "most closely resembled" what the exact match data for that EM record were thought to be. That is, for each EM record which contained income tax return data,[20] the AF was searched for the observation which was thought to most closely resemble the tax return actually filed by that EM unit and that unit's SSA data.

Two basic steps in a statistical match can be identified: (1) for each base file record, the determination of what (part or all of) the data from an exact match would be, and (2) the search for the record which most closely approximates that estimated exact match information.[21] In step (1), for each EM record, we tried to estimate the characteristics which the AF record should have had, and in step (2) we looked for an AF record which approximated those characteristics. In this match, unlike many statistical matches which have been carried out, there were several variables which were defined (almost) identically in the two files and which were obtained from the same data source. (The AF was designed with this comparability in mind.) For those variables, the AF values searched for would be identical to (or very close to) the EM values and those searched for values could be determined with accuracy. Thus, in this match, step (2), the search for the "best" AF record for each EM record, was the major step. The variables used and the steps carried out in the statistical match between the EM and AF are summarized briefly in Appendix A.

III. Definitions

The population covered by these estimates consists of the civilian noninstitutional population residing in the 50 states and the District of Columbia plus military personnel in those areas living off post or on post with their families. The recipient units used are "family units." A family unit is either a "family" or an "unrelated individual." A family is defined as "a group of two or more persons related by blood, marriage, or adoption and residing together; all such persons are considered as members of the same family."[22/] An unrelated individual is defined as a person 14 years old or over who is not living with any relatives. An unrelated individual may live by himself, or may reside with one or more other persons not related to him.[23/]

Both income and taxes are defined to include only what might be called "cash" amounts. Income is defined to include all regularly received cash income for calendar 1972. Some of these income amounts are obtained from the data in the AF portion of the matched record, while other types are obtained from the CPS. Total money income consists of the following income types (where the source of the data is shown in parentheses): (1) wages and salaries (SOI);[24/] (2) net income from nonfarm unincorporated business or partnership (SOI); (3) net income from farm self-employment (CPS); (4) property income (interest, dividends, rent, royalty, estate and trust) (SOI); (5) social security and railroad retirement benefits (CPS); (6) public assistance (CPS); (7) other government transfer payments (unemployment compensation, workmen's compensation, government pensions, veterans' benefits) (CPS); (8) other income types (private pensions and annuities, alimony, contributions from persons outside the household, miscellaneous types) (CPS).

At this point, a word about the accuracy of the income data is needed. For most income types, income tax return data are generally considered to be more accurate than CPS income data; therefore, income tax return income amounts were used in place of the CPS amounts where it was feasible. One indication of accuracy is a comparison of aggregate amounts of income. The CPS found roughly 89 percent of total aggregate money income as estimated in independent aggregate control totals. The percentage found varied widely by income type; e.g., 96 percent for wages and salaries, 45 percent for property income. In the income estimates used in this paper, in which the CPS amounts of wages and salaries, nonfarm business and partnership income, and property income were replaced by the SOI amounts, and the sample was reweighted, aggregate total money income was roughly 93 percent of the control total.[25] For example, the amount of property income increased to 63 percent of the control total, which was a significant improvement, although still not satisfactory.[26]

The types of taxes included here are total Federal individual income tax and social security employee and self-employment taxes. The income tax estimates were made using amounts which appear on unaudited tax returns; no attempt was made to estimate the liability after audit. Total Federal individual income tax consists of income tax after credits plus additional tax for tax preferences ("minimum tax").[27][28] The aggregate amount of total income tax in the Statistical Match File is $90.2 billion, which is quite close to the SOI figure ($93.3 billion) after appropriate adjustments (e.g., exclusion of decedents and of some military personnel) have been made.

The social security employee tax rate in 1972 was 5.2 percent of taxable wages up to $9,000. Although the employer contributed roughly an equal amount, the employer's share was not included in this analysis. This exclusion is in keeping with the "cash" basis and estimation of impact incidence used. Social security self-employment tax, which was levied at a rate of 7.5 percent of taxable earnings up to $9,000, was included.

In the remainder of this paper we will refer to "social security taxes", which are defined as the sum of employee and self-employment taxes. The aggregate amount of social security tax in the Statistical Match File is $23.5 billion, which is quite close to the administrative total of $25.4 billion, after appropriate adjustments have been made.[29]

In summary, the aggregate amounts of income and social security taxes are quite accurate, while the income amounts on average are underreported by roughly 7 percent. Thus, the estimates of effective tax rates shown in the next section are, on average, slightly overstated.

IV. Estimates[30] [31]

In 1972 mean total family income for all family units was $11,135 (see Table 1).[32] Those 73.2 million units paid a mean amount of total income tax of $1,234, which amounted to 11.1 percent of their total money income. Roughly 70 percent of all units paid total income tax; for those that did, the mean amount of tax was $1,772. As would be expected, the effective income tax rate rose throughout almost the entire income range--from the $2,000-2,999 class through the top class.[33] The effective rate rose above 10 percent beginning with the $16,000-17,999 class and reached the mean in the $18,000-19,999 class.

All family units paid a mean amount of social security tax of $321, which amounted to 2.9 percent of their total money income. About 77 percent of all units paid social security tax; for those that did, the mean amount of tax was $416. The effective social security tax rate rose from the $2,000-2,999 class (1.9 percent) through the $8,000-8,999 class (3.9 percent) and declined steadily beginning with the $10,000-11,999 class, reflecting the limit of $9,000 on per worker taxable earnings. The mean amount of social security tax for all units in the class continued to increase through the $25,000-29,999 class, despite the declining effective rate, reflecting primarily the increasing proportion of units in that income range with more than one earner. In the $10,000-11,999 class, just over 40 percent of all units had two or more persons with SSA taxable earnings, while in the $25,000-29,999 class more than 60 percent had two or more persons with SSA taxable earnings.

Table 1 shows that for non-negative incomes below the $5,000-5,999 class, mean social security taxes (for all units in the class) exceeded mean income taxes in each class. The differences between the means in these classes were quite small--the largest difference was $23 in the $2,000-2,999 class. For the $5,000-5,999 class and above, mean income taxes were higher, by an increasing amount as income rose. By the $10,000-11,999 class, the effective income tax rate was more than double the effective social security tax rate. As might be expected, the combined effective rate was dominated by the income tax effective rate.

The means and effective rates are shown separately for families and unrelated individuals in Tables 2 and 3, respectively. The patterns were generally similar to those for all family units, although for families mean social security taxes exceeded mean income taxes through the $5,000-5,999 class, and the differences between those means in classes in which social security taxes were higher were larger, reaching a peak of $54 in the $4,000-4,999 class. For unrelated individuals, mean social security taxes exceeded mean income taxes through the $2,000-2,999 class.

When the estimates are broken down by size of family (not shown), some differences can be seen. For families with two persons, mean social security tax exceeded mean income tax through the $4,000-4,999 class; for three-person families, through the $5,000-5,999 class; for four-person families, through the $6,000-6,999 class; and for families with five persons or more, through the $9,000-9,999 class.

When the estimates are broken down by number of persons in the unit with SSA taxable earnings (not shown), the results are as might be expected--the greater the number of persons with SSA taxable earnings, the greater the importance of social security taxes relative to income taxes (holding income constant). For families with one SSA earner, mean social security tax was higher than mean income tax through the $5,000-5,999 class. For families with two SSA earners, mean social security tax was higher through the $6,000-6,999 class, while for families with three or more SSA earners, mean social security tax was higher through the $9,000-9,999 class.

The distribution by size of total family income among quintiles, for all family units, is shown in Table 4. The bottom quintile received 3.24 percent of aggregate total family income,[34/] paid 0.80 percent of aggregate total income tax and 2.78 percent of aggregate social security tax. In contrast, the top quintile received 46.07 percent of income, paid 63.38 percent of total income tax, and 35.86 percent of aggregate social security tax. While the top quintile paid a much greater proportion of total income tax than of social security tax, the other four quintiles paid greater proportions of social security tax than of income tax.

The bottom quintile (ranked by before-tax income) received 3.57 percent of income after income and social security taxes, while the top quintile (ranked by before-tax income) received 44.18 percent of after-tax income. The share of those units in the bottom quintile before tax was 10 percent higher in terms of after-tax income than in before-tax income. For the second quintile, the after-tax share was 8 percent higher, for the third quintile it was 3 percent higher, and for the fourth quintile it was roughly one percent higher. For units in the top quintile before tax, the after-tax share was 4 percent lower than the before-tax share.

For all family units, mean total family income after income and social security taxes was $9,581. When all family units are reranked according to income after income and social security taxes (see Table 5), the changes by quintile are not very large. The bottom quintile received 3.46 percent of after-tax income, and the top quintile received 44.23 percent. Compared to the share of before-tax income, the share of after-tax income of the bottom quintile (after reranking) was 7 percent higher, the share of the second was 7 percent higher, the share of the third was 3 percent higher, the share of the fourth was about one percent higher, and the share of the top quintile was 4 percent lower.[35/] The distribution of after-tax income by income size classes is shown in Table 6.

Roughly 70 percent of all units paid total income tax--78 percent of families and 46 percent of unrelated individuals (Table 7).[36/] About 77 percent paid social security tax--85 percent of families and 52 percent of unrelated individuals. Approximately 17 percent of all units paid neither tax--9 percent of families and 39 percent of unrelated individuals. For all family units, 64 percent paid both taxes--73 percent of families and 36 percent of unrelated individuals. Of the units which paid both taxes, roughly six out of seven paid more income tax than social security tax. About six percent of all units paid income tax but no social security tax, while approximately 13 percent paid social security tax but no income tax.

Altogether, roughly 60 percent of all family units paid more income tax than social security tax, more than 20 percent paid more social security tax than income tax, and more than 15 percent paid neither tax.[37/]

Table 1--Size Distribution of Total Family Income, Mean Income Tax, Mean Social Security Tax, and Effective Tax Rates, All Family Units [a]

Size of Total Family Income ($)	Thousands of Units	Mean Total Family Income ($)	Mean Income Tax (All Units) ($)	Mean Social Security Tax (All Units) ($)	Income Tax as a Percent of Income	Social Security Tax as a Percent of Income
Negative	206	-12,496	1,517	124	--	--
0-999	2,365	455	20	27	4.4	5.9
1,000-1,999	4,393	1,530	19	28	1.2	1.9
2,000-2,999	4,981	2,483	23	46	0.9	1.9
3,000-3,999	4,410	3,480	65	86	1.9	2.5
4,000-4,999	4,243	4,488	115	126	2.6	2.8
5,000-5,999	3,900	5,487	213	181	3.9	3.3
6,000-6,999	3,669	6,496	314	231	4.8	3.6
7,000-7,999	3,893	7,485	448	280	6.0	3.7
8,000-8,999	3,716	8,499	599	329	7.0	3.9
9,000-9,999	3,796	9,495	724	372	7.6	3.9
10,000-11,999	6,823	10,989	984[b]	408	9.0[b]	3.7
12,000-13,999	6,135	12,966	1,201	437	9.3	3.4
14,000-15,999	5,125	14,968	1,461	491	9.8	3.3
16,000-17,999	3,737	16,968	1,786	525	10.5	3.1
18,000-19,999	2,905	18,920	2,103	537	11.1	2.8
20,000-24,999	4,300	22,143	2,664	589	12.0	2.7
25,000-29,999	2,042	27,210	3,641	637	13.4	2.3
30,000-39,999	1,423	33,965	5,090	615	15.0	1.8
40,000-49,999	478	44,117	8,268	588	18.7	1.3
50,000-99,999	517	65,095	16,827	551	25.8	0.8
100,000 +	103	167,533	69,075	504	41.2	0.3
Total	73,160	11,135	1,234	321	11.1	2.9

[a] No attempt has been made to suppress estimates for groups with small numbers of observations; such estimates should be used with caution. A very rough estimate of the number of observations in a group can be obtained by dividing the weighted number by 1,568, the mean sample weight for all family units.

[b] If one record with a large capital gain and a large income tax liability on that gain (roughly $150,000) is excluded, mean income tax for this income size class falls by $49 and income tax as a percent of income falls by 0.5.

Table 2--Size Distribution of Total Family Income, Mean Income Tax, Mean Social Security Tax, and Effective Tax Rates, Families [a]

Size of Total Family Income ($)	Thousands of Units	Mean Total Family Income ($)	Mean Income Tax (All Units) ($)	Mean Social Security Tax (All Units) ($)	Income Tax as a Percent of Income	Social Security Tax as a Percent of Income
Negative	152	-14,551	1,920	155	--	--
0-999	591	448	54	63	12.1	14.1
1,000-1,999	1,157	1,567	60	61	3.8	3.9
2,000-2,999	2,052	2,513	26	61	1.0	2.4
3,000-3,999	2,523	3,500	54	89	1.5	2.5
4,000-4,999	2,750	4,496	72	126	1.6	2.8
5,000-5,999	2,622	5,491	136	180	2.5	3.3
6,000-6,999	2,741	6,500	246	236	3.8	3.6
7,000-7,999	2,875	7,482	369	290	4.9	3.9
8,000-8,999	2,870	8,505	520	340	6.1	4.0
9,000-9,999	3,151	9,495	641	382	6.8	4.0
10,000-11,999	5,873	11,005	923[b]	422	8.4[b]	3.8
12,000-13,999	5,522	12,972	1,145	450	8.8	3.5
14,000-15,999	4,822	14,975	1,423	503	9.5	3.4
16,000-17,999	3,545	16,974	1,752	537	10.3	3.2
18,000-19,999	2,767	18,919	2,069	549	10.9	2.9
20,000-24,999	4,125	22,146	2,627	603	11.9	2.7
25,000-29,999	1,982	27,218	3,604	646	13.2	2.4
30,000-39,999	1,371	33,961	5,067	627	14.9	1.8
40,000-49,999	442	44,092	7,994	602	18.1	1.4
50,000-99,999	484	64,786	16,698	561	25.8	0.9
100,000 +	100	168,591	69,333	510	41.1	0.3
Total	54,516	13,111	1,474	385	11.2	2.9

[a] See footnote a, Table 1.

[b] If one record with a large capital gain and a large income tax liability on that gain (roughly $150,000) is excluded, mean income tax for this income size class falls by $57 and income tax as a percent of income falls by 0.5.

Table 3--Size Distribution of Total Family Income, Mean Income Tax, Mean Social Security Tax, and Effective Tax Rates, Unrelated Individuals a/

Size of Total Family Income ($)	Thousands of Units	Mean Total Family Income ($)	Mean Income Tax (All Units) ($)	Mean Social Security Tax (All Units) ($)	Income Tax as a Percent of Income	Social Security Tax as a Percent of Income
Negative	54	-6,754	391	38	--	--
0-999	1,773	457	8	15	1.8	3.3
1,000-1,999	3,236	1,517	4	17	0.3	1.1
2,000-2,999	2,929	2,462	21	36	0.9	1.5
3,000-3,999	1,888	3,452	80	80	2.4	2.4
4,000-4,999	1,493	4,472	195	125	4.4	2.8
5,000-5,999	1,278	5,477	370	181	6.8	3.3
6,000-6,999	928	6,481	515	219	7.9	3.4
7,000-7,999	1,018	7,491	670	255	8.9	3.4
8,000-8,999	846	8,480	865	291	10.2	3.4
9,000-9,999	645	9,496	1,130	323	11.9	3.4
10,000-11,999	950	10,892	1,360	322	12.5	3.0
12,000-13,999	613	12,921	1,707	320	13.2	2.5
14,000-15,999	303	14,863	2,063	298	13.9	2.0
16,000-17,999	192	16,866	2,418	299	14.3	1.8
18,000-19,999	138	18,943	2,788	284	14.7	1.5
20,000-24,999	175	22,074	3,524	254	16.0	1.2
25,000-29,999	60	26,930	4,863	347	18.1	1.3
30,000-39,999	52	34,094	5,685	293	16.7	0.9
40,000-49,999	35	44,425	11,697	404	26.3	0.9
50,000-99,999	34	69,533	18,693	400	26.9	0.6
100,000 +	3	135,086	61,159	316	45.3	0.2
Total	18,644	5,359	531	133	9.9	2.5

a/ See footnote a, Table 1.

Table 4--Distribution of Total Family Income by Quintiles

Quintile	Percent of Total Family Income	Percent of Total Income Tax	Percent of Social Security Tax	Mean Total Family Income ($)	Mean Total Income Tax (All Units) ($)	Mean Social Security Tax (All Units) ($)	Income Tax as a Percent of Income	Social Security Tax as a Percent of Income	Percent of Total Family Income After tax
					ALL FAMILY UNITS				
Bottom	3.24	0.80	2.78	1,804	49	45	2.7	2.5	3.57
2	9.65	3.33	10.84	5,373	205	174	3.8	3.2	10.43
3	16.55	11.56	21.93	9,214	713	352	7.7	3.8	17.01
4	24.49	20.95	28.61	13,635	1,293	459	9.5	3.4	24.81
Top	46.07	63.38	35.86	25,649	3,911	576	15.2	2.2	44.18
Total	100.00	100.00	100.00	11,135	1,234	321	11.1	2.9	100.00
					FAMILIES				
Bottom	4.92	1.22	5.32	3,225	90	102	2.8	3.2	5.39
2	11.59	5.35	15.18	7,598	394	292	5.2	3.8	12.29
3	17.22	12.62	21.93	11,289	930	422	8.2	3.7	17.66
4	23.57	20.52	26.42	15,451	1,512	509	9.8	3.3	23.87
Top	42.70	60.28	31.14	27,992	4,443	599	15.9	2.1	40.79
Total	100.00	100.00	100.00	13,111	1,474	385	11.2	2.9	100.00
					UNRELATED INDIVIDUALS				
Bottom	2.95	0.42	2.39	790	11	16	1.4	2.0	3.25
2	8.09	0.56	4.16	2,168	15	28	0.7	1.3	9.05
3	13.82	4.18	13.59	3,703	111	90	3.0	2.4	14.92
4	24.18	19.47	32.95	6,479	517	219	8.0	3.4	24.46
Top	50.96	75.38	46.92	13,655	2,001	312	14.7	2.3	48.32
Total	100.00	100.00	100.00	5,359	531	133	9.9	2.5	100.00

Table 5--Distribution of Total Family Income After Tax by Quintiles of After-Tax Income

Quintile	Percent of Total Family Income After Tax	Mean Total Family Income After Tax ($)
	ALL FAMILY UNITS	
Bottom	3.46	1,658
2	10.36	4,963
3	17.04	8,163
4	24.80	11,880
Top	44.23	21,188
Total	100.00	9,581
	FAMILIES	
Bottom	5.26	2,959
2	12.23	6,881
3	17.69	9,952
4	23.87	13,429
Top	40.95	23,038
Total	100.00	11,252
	UNRELATED INDIVIDUALS	
Bottom	3.18	747
2	9.03	2,120
3	14.85	3,486
4	24.36	5,719
Top	48.58	11,404
Total	100.00	4,695

Table 6--Size Distribution of Total Family Income After Tax,
All Family Units, Families, Unrelated Individuals a/

Size of Total Family Income After Tax ($)	All Family Units		Families		Unrelated Individuals	
	Thousands of Units	Mean Income After Tax ($)	Thousands of Units	Mean Income After Tax ($)	Thousands of Units	Mean Income After Tax ($)
Negative	448	-7,966	293	-10,683	155	-2,825
0-999	2,312	478	585	492	1,727	474
1,000-1,999	4,505	1,532	1,201	1,574	3,305	1,516
2,000-2,999	5,258	2,489	2,177	2,528	3,081	2,461
3,000-3,999	4,885	3,492	2,780	3,517	2,105	3,458
4,000-4,999	4,800	4,498	3,039	4,509	1,761	4,481
5,000-5,999	4,543	5,500	3,196	5,507	1,348	5,484
6,000-6,999	4,618	6,503	3,447	6,507	1,171	6,491
7,000-7,999	4,563	7,486	3,521	7,492	1,042	7,463
8,000-8,999	4,405	8,502	3,659	8,500	746	8,514
9,000-9,999	4,091	9,492	3,561	9,499	529	9,445
10,000-11,999	7,455	10,973	6,729	10,976	726	10,939
12,000-13,999	6,056	12,953	5,721	12,956	335	12,897
14,000-15,999	4,478	14,955	4,279	14,959	200	14,860
16,000-17,999	3,053	16,917	2,921	16,916	132	16,936
18,000-19,999	2,136	18,943	2,059	18,939	77	19,044
20,000-24,999	2,939	22,109	2,840	22,114	100	21,945
25,000-29,999	1,202	27,161	1,176	27,115	26	27,391
30,000-39,999	831	34,156	783	34,098	48	35,101
40,000-49,999	326	44,427	308	44,416	18	44,618
50,000-99,999	223	65,506	210	65,026	13	73,579
100,000 +	31	154,919	31	154,919	0	--
Total	73,160	9,581	54,516	11,252	18,644	4,695

a/ See footnote a, Table 1.

Table 7--Type of Tax Paid, All Family Units, Families, Unrelated Individuals

Type of Tax Paid	All Family Units		Families		Unrelated Individuals	
	Millions of Units	Percent of all Units	Millions of Units	Percent of all Units	Millions of Units	Percent of all Units
All Units	73.2	100	54.5	100	18.6	100
Income tax	51.0	70	42.5	78	8.5	46
Social Security tax	56.3	77	46.6	85	9.7	52
Neither tax	12.4	17	5.2	9	7.2	39
Both taxes a/	46.5	64	39.7	73	6.8	36
Income tax greater b/	40.1	55	34.2	63	5.8	31
Social security tax greater	6.4	9	5.5	10	1.0	5
Income tax only	4.4	6	2.7	5	1.7	9
Social security tax only	9.8	13	6.9	13	2.9	16
Income tax greater than social security tax	44.5	61	37.0	68	7.5	40
Social security tax greater than income tax	16.3	22	12.4	23	3.9	21

a/ Includes a few units for which the amounts were equal.

b/ The comparisons in this table were made using tax amounts rounded to the nearest dollar.

FOOTNOTES

*This paper was presented at the 15th General Conference of the International Association for Research in Income and Wealth, University of York, England, August 23, 1977. The author is greatly indebted to Sharon Johnson, who prepared the estimates, and to Dorothy Projector, Benjamin Bridges, John Hambor, Fritz Scheuren, Tom Petska, and Penny Johnston for their many helpful comments.

1/ A family unit is either a family (two or more persons) or an unrelated individual. See Section III for a more detailed definition.

2/ For two closely related papers, see Bridges and Johnston (1976) and Johnston and Wixon (1977). In these papers, tax liabilities are estimated directly from household survey data, without matching in other data sets.

3/ For 1972, the following types of units had these specified filing limits: (a) single person under age 65--$2,050; (b) married couple or single person age 65 or older--$2,800; (c) married couple with one spouse 65 or older--$3,550; (d) married couple with both spouses 65 or older--$4,300. See Internal Revenue Service (1974) for several exceptions to these requirements. Some returns were filed by units which were not required to file, for example, in order to obtain a refund of taxes which had been withheld.

4/ See Social Security Administration (1973) for exceptions and for other details of the program.

5/ The social security tax rates used in this paper are the sum of the rates for Old-Age and Survivors Insurance, Disability Insurance, and Hospital Insurance.

6/ If a person had both covered wage and salary income and self-employment net income, the wage and salary income was applied toward the $9,000 limit first.

7/ The Bureau of Economic Analysis of the U.S. Department of Commerce formerly published annual estimates of federal individual income tax liability by size of "family personal income," but estimates of social security tax liability were not shown (their family personal income concept is defined to be net of social security tax), and those estimates have not been available since 1964. See Fitzwilliams (1964) for the most recent estimates.

8/ For example, see Bridges (1971); Herriot and Miller (1971); Musgrave, Case, and Leonard (1974); Pechman and Okner (1974); and Reynolds and Smolensky (1974).

9/ The term "exact" match does not imply that such matches are without error--errors in the identifying information or processing errors can cause mismatches and nonmatches. Also, the identifying information is not always unique, even in the absence of errors.

[10] Later versions of the EM include a fourth set of data, SSA benefit information. Only the first three sets of data were used in this work.

[11] For more details regarding earlier versions of the EM, see Scheuren and Tyler (1975), Scheuren and Oh (1976), Scheuren et al. (1975), and other reports in the Studies from Interagency Data Linkage series.

[12] See Internal Revenue Service (1974), pp. 288-9, for a description of the SOI sample.

[13] See Budd and Radner (1969, 1975), Budd (1971), Budd, Radner and Hinrichs (1973), and Radner (1974) for descriptions of this work of varying levels of detail.

[14] See Okner (1972).

[15] See Alter (1974).

[16] See Ruggles and Ruggles (1974).

[17] See Turner and Gilliam (1975).

[18] See Armington and Odle (1975).

[19] For readers who are interested in this topic, the July 1972 and April 1974 issues of the Annals of Economic and Social Measurement contain several comments and replies and an overview paper on matching. Kadane (1975) and Wolff (1974) are somewhat more theoretical papers on statistical matching. Radner and Muller (1978) contains an overview of exact and statistical matching.

[20] The EM record would not have income tax return data if no return was filed or if the return filed was not located in the exact match. Roughly 42,000 EM records were used in the statistical match.

[21] See Appendix B for further discussion.

[22] U.S. Bureau of the Census (1973), p. 12.

[23] Ibid.

[24] For EM records with no SOI return, CPS amounts were used for all income types.

25/ The file was reweighted for several reasons, e.g., adjustment for the Decennial Census undercount and adjustment for records not matched in the exact matches by which the EM was constructed. The sample weights used, which are preliminary weights, were adjusted for consistency with SSA and Internal Revenue Service data. See Hirschberg, Yuskavage, and Scheuren (1978) for more details on the type of weight adjustments being done.

26/ The percentage of control aggregate in the Statistical Match File for each of the 8 income types is as follows: (1) 101%; (2) 89%; (3) 64%; (4) 63%; (5) 94%; (6) 72%; (7) 69%; (8) 57%. (The control aggregates were obtained from the Bureau of Economic Analysis, U.S. Department of Commerce.) The next step in this work will be to adjust the income amounts for response and reporting errors. See Budd, Radner, and Hinrichs (1973) for an example of the type of adjustment work that will be done.

27/ See Internal Revenue Service (1974) for a more detailed definition.

28/ It should be noted that the amounts of income tax reflect capital gain income (or loss), although capital gain income is excluded from the definition of income.

29/ Social security taxes were estimated by applying the appropriate tax rate to taxable earnings up to the $9,000 limit. Thus, excess tax paid by a person is excluded from these estimates.

30/ These estimates should be considered as preliminary for several reasons; the two most important reasons are mentioned here. First, work on the correction of response and reporting errors has not been completed. (See Budd and Radner (1975) for a discussion of this problem.) Second, new sample weights are being constructed for the family units.

31/ Since 1972, several changes have taken place which might be expected to affect the relationships described in this section. For example, the social security tax rate has been raised to 5.85 percent for employees and 7.9 percent for the self-employed, and the limit on taxable earnings has been raised to $16,500 for 1977. The income level above which the filing of an income tax return is required has been raised by several hundred dollars (e.g., for 1976 the limit for a single individual was $400 higher than for 1972). Also, a personal exemption credit consisting of the greater of $35 for each regular and dependent exemption, or 2 percent of taxable income (but not more than $180), was in effect for 1976.

32/ The Statistical Match File used in this paper contained 46,654 observations of family units. This number differs from the roughly 42,000 tax returns assigned in the statistical match because some family units had no returns, while some had more than one return.

33/ The definitions of income and income tax should be kept in mind, particularly when looking at low-income units. A family unit which had only capital gain income would have been in the $0-999 class, but could have had substantial income tax liability.

34/ The share of income received by the bottom quintile will probably be substantially higher after adjustment for response and reporting errors. Radner and Hinrichs (1974), after making such adjustments, estimated a share of 4.2 percent for 1971.

35/ Tabulations were made to examine the extent of reranking of units which occurred between before-tax and after-tax rankings. When all family units are ranked according to size of before-tax income and separated into 20 quantiles of five percentiles each, then reranked according to size of after-tax income and separated into 20 quantiles of five percentiles each, roughly 30 percent of the units are in different quantiles in the two rankings.

36/ At least part of this large difference between families and unrelated individuals can be explained by the different composition of the groups. For example, many unrelated individuals are retired aged persons. Based upon data from the March 1973 CPS, roughly 30 percent of all unrelated individuals were nonworkers age 65 or older.

37/ Very few family units paid the same (nonzero) amount of the two taxes.

REFERENCES

Alter, Horst E. (1974). "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey 1970." Annals of Economic and Social Measurement (April) 2: 373-394.

Armington, Catherine, and Odle, Marjorie (1975). "Creating the MERGE-70 File: Data Folding and Linking." Research on Microdata Files Based on Field Surveys and Tax Returns, Working Paper I, The Brookings Institution (June). Mimeographed.

Bridges, Benjamin, Jr. (1971). "Family Need Differences and Family Tax Burden Estimates." National Tax Journal (December) XXIV: 423-447.

Bridges, Benjamin, Jr., and Johnston, Mary P. (1976). "Estimation of Social Security Taxes on the March Current Population Survey." Studies in Income Distribution, No. 4, Office of Research and Statistics, Social Security Administration (March).

Budd, Edward C. (1971). "The Creation of a Microdata File for Estimating the Size Distribution of Income." Review of Income and Wealth (December) 17: 317-33.

Budd, Edward C., and Radner, Daniel B. (1969). "The OBE Size Distribution Series: Methods and Tentative Results for 1964." American Economic Review (May) LIX: 435-449.

Budd, Edward C., and Radner, Daniel B. (1975). "The Bureau of Economic Analysis and Current Population Survey Size Distributions: Some Comparisons for 1964," in James D. Smith, ed., The Personal Distribution of Income and Wealth, Studies in Income and Wealth, 39: 449-558.

Budd, Edward C.; Radner, Daniel B.; and Hinrichs, John C. (1973). "Size Distribution of Family Personal Income: Methodology and Estimates for 1964." Bureau of Economic Analysis Staff Paper No. 21. U.S. Department of Commerce (June).

Fitzwilliams, Jeannette M. (1964). "Size Distribution of Income in 1963." Survey of Current Business (April) 44:3-11.

Herriot, Roger A., and Miller, Herman P. (1971). "The Taxes We Pay." Conference Board Record (May) 8: 31-40.

Hirschberg, David; Yuskavage, Robert; and Scheuren, Fritz (1978). "The Impact on Personal and Family Income of Adjusting the Current Population Survey for Undercoverage." Proceedings of the 1977 Meetings of the American Statistical Association, Social Statistics Section, 70-80.

Internal Revenue Service (1974). Statistics of Income--1972, Individual Income Tax Returns. Washington, D.C.

Johnston, Mary P., and Wixon, Bernard (1977). "Payroll Tax Liability and Its Relation to Family Unit Income: 1971, 1973, 1974." Studies in Income Distribution, No. 8, Office of Research and Statistics, Social Security Administration (Forthcoming).

Kadane, Joseph B. (1975): "Statistical Problems of Merged Data Files," OTA Paper 6, Office of Tax Analysis, U.S. Treasury Department (December 12).

Musgrave, Richard A.; Case, Karl E.; and Leonard, Herman (1974). "The Distribution of Fiscal Burdens and Benefits," Public Finance Quarterly (July) 2: 259-311.

Okner, Benjamin A. (1972). "Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File." Annals of Economic and Social Measurement (July) 1: 325-42.

Pechman, Joseph A., and Okner, Benjamin A. (1974). Who Bears the Tax Burden, The Brookings Institution.

Radner, Daniel B. (1974). "The Statistical Matching of Microdata Sets: The Bureau of Economic Analysis 1964 Current Population Survey--Tax Model Match." Ph.D. dissertation, Department of Economics, Yale University. Microfilm.

Radner, Daniel B., and Hinrichs, John C. (1974). "Size Distribution of Income in 1964, 1970, and 1971." Survey of Current Business (October) 54: 19-31.

Radner, Daniel B., and Muller, Hans J. (1978). "Alternative Types of Record Matching: Costs and Benefits." Proceedings of the 1977 Meetings of the American Statistical Association, Social Statistics Section, 756-61.

Reynolds, Morgan, and Smolensky, Eugene (1974). "The Post Fisc Distribution: 1961 and 1970 Compared." National Tax Journal (December) XXVII: 515-530.

Ruggles, Nancy, and Ruggles, Richard (1974). "A Strategy for Merging and Matching Microdata Sets." Annals of Economic and Social Measurement (April) 2: 353-72.

Scheuren, Frederick J., and Oh, H. Lock (1976). "Fiddling Around with Nonmatches and Mismatches," Proceedings of the 1975 Meetings of the American Statistical Association, Social Statistics Section, 627-633.

Scheuren, Frederick J., and Tyler, Barbara (1975). "Matched Current Population Survey and Social Security Data Bases," Public Data Use (July) 3: 7-10.

Scheuren, Frederick J.; Herriot, Roger; Vogel, Linda; Vaughan, Denton; Kilss, Beth; Tyler, Barbara; Cobleigh, Cynthia; and Alvey, Wendy (1975). "Exact Match Research Using the March 1973 Current Population Survey - Initial States". Studies from Interagency Data Linkages, No. 4, Office of Research and Statistics, Social Security Administration (July).

Social Security Administration (1973). Social Security Handbook (July).

Turner, J. Scott, and Gilliam, Gary B. (1975). "Reducing and Merging Microdata Files," OTA Paper 7, Office of Tax Analysis, U.S. Treasury Department (October).

U.S. Bureau of the Census (1973). Current Population Reports, Series P-60, No. 90, "Money Income in 1972 of Families and Persons in the United States."

Wolff, Edward N. (1974). "The Goodness of Match," National Bureau of Economic Research Working Paper No. 72 (December).

APPENDIX A

The Statistical Match between the Exact Match and Augmentation Files

This statistical match was made by separating both files into comparable cell categories and using a "distance function" to choose the best match within a cell. The variables used to make the match are shown in Table A1. The first 14 variables can be considered to be common to the two files--that is, they have the same (or very nearly the same) definition and can be expected to have the same (or very nearly the same) error pattern in the two files. In other words, in an error-free exact match, values for the pair in the two files would be identical (or very nearly the same). The first ten variables in Table A1 were used as cell classifiers (see Table A2). Age and Adjusted Gross Income (AGI) were used as ranges around the EM value. The age range was the EM value plus or minus five years. For most records, the AGI range was the EM value plus or minus 10 percent, with a minimum range of $1,000. Nineteen variables (all variables except number of taxpayers, sex, and Adjusted Gross Income) were used in the distance function which was used to choose the "best" AF record among all those eligible on the basis of cells and ranges. The AF records were used with replacement.

Four levels were used in the match. In Level 1, to be eligible to be matched to a given EM record, the AF record had to have values for all ten cell classifiers which were identical to the EM record's values and be within the AGI and age ranges. Each eligible record had a distance computed. In order to be an acceptable match at this level, the distance had to be below a specified maximum; of the records with distances below that limit, the record with the smallest distance was chosen as the match. In order to have a distance below the limit, the AF record had to have values identical to the EM values for the codes showing the existence of Schedules C, E, D, and SE, and for the numbers of dependent and age and blind exemptions. Roughly 78 percent of the EM records were matched at Level 1.

If no acceptable match was found at Level 1, only the first seven cell classifiers and the AGI range were used at Level 2. At that level, the record with the smallest computed distance was acceptable; 21 percent of the records were matched at Level 2. Level 3 used only the first 5 cell classifiers, while Level 4 used only the first 3. At both levels the AGI range and the distance function were also used. The record with the smallest distance was chosen as the match.

The distance function for a given EM record was of the following general form:

$$D_k = \sum_{j=1}^{19} W_j \left[g_j (a_{jk} - e_j) \right]$$

where

D_k = distance for the kth AF record

W_j = weight applied to the jth matching pair of variables (j=1,...., 19)

a_{jk}= value of the jth matching variable for the kth AF record

e_j = value of the jth matching variable for the EM record

g_j = function which transforms the difference $(a_{jk}-e_j)$ into a distance.

A discussion of the values used for W_j and the forms used for g_j would be far too lengthy for this paper. Very briefly, the W_j reflect primarily the importance of the pair of variables in the results of the match and the comparability of the pair. (See Radner (1974) for further discussion.) The forms used for g_j were simple ones (e.g., absolute value, and square of the value).[1/]

[1/] Space does not permit a discussion of the very important and very complex question of the reliability of the results obtained from a statistical match. However, several sources of error in such matches will be mentioned here. First, because of lack of comparability between matching variables in the two sets (i.e., the variables are not defined identically and/or have different error patterns), we cannot know with certainty the values of the matching variables that we are searching for in the non-base set. Second, even if we knew those values with certainty, often we could not find a non-base set unit with such values because the non-base set is a sample. Third, even if we could find a non-base set unit with such values (assuming it is not the true match), the values for nonmatching variables in the non-base set would probably differ from the true values because those nonmatching variables are not "completely explained" by the matching variables.

Table A1--Variables Used in the Statistical Match

	Variable	EM Source of Data a/	AF Source of Data
1.	Number of Taxpayers b/	IRS	IRS
2.	Sex b/	SSA	SSA
3.	Race	SSA	SSA
4.	Marital Status	IRS	IRS
5.	Number of Dependent Exemptions	IRS	IRS
6.	Type of Earnings	SSA	SSA
7.	Size of Earnings	SSA	SSA
8.	Wage and Salary Income	IRS	IRS
9.	Dividend Income (after exclusion)	IRS	IRS
10.	Interest Income	IRS	IRS
11.	Age	SSA	SSA
12.	Adjusted Gross Income b/	IRS	IRS
13.	Net Adjusted Gross Income c/	IRS	IRS
14.	Number of Age and Blind Exemptions	IRS	IRS
15.	Existence of Schedule C (nonfarm business income)	IRS	IRS
16.	Existence of Schedule E (supplemental income)	IRS	IRS
17.	Existence of Schedule D (capital gain or loss)	IRS	IRS
18.	Existence of Schedule SE (self-employment income)	IRS	IRS
19.	Existence of Schedule F (farm income)	IRS	IRS
20.	Existence of Rent and/or Royalty Income	CPS	IRS
21.	Existence of Pension Income	CPS	IRS
22.	Home Ownership	CPS	IRS

a/ IRS = Internal Revenue Service
SSA = Social Security Administration
CPS = Current Population Survey

b/ Not used in the distance function.

c/ Defined as Adjusted Gross Income minus $750 times the total number of exemptions.

Table A2--Cell Categories Used in the Statistical Match

Variable	Cell Categories	Levels at which Cell Categories Were Used
1. Number of Taxpayers	a. One b. Two	1,2,3,4
2. Sex	a. Male b. Female	1,2,3,4
3. Race	a. Black b. White c. Other	1,2,3,4 a/
4. Marital Status	For records with 1 taxpayer: a. Separate return with 1 taxpayer exemption b. Surviving spouse return c. Head of household return d. Single return	1,2,3
	For records with 2 taxpayers: a. Joint return b. Separate return with 2 taxpayer exemptions	1,2,3
5. Number of Dependent Exemptions	For records with 1 taxpayer: a. None b. One or more	1,2,3
	For records with 2 taxpayers: a. None b. One c. Two d. Three e. Four or more	1,2,3
6. Type of Earnings (SSA)	a. None b. Wage and Salary only c. Self-employment only d. Both Wage and Salary and Self-employment	1,2

7.	Size of Earnings (SSA)	a. $0	1,2
		b. $1-8,999	
		c. $9,000	
		d. $9,001 or more	
8.	Wage and Salary Income	a. Zero	1
		b. Nonzero	
9.	Dividend Income (after exclusion)	a. Zero	1
		b. Nonzero	
10.	Interest Income	a. Zero	1
		b. Nonzero	

a/ At Level 4, the "White" and "Other" categories were combined.

APPENDIX B

A Suggested Framework for Statistical Matching 1/

In this appendix a brief summary of the theoretical steps involved in a statistical match will be followed by a somewhat more detailed discussion of those steps. An example involving household survey and income tax data will be used to clarify the concepts as the discussion proceeds.

In summarizing the matching steps, we begin with a universe, "U," for which we want to make estimates of variables and their relationships to each other. We have two microdata sets, "A" and "B," samples which provide observations on the universe; each set contains some variables which are not included in the other set. We then define a hypothetical exact match result which we want the statistical match to approximate. However, we do not know the hypothetical exact match result; therefore we estimate it, either explicitly or implicitly, using whatever information is available. The appropriate matched pairs of units are then chosen in a way which minimizes deviations from the estimate of the exact match result.

Universe

We will begin the detailed discussion of the framework by considering the universe U for which we want to estimate various relationships. U consists of a set of N units; for each unit there are values for R variables. By definition all information in U is error-free, and it is assumed that all information relevant to the estimates we want to make is contained in the R variables. U can be represented by an N x R matrix in which each of the N rows contains the values of the R variables for one unit.

Two Data Sets

We will assume that we have two microdata sets of observations on variables for units in U; these sets, A and B, are the sets we want to match statistically. A and B will be assumed to be samples from U. A contains n units, while B contains r units, where both n and r are less than N; r does not necessarily equal n. It will also be assumed that very few units from U are represented in both A and B; A and B could be independent samples for which n/N and r/N are small. For example, set A might be the persons interviewed in a household sample survey for a given year, and set B might be a sample of income tax returns for that same year.

It will be assumed that A contains observations on k variables, while B contains observations on m variables. By assumption, both k and m are less than R, and all of the variables are contained in U. Some variables from U may be contained in both A and B, while at least some will be contained in only one set.

The i^{th} unit in A, which will be denoted A_i, contains k observed variables, as shown below:

$$A_i = (a_{i1}\ a_{i2}\ \dots\ a_{ik})$$

Similarly, the i^{th} unit in B contains m observed variables:

$$B_i = (b_{i1}\ b_{i2}\ \dots b_{im})$$

It will be assumed that at least some of the variables in A and B can contain errors, while in U they do not. Because of different error components, a variable from U which appears in both A and B can have different values in the two sets for the same underlying unit in U. For example, even if wage income were defined identically in the household survey and the tax return, the survey response might differ from the amount shown on the tax return.

Hypothetical Exact Match

At this point we have defined the universe and the two data sets which will be matched statistically. We will now define "C," a hypothetical data set which represents the result of an exact match between A and B, if the underlying units represented in A were also represented in B. The set C is hypothetical because that exact match cannot be carried out. The exact match is impossible because very few of the units represented in A are also represented in B. By assumption C contains all k variables from A and all m variables from B, including their error terms. Because a statistical match is an approximation of an exact match, C is the data set which we try to approximate when we perform a statistical match. 2/

For the i^{th} unit in A, the information in C will be denoted C_i, and can be expressed as follows:

$$C_i = (a_{i1}\ a_{i2}\ \dots\ a_{ik}\quad b^*_{i1}\ b^*_{i2}\ \dots\ b^*_{im})$$

Using the previously mentioned example, C_i contains the survey response given by A_i and the data from the tax return filed by A_i . As noted above, that tax return does not appear in B, except in rare cases.

Estimate of C

When we actually want to make a match, we do not know C (i.e., we do not know the b^*_{ij}). We therefore make (either explicitly or implicitly, depending upon the matching method) an estimate of C, called "L," using whatever information is available. It is not necessary for all B variables to be estimated in L. Estimated values can be obtained by assumption. For example, for a given A unit, it might be assumed that the value for a given B variable should be equal to the value for a given A variable (say, $a_{11} = b*_{11}$). We could say that wage income in B should be identical to wage income in A. This would be valid if wage income were defined identically and had an identical error pattern in A and B, which ordinarily is not true. Estimated values can also be obtained by other means, for example, by regression techniques or by using cross-tabulations from an exact match between sets similar to A and B.

For the i^{th} unit in A, the information in L will be denoted L_i, and can be expressed as follows:

$$L_i = (a_{i1}\ a_{i2}\ \ldots\ a_{ik}\ \ \hat{b}*_{i1}\ \hat{b}*_{i2}\ \ldots\ \hat{b}*_{im})$$

Using the continuing example, for each unit in A, L contains that unit's survey response data and estimates of some or all of the variables in the tax return filed by that A unit. 3/

Statistical Match Result

We will now introduce "M," the result of statistically matching sets A and B in some unspecified way. For the i^{th} unit in A, the information in M will be denoted M_i, and can be expressed as follows:

$$M_i = (a_{i1}\ a_{i2}\ \ldots\ a_{ik}\ b^o_{i1}\ b^o_{i2}\ \ldots\ b^o_{im})$$

In our example, for each unit in A, M contains that unit's survey response data and the tax return data from the B unit assigned to that A unit in the statistical match. 4/

Not every B unit has to be used in the match solution, and some B units can be used more than once in the solution. 5/ It follows from the definition of a statistical match that the m variables assigned to a given A unit in the match are all from one B unit.

In making a statistical match we choose among alternative solutions; each alternative solution is characterized by the particular set of B units assigned and the particular A unit(s) to which each is assigned. We choose the solution in which M approximates L as closely as possible, in terms of the variables and relationships of greatest importance in the results of the match. This approximation can be viewed in terms of a "distance function." We can define in general terms a distance function, "D," which measures the distance (D_M) of M from L. The distance D_M is defined in a subjective way according to the purpose of the match. Thus,

$$D_M = D(M, L, P)$$

where P denotes the purpose of the match. The statistical match solution which minimizes D_M is the optimal match result. 6/

FOOTNOTES

1/ See Radner (1974) for a more detailed discussion of this topic.

2/ It is important to note that C is not unique. The form of C depends upon which data set, A or B, is taken as the base. We are assuming that A is the base set. We are also assuming that the hypothetical exact match by which C was constructed was carried out without error.

3/ L can also include constructed variables for either set, or both.

4/ We are assuming that only one B unit is assigned to each A unit in the match. In some statistical matches more than one B unit was assigned to some or all A units, and sample weights were adjusted to account for this "splitting." See Budd, Radner, and Hinrichs (1973) and Turner and Gilliam (1975) for examples of such procedures.

5/ Some matching procedures do require every B unit to be used in the match solution, and used with its before-match sample weight. For example, see Radner (1974) and Turner and Gilliam (1975).

6/ This is not meant to suggest that any given statistical match should be carried out using a distance function, or that using a distance function is the best way of matching in theory.