NUMBER 18

AN EXAMPLE OF THE USE OF STATISTICAL MATCHING IN THE ESTIMATION AND ANALYSIS OF THE SIZE DISTRIBUTION OF INCOME*

Daniel B. Radner

Division of Economic Research

OCTOBER 1980

Social Security Administration
Office of Policy
Office of Research and Statistics

*The author is greatly indebted to Sharon Johnson, who prepared the estimates, and to Benjamin Bridges, Thomas Petska, Sharon Johnson, and Peter Petri for their helpful comments.

I. INTRODUCTION

This paper discusses the use of statistical matching in the estimation and analysis of the size distribution of family unit personal income. Statistical matching is a relatively new technique which has been used 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. 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. In contrast, in an "exact" match, information for the same person from two or more files is brought together using personal identifying information.

The paper begins with a brief discussion of data on the size distribution of income in the U.S. and their limitations. Several methods of improving or augmenting those data are described, and earlier examples of statistical matching for that purpose are mentioned. A brief summary of the types of statistical matching methods which have been used is also presented. Then a recent example of statistical matching carried out at the Office of Research and Statistics, Social Security Administration, with the cooperation of the Bureau of Economic Analysis, U.S. Department of Commerce, is described, and the effects on the size distribution of adjusting and augmenting the initial data using the statistically matched data from that example are shown. Material relating to the accuracy of that statistical match is presented in the appendix.

II. DATA ON THE SIZE DISTRIBUTION OF INCOME IN THE U.S.

The focus of this paper is on the size distribution of annual income of family units in the U.S. The two major annual sources of data of this type which utilize at least fairly comprehensive definitions of income are the Current Population Survey (CPS) of the Bureau of the Census and the Statistics of Income (SOI) of the Internal Revenue Service. The CPS data are based upon a household sample survey (e.g., U.S. Bureau of the Census 1980). Information on recipiency of many different income types is obtained; amounts are obtained for fewer types or combinations of types. Published distributions are shown primarily using total money income, but microdata files which allow different (less comprehensive) definitions of income can be obtained. Data on a family unit basis are available by several types of demographic classifications, both in published form and on microdata files.

The SOI is a stratified sample of unaudited Federal individual income tax returns (e.g., Internal Revenue Service 1980). The sample is heavily weighted toward high-income and business returns. Most published distributions are shown using adjusted gross income, but microdata files which permit other definitions of income can be obtained.

Recent other sources of data on the size distribution of annual income which are not available annually include the 1976 Survey of Income and Education (U.S. Bureau of the Census 1978), the 1972-73 Consumer Expenditure Survey (Bureau of Labor Statistics 1978), and the 1970 Decennial Census (U.S. Bureau of the Census 1973a).

The Bureau of Economic Analysis (BEA) of the U.S. Department of Commerce formerly published an annual series on the size distribution of family personal income (e.g., Fitzwilliams 1964), but only one set of those estimates has been published since 1964 (Radner and Hinrichs 1974). Although work on resuming that series is continuing, it is difficult to say when an annual series will be available. The BEA work involves combining data from several sources in an attempt to build a more accurate estimate of the size distribution.[1/]

Both major annual sources of data suffer from serious limitations. The CPS regularly collects data only on cash income before tax in its March interview. Thus, noncash income and tax liabilities are not collected. Also, the CPS, like most household surveys, suffers from serious problems of misreporting and nonreporting of income. For example, for 1978 the CPS shows about 90 percent of total money income as estimated in independent control aggregates, and less than 50 percent of interest, dividends, and workmen's compensation (U.S. Bureau of the Census 1980, p. 297). A substantial proportion of the CPS aggregate (about 20 percent in 1978) consists of amounts assigned to persons from whom responses were not obtained. Another limitation is that the CPS sample contains few high-income observations.

The SOI data also have several limitations. One major problem is that many persons are omitted because they do not file tax returns. Also, nontaxable income types are omitted for all persons,

and the demographic information included is very limited. In addition, the unit used is the tax unit, and family units cannot be constructed. Misreporting of income is also a problem for some income types, particularly those in which income is net of expenses.[2/]

III. IMPROVING THE DATA

Improving data collection in existing surveys and mounting new, better surveys are obvious examples of ways to improve the data. However, we will confine the discussion here to methods of improvement which utilize sources in which the data have already been collected. Some methods utilize distribution data from only one data source. Procedures such as assignment of amounts to nonrespondents in surveys and reweighting to make the population data conform to independent control totals are often applied, usually before the data are made generally available (e.g., the CPS). Additional reweighting or adjustment of amounts can also be applied to make the data conform to independent control totals, such as income aggregates. Also, other variables can be added by various imputation techniques (e.g., regressions).

Another way to improve the data is to combine two or more sources of distribution data. For example, if the CPS and SOI data could be combined and the strong points of each could be used, a better estimate of the size distribution could be produced. The CPS collects data which are on a family unit basis and which include nonfilers, many nontaxable income types, and much demographic information. The SOI contains data on tax liabilities and more accurate data for several income types, as well as more high-income observations.

Sources of distribution data can be combined either on an aggregated basis or on a record-by-record basis. In the pre-1964 BEA series, different data sources were combined on an income-size-class basis to produce a "corrected" distribution which incorporated the strong points of each data source (Goldsmith 1958). However, matching on a record-by-record basis allows much greater detail and much greater flexibility in the use of the matched results than can be obtained by combining on an aggregated basis. For example, the best estimates of some income types might be used from one data source, while the best estimates of other income types might be used from other data sources in the matched record.[3/]

As noted earlier, there are two basic types of matches, exact and statistical. Exact matching has been used extensively to assess the accuracy of income data. Matching of survey data with other survey data and with administrative record data and matching of administrative data with other administrative data have been used. For example, the income data from the 1950, 1960, and 1970 Decennial Censuses have been assessed in this way (e.g., U.S. Bureau of the Census 1970). Recent examples of exact matching used to assess the accuracy of income data include the 1963 Pilot Link Study (Scheuren, Bridges, and Kilss 1973), and the 1973 Current Population Survey--Internal Revenue Service--Social Security Administration (CPS-IRS-SSA) Exact Match Study.[4/]

Although exact matching has been used to assess the accuracy of income data, it has rarely been used directly to correct

estimates of the annual size distribution of income. That is, more accurate amounts have rarely been used to replace less accurate amounts in an exact match file. One attempt along those lines was never completed (Steinberg 1973). Mean incomes and the composition of total money income by type have been presented from the 1973 CPS-IRS-SSA Exact Match (Radner 1978).[5/]

Exact matching has also been used to add other variables to ("augment") a given data source. The 1963 Pilot Link Study and the 1973 CPS-IRS-SSA Exact Match Study are examples of that use.[6/]

IV. STATISTICAL MATCHING AS A METHOD OF IMPROVING THE DATA

In contrast to exact matching, statistical matching has been used in several attempts to construct more accurate, more detailed, and/or more comprehensive income size distribution data. More accurate income data have been added to a file, data items not present in a file have been added, and more observations have been added. Several examples of statistical matching for this purpose are mentioned below.[7/]

The first example was in connection with the construction of "corrected" estimates of the size distribution of family personal income at BEA (Budd and Radner 1969, 1975; Budd 1971; Budd, Radner, and Hinrichs 1973). As noted above, in earlier work BEA had combined data sources on an income-size-class level. Combining microdata files containing income data on a record-by-record basis through matching was a logical next step at BEA as microdata files and modern computers became available. Because exact matching could not be used, statistical matching was applied.[8/] In the BEA

work, the March 1965 CPS file was statistically matched with a sample of 1964 Federal individual income tax returns. That matched file was then statistically matched with the 1972 Survey of Financial Characteristics of Consumers (SFCC). The first of those matches emphasized the "correction" of CPS income amounts and the addition of more high-income observations, while the second match was primarily for the purpose of adding several SFCC variables which were used to assign amounts of noncash income types. The BEA work can also be viewed as one step in the construction of a microdata file which was consistent with and nested within the personal income aggregate estimates from the National Income and Product Accounts (Ruggles, Ruggles, and Wolff 1977).

Other early work took place at the Brookings Institution in connection with analysis of the tax system (Okner 1972). Brookings was interested in putting a sample of tax returns on a family unit basis and adding information for nonfilers and for nontaxable income types. That Brookings match was between the 1967 Survey of Economic Opportunity (SEO)and the 1966 Internal Revenue Service Tax File of individual Federal income tax returns. The match was one step in the donstruction of a corrected and more detailed microdata base for policy analysis, particularly tax policy analysis. The match was intended to provide both correction and addition of variables and observations to the SEO data. A later match between the March 1971 CPS and the 1970 Tax Model was also performed at Brookings (Armington and Odle 1973).

A statistical match between the 1970 Canadian Survey of Consumer Finances and the 1970 Family Expenditure Survey was carried out at Statistics Canada in connection with work on the measurement and comparison of relative distributions of income for several countries (Alter 1974). Neither survey contained all of the information needed for the desired definition of income. Addition of variables was the purpose of this match.

In work closely related to the size distribution of income, a file produced by three statistical matches has been used to estimate and analyze the size distribution of household wealth in the U.S. (Wolff 1977; Ruggles and Ruggles 1974; Ruggles, Ruggles, and Wolff 1977). The basic match was between the 1969 Internal Revenue Service Tax Model and the 1970 Decennial Census Public Use Sample 15 percent file. Other matches were between the 1969 and 1970 Tax Models and between the 5 percent and 15 percent Public Use Samples. Addition of variables was the purpose of these matches.

Several statistical matches have been carried out at the Office of Tax Analysis of the U.S. Treasury Department in connection with analysis of the tax system. The files matched include the 1973 SOI and CPS files, the 1975 SOI and 1976 Survey of Income and Education files, and the 1977 SOI and 1978 CPS files (Barr and Turner 1978, 1980). The purpose was the addition of nonfilers and variables to the tax return samples.

Mathematica Policy Research has performed several matches related to the size distribution of income in connection with policy analysis. Completed work includes matches between a 1970 Decennial

Census Public Use Sample and the 1973 Aid to Families with Dependent Children Survey (Springs and Beebout 1976) and between the March 1975 CPS and the Survey of Household Characteristics (Beebout, Doyle, and Kendall 1976). Addition of variables was the purpose of these matches.[9/]

V. STATISTICAL MATCHING METHODS[10/]

At this point it will be useful to summarize types of statistical matching methods which have been used. Most statistical matches have been between a "base" file, which remained essentially unchanged in the match, and a "nonbase" file which was matched to the base file.

Many different statistical matching methods have been used. In most cases the variables in both files were separated into "matching variables" (which were similar in the two files and were used to carry out the match) and "nonmatching variables" (which were "added" variables). Values of matching variables sometimes were adjusted to take account of noncomparabilities, e.g., differences in definition and/or differential reporting errors in the two files.

In most matches both files were separated into comparable subsets of units. Within each subset, rules were specified for the choice of a nonbase file record (or records) to be assigned to each base file record. The selection of the record within the subset often was based upon a distance function by which a distance was computed between a given base file record and each potential

match in the nonbase file, using differences between values of matching variables in the two files. In some cases, these differences were weighted according to the relative importance of the matching variables as explainers of important nonmatching variables or the relative importance and the comparability of the pairs of matching variables. The potential match with the smallest distance ordinarily has been chosen as the match; a maximum distance has been used to define a subset of potential matches from which a random choice was made. In some cases, subsets were defined so narrowly that most subsets contained only one record. In other cases, the choice within subsets was random.

Statistical matches have been separated into two basic types, constrained and unconstrained, according to what restrictions, if any, are placed on the use of nonbase set records. In a constrained match, every nonbase set record appears in the matched result and has a sample weight identical (or very close) to its sample weight before matching. In an unconstrained match there is no such restriction on the nonbase set records. A constrained match can be viewed as choosing nonbase set records without replacement (until all nonbase set records are used), while an unconstrained match can be viewed as choosing with replacement.[11/]

VI. THE EXACT MATCH FILE--AUGMENTATION FILE STATISTICAL MATCH

The statistical match which is described here was between the 1973 CPS-IRS-SSA Exact Match (EM) file and the Augmentation File (AF), which contained detailed Federal individual income tax return

information. The EM was the base file in this match. This statistical match was the first step in the construction of a microdata file in which the income data are being adjusted to be consistent with independent recipient and aggregate control totals and which contains data on tax liabilities. The matched file is being used to examine the role of social security in the tax-transfer system. Some preliminary analyses have already been carried out with an early version of the statistically matched file (Radner 1978, 1979a).

A. Files Matched

The EM file was constructed in a joint project by the Social Security Administration (SSA) and the Bureau of the Census (Kilss and Scheuren 1978). The EM sample was based on the March 1973 CPS; that file contained roughly 50,000 households. Persons age 14 and over in the March 1973 CPS had their survey data exactly matched with their SSA earnings (SER) and benefit administrative records and with selected items from their 1972 Federal individual income tax returns. All EM records which had good CPS-SER exact matches and for which a tax return had been found were used in the initial match; there were 42,293 such records.[12/]

Although the EM is an extremely valuable file for many purposes, it has several limitations for use in research on the size distribution of income. First, the EM contains no data on income tax liabilities; only a few tax return items are included.[13/] Second, some of the CPS income information in the EM suffers from serious response errors. Third, the EM sample contains few high-income

observations. This statistical match was designed to produce a matched file which was improved in all of these areas.

The starting point for the construction of the AF was a sample of roughly 106,000 Federal individual income tax returns chosen by subsampling the 1972 Statistics of Income (SOI) sample (Internal Revenue Service 1974). A subsample of the SOI was used because processing of the entire file would have been too expensive. Subsampling rates differed among the various strata; relatively more high-income and business returns were eliminated. However, even after subsampling, high-income returns were over-sampled.

The next step in the construction of the AF was an exact match between the SOI subsample and SSA's SER file which contained earnings and demographic data. The SER information was added primarily to improve the quality of the EM-AF statistical match by adding more good matching variables.[14/] The file was then modified for use in the statistical match. A small number of records with missing or invalid values for important variables was eliminated, the file was subsampled and reweighted so that sample weights varied only with size of the absolute value of AGI, and, where possible, returns filed by persons outside the EM universe were eliminated (e.g., some military personnel). The version of the AF used in the statistical match contained 95,159 records.

For both the EM and the AF, the basic unit used in the match was the tax filing unit. However, due to data limitations, in the case of a joint return, the SER data used were only for the principal taxpayer.

B. Choice of Matching Methods

The EM-AF statistical match consisted of three parts, each of which was a statistical match: the initial match, the rematch; and the high-income match. The initial match and rematch were basically similar matches which focused on adding tax liabilities and more accurate income data to the EM. The high-income match was a different type of match which focused on adding more high-income AF returns to the statistically matched file.

The statistical matching methods chosen in the EM-AF match were influenced by several factors. One constraint was the amount and type of computer resources available. A second important influence was the purpose of the match--the file which would result from this match was expected to be used for at least several different purposes, some of which could not be specified when the match was carried out. A third important influence was the characteristics of the files being matched. Many good matching variables were available, and the final sample weights and exact match rules for the EM were not available when the matching was carried out. A fourth influence was that there was a desire to learn more about the accuracy of statistical matching procedures.

An unconstrained method was chosen for the initial match and rematch for two principal reasons. First, an unconstrained method was expected to be much less expensive than a constrained method. Second, we wanted to see how accurate an unconstrained match could be under favorable circumstances--a large number of good matching variables and a relatively large number of records in the nonbase

file. The information obtained would be used to help determine whether we would do statistical matches in the future. A constrained method was chosen for the high-income match because we wanted to preserve the AF data for high-income returns.[15/]

C. Matching Data Used

In choosing the variables to be used in carrying out the match, it was necessary to select pairs of matching variables, to determine the comparability of the variables in each pair, and to determine the relative importance of each pair.

The matching variables used are shown in table 1. Three basic sets of variables were used: (1) SER variables which appeared in both files; (2) IRS variables which appeared in both files; (3) roughly comparable CPS (EM) and SOI (AF) variables. The variables in group (1) were considered to be "identical." That is, their response and processing error patterns, as well as their definitions, were assumed to be the same. Thus, in an exact match carried out without error, the values in the two files would be the same. The variables in group (2) were considered to be very similar, but in general not "identical." The variables in group (3) were not very similar.

The relative importance of a matching pair depended upon its usefulness in explaining nonmatching variables of interest and its own usefulness in the results.[16/] The pair's own usefulness was important because, for some purposes, the AF data would be used as an entity (e.g., all tax return items would be taken from the

Table 1.-- Pairs of Matching Variables Used in the Match

Variable Pair	EM Source of Data a/	AF Source of Data
1. Number of Taxpayers	IRS	IRS
2. Sex	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	IRS	IRS
13. Net Adjusted Gross Income b/	IRS	IRS
14. Number of Age and Blind Exemptions	IRS	IRS
15. Presence of Schedule C (nonfarm business income)	IRS	IRS
16. Presence of Schedule E (supplemental income)	IRS	IRS
17. Presence of Schedule D (capital gain or loss)	IRS	IRS
18. Presence of Schedule SE (self-employment income)	IRS	IRS
19. Presence of Schedule F (farm income)	IRS	IRS
20. Presence of Rent and/or Royalty Income	CPS	IRS
21. Presence of Pension Income	CPS	IRS
22. Home Ownership	CPS	IRS

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

b/ Defined as adjusted gross income minus $750 times the total number of exemptions.

SOI). The amount of Federal individual income tax liability perhaps was the single most important AF nonmatching variable being added. Many of the IRS matching variables were expected to be useful in adding the amounts of tax liability. The SER variables provided demographic categories and information about social security coverage of earnings.

D. Initial Match

A brief summary of the matching procedure is followed by descriptions of the cells, ranges, distance function, reference distance, and pseudo-cells used. In the initial match procedure, for each EM record, a set of cell categories and acceptable ranges of adjusted gross income (AGI) and age were defined. For each AF record in those cell categories and within the AGI and age ranges (with some exceptions), a distance between the EM record and that AF record was computed using a distance function. The AF record with the smallest distance was chosen as the tentative match. If that distance was below a specified maximum (the reference distance), then that AF record was the final match for that EM record (Level 1). If there was no final match, then several cells were collapsed and the age range and maximum distance were eliminated (Level 2). Further collapsing of cells was necessary in some cases (Levels 3 and 4). A few records still unmatched after those steps were matched after their AGI ranges were expanded. Each EM record was matched at the earliest level possible. AF records were used with replacement. The various parts of the procedure are described below.

1. Cells

In general the cells used were based upon pairs of SER and IRS categorical variables which were considered to be very important in the match. The cells used at each level are shown in table 2. At Level 1 there were 993 cells which contained at least one EM record, at Level 2 there were 360 cells, at Level 3 there were 82 cells, and at Level 4 there were eight cells.

2. AGI and Age Ranges

The AGI range was used at all levels. The absolute size of the AGI range depended upon the size of AGI of the EM record. The relationship between the size of AGI and the size of the range is shown in table 3.

For Level 1 only, an age range consisting of the EM age plus or minus five years was used. AF records outside that range were not eligible for matching at Level 1. No age range was used at the other levels.

3. Distance Function

The distance function was used to select the AF record which fit each EM record best, given the cell categories and ranges.[17/] The distance function used for the i^{th} EM record had the following form:

$$D_{ij} = \sum_{k=1}^{m} W_k[g_k(a_{jk}-e_{ik})]$$

where

D_{ij} = the distance between the i^{th} EM record and the j^{th} AF record.

Table 2---Cell Categories Used in the Initial 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

Table 2 (continued)

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

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

Table 3.--AGI Ranges in the Initial Match

Size of EM AGI	Top of AGI Range	Bottom of AGI Range	Width of AGI Range [a/]
-$5,000 or less	90% of EM AGI	110% of EM AGI	20% of |EM AGI|
-$4,999 to -$501	EM AGI + $500	EM AGI - $500	$1,000
-$500 to -$1	-$1	EM AGI - $500	-EM AGI + $500
Zero	Zero	-$1,000	$1,000
$1 to $500	EM AGI + $500	$1	EM AGI + $500
$501 to $5,000	EM AGI + $500	EM AGI - $500	$1,000
$5,001 or more	110% of EM AGI	90% of EM AGI	20% of EM AGI

a/ The widths shown here are the widths used in the computation of the standardized AGI range used in the distance function. In some cases the widths shown here differ by $1 from the actual range used in defining eligibility of AF records for matching.

$g_k(a_{jk}-e_{ik})$ = the distance between the values of the k^{th} pair of variables in the i^{th} EM record and the j^{th} AF record (k=1,...,m).

a_{jk} = the value of the variable in the k^{th} pair in the j^{th} record in the AF.

e_{ik} = the value of the variable in the k^{th} pair in the i^{th} record in the EM.

g_k = the function which transformed differences between values into distances for the k^{th} pair.

W_k = the weight applied to distances for the k^{th} pair.

Nineteen pairs of variables were included in the distance function. Those variables are shown in table 4. Number of taxpayers, sex, and AGI were the only matching variables not used in the distance function--number of taxpayers and sex because they were always used as cell classifiers and AGI because it was replaced by net AGI in the distance function.

There were four different forms used for g_k, as shown in table 4. The form used depended upon which pair of variables was being considered. The "0-1" form was simply a distance of zero if the values were equal and one if the values were unequal. This form was used where the size of differences between the values was not considered to be meaningful. The "absolute value" form was the absolute value of the difference between values. In this case the desired effect was proportional to the size of the differences. The "square" form was the square of the difference between values. This form was used where larger differences were desired to have a more than proportional effect on the match compared to small differences. The "SAR" form was the absolute value of

Table 4.--W_k and Forms of g_k Used in the Initial Match

Variable Pair	W_k	Form of g_k a/
Race	10,000	0-1
Marital Status	10,000	0-1
Number of Dependent Exemptions	10,000	Absolute Value
Type of SSA Earnings	10,000	0-1
Size of SSA Earnings	1	SAR
Wage and Salary Income	1	SAR
Dividend Income (after exclusion)	25	SAR
Interest Income	1	SAR
Age	5	Square
Net Adjusted Gross Income	1	SAR
Number of Age and Blind Exemptions	10,000	Absolute Value
Presence of Schedule C	10,000	0-1
Presence of Schedule E	10,000	0-1
Presence of Schedule D	10,000	0-1
Presence of Schedule SE	10,000	0-1
Presence of Schedule F	50	Square
Presence of Rent and/or Royalty Income	30	Square
Presence of Pension Income	40	Square
Home Ownership	25	Square

a/ The forms of g_k are defined as:

Form	Value of $g_k(a_{jk}-e_{ik})$
0-1	$\begin{cases} 0 \text{ if } a_{jk} = e_{ik} \\ 1 \text{ if } a_{jk} \neq e_{ik} \end{cases}$
Absolute Value	$\lvert a_{jk} - e_{ik} \rvert$
Square	$(a_{jk} - e_{ik})^2$
SAR	$\lvert a_{jk} - e_{ik} \rvert \div \frac{\text{Width of EM AGI Range}}{1{,}000}$

the difference between values divided by the "standardized AGI range" or SAR. SAR was defined to be the width of the EM record's AGI range in thousands of dollars. Income amount differences were scaled by dividing by the SAR. The form used for each pair of variables is shown in table 4.

One weight was applied to each pair of variables in the distance function; that is, the weight did not vary among EM records. The higher the weight, the closer the matched values for that pair of variables would be expected to be, ceteris paribus.

A tentative set of weights was specified initially. Each tentative weight reflected the comparability, the importance, and the scale of the pair. The more comparable and the more important the pair of variables was, the higher the weight was. It was necessary to adjust for scale so that some variables would not "overwhelm" other variables in the distance function. The tentative weights were modified as a result of testing using subsamples from the EM file.[18/] The final weights used are shown in table 4.

4. Reference Distance and Pseudo-cells

The reference distance was the distance all Level 1 matches had to be below; 10,000 was used as the reference distance. The testing mentioned above gave rise to the use of what we have called "pseudo-cells," categories which were not treated as cells in the computer program, but which operated as Level 1 cells in the match. Presence of Schedules C, D, E and SE and the exact number of dependent and age plus blind exemptions were used as pseudo-cells. Each of those pairs of variables

was given a weight of 10,000 in the distance function. Thus, any difference between EM and AF values implied that the distance could not be below the reference distance of 10,000. These variables were chosen as pseudo-cell variables because in the testing the EM and AF values disagreed too often.

5. Processing

More than 77 percent of the EM records were matched at Level 1, and more than 21 percent of the EM records were matched at Level 2. Less than two percent of the EM records were matched at Levels 3 and 4. The matching through Level 4 with the AGI ranges shown in table 3 matched 42,278 records. The remaining 15 records were matched after the AGI ranges were expanded.

E. Rematch

Part of the EM file was rematched with the AF because we were not fully satisfied with the results of the initial match. The dissatisfaction was primarily with the underestimates of numbers of recipients and aggregate amounts for several income types in the AF (see appendix, table A-4). These underestimates were also considered to be a problem by BEA, which became closely involved in the work.[19/] The principal differences between the initial match and the rematch were that the presence of several income types was given a larger role in the rematch and a much simpler distance function was used in the rematch.

EM records which were considered to have an inconsistent initial match were rematched.[20/] A match was inconsistent if there was a

discrepancy between EM and AF information for presence of Schedules C, D, E, SE, or F, presence of wages and salaries, interest, dividends in AGI, or social security taxable earnings.[21/] A total of 6,861 EM records (about 16 percent of the EM) were rematched.

The rematching was carried out using cell categories, an AGI range, and a distance function. A total of eight levels was used. At each level, cell categories and an AGI range above and below the EM AGI amount were defined; matching at that level had to take place within those cell categories and range. The same distance function was used at each level; the distance was defined as the absolute value of the difference between EM AGI and AF AGI. The cell categories and AGI range used at each level are shown in table 5.

An EM record was considered to be rematched when at least one AF record within the relevant cell categories and the relevant AGI range was found. If more than one such AF record was found, the AF record with the smallest distance was chosen. AF records were used with replacement. More than 63 percent of the EM records were matched at Level 1, and more than 96 percent were matched in the first four levels. For the EM records used in the rematch, matches from the rematch replaced matches from the initial match.

F. High-Income Match

Because the EM sample is not stratified by size of income, that sample contains few high-income records. Thus, estimates from that sample contain large sampling errors for high-income groups and for aggregates of items which are concentrated in the high-income groups. In the initial match and the rematch, only one AF record was used for

Table 5.--Cell Categories and AGI Ranges Used in the Rematch

Cell Categories or AGI Range Used in the Rematch	Levels in Which the Cell Category or Range Was Used
Schedule C	1, 2, 3, 4, 5, 6, 7, 8
Present	
Absent	
Schedule E	1, 2, 3, 4, 5, 6, 7, 8
Present	
Absent	
Type of Return	1, 2, 3, 4, 5, 6, 7 a/
Joint	
Nonjoint	
Schedule SE	1, 2, 3, 4, 5, 6, 7 a/
Present	
Absent	
Schedule F	1, 2, 3, 4, 5, 6 a/, 7 a/
Present	
Absent	
Age of Taxpayer	1, 2
Joint Returns	
Less than 35	
35-64	
65 and over	
Nonjoint Returns	
Less than 45	
45-64	
65 and over	
Age of Taxpayer (recoded)	3, 4
Less than 65	
65 and over	
Sign of AGI	1, 2, 3, 4
Positive	
Zero	
Negative	
Sample Weight Class (Absolute Value of Size of AGI)	1, 2, 3, 4
Less than $10,000	
$10,000-$14,999	
$15,000-$19,999	
$20,000-$49,999	
$50,000-$99,999	
$100,000-$199,999	
$200,000-$499,999	
$500,000-$999,999	
$1,000,000 and over	
Sex of Taxpayer	1, 2, 3
Joint Return	
Nonjoint Return, Male	
Nonjoint Return, Female	
Wages and Salaries	1, 2, 3
Zero	
Nonzero	

Table 5.--Cell Categories and AGI Ranges Used in the Rematch--Continued

Cell Categories or AGI Range Used in the Rematch	Levels in Which the Cell Category or Range Was Used
Interest	1, 2, 3
Zero	
Nonzero	
Dividends in AGI	1, 2, 3
Zero	
Nonzero	
1972 Social Security Taxable Earnings	1, 2, 3
Zero	
Nonzero	
Schedule D	1
Present	
Absent	
AGI Range	
± 10%, with a minimum of ±$500	1, 2, 3, 4
± 20%, with a minimum of ±$1,000	5
± 30%, with a minimum of ±$2,000	6, 7, 8

a/ Not used for all records.

each EM record.[22/] Thus, that sampling error problem still existed after those steps had been completed.

Because better estimates for high-income groups and for aggregates were desired, it was decided to add more high-income AF returns to the statistical match file in another statistical match. As described earlier, the AF was highly stratified by size of AGI and thus contained far more high-income records than the EM did. After examining the results of the initial match and rematch, it was decided that the estimates would be improved substantially by adding more AF records at $30,000 AGI (absolute value) and above. There were 1,201 EM records (less than 3 percent of the EM) and 26,414 AF records used in the high-income match. For those EM records, the matches from the high-income match replaced the matches from the initial match or rematch. The matching method chosen was basically a constrained one, unlike the initial match and rematch which were unconstrained. A constrained method was chosen in order to preserve the high-income AF information. The high-income match was carried out using cells and ranking of records in both files within those cells. The AF records were reweighted, sample weights of records in both files were split, and records were duplicated, as discussed below.

The cells used are shown in table 6. The constrained matching method required that the weighted number of records in each cell must be equal in the two files. The AF records were reweighted slightly in order to accomplish this.

Table 6.--Cells Used in the High-Income Match

Cell No.	Return Type	Sign of AGI	Race	Presence of Social Security Taxable Earnings	Age	Sex	Number of Dependent Exemptions
				Absolute Value of AGI $30,000-$49,999			
1	Joint	Positive	White	Present	65+	-	-
2	"	"	"	"	40-64	-	0-1
3	"	"	"	"	"	-	2-3
4	"	"	"	"	"	-	4+
5	"	"	"	"	<40	-	-
6	"	"	"	Absent	65+	-	-
7	"	"	"	"	40-64	-	-
8	"	"	"	"	<40	-	-
9	"	"	Black and Other	Present	-	-	-
10	"	"	"	Absent	-	-	-
11	Nonjoint	"	-	Present	65+	-	-
12	"	"	-	"	<65	Male	-
13	"	"	-	"	<65	Female	-
14	"	"	-	Absent	65+	-	-
15	"	"	-	"	<65	-	-
16	-	Negative	-	-	-	-	-
				Absolute Value of AGI $50,000-$99,999			
1	Joint	Positive	White	Present	65+	-	-
2	"	"	"	"	40-64	-	0-1
3	"	"	"	"	"	-	2-3
4	"	"	"	"	"	-	4+
5	"	"	"	"	<40	-	-
6	"	"	"	Absent	65+	-	-
7	"	"	"	"	<65	-	-
8	"	"	Black and Other	-	-	-	-
9	Nonjoint	"	-	Present	65+	-	-
10	"	"	-	"	<65	-	-
11	"	"	-	Absent	65+	-	-
12	"	"	-	"	<65	-	-
13	-	Negative	-	-	-	-	-

Table 6.--Cells Used in the High-Income Match--Continued

Cell No.	Return Type	Sign of AGI	Number of Age and Blind Exemptions	Presence of Wage and Salary Income (IRS)
		Absolute Value of AGI $100,000 and Over		
1	Joint	Positive	0	Absent
2	"	"	0	Present
3	"	"	1+	Absent
4	"	"	1+	Present
5	Nonjoint	"	0	-
6	"	"	1+	-
7	-	Negative	-	-

Within each cell, EM and AF records were ranked by size of AGI, and the sample weights were cumulated for each file. Then EM and AF records of equal rank were matched, with sample weights being split whenever a cumulated weight amount in either file was reached. For example, assume that a cell contained two EM records (a and b, ranked in that order), each with a weight of 1500, and three AF records (I, II, and III, ranked in that order), each with a weight of 1000. Then the cumulated weights would be 1500 and 3000 in the EM, and 1000, 2000, and 3000 in the AF, and the match would produce the following four matched records: a-I (with a weight of 1000); a-II (weight = 500); b-II (weight = 500); b-III (weight = 1000).[23/]

G. Effects on the Size Distribution of Income

In this section the CPS size distribution of total money income for all family units[24/] is compared with what will be called the AF-CPS size distributions of total money income and total money income minus total Federal income tax.[25/] The effects of adjusting the estimate of total money income by replacing CPS amounts with tax return amounts for several income types and subtracting estimates of total income tax, both from the statistically matched in data, are examined. The AF-CPS estimates incorporate the initial match, the rematch, and the high-income match.

The comparison between the two before-tax distributions can be interpreted as a comparison between an original estimate and a "corrected" estimate, or just as a sensitivity analysis. The after-tax distribution is an example of an estimate which can be produced only by adding one or more variables to an existing file, either by statistical matching or by some other technique.

The CPS estimates shown here are not identical to the published estimates (U.S. Bureau of the Census 1973b) primarily because the sample weights are different.[26/]

For AF-CPS income, SOI amounts replace CPS amounts for wages and salaries, nonfarm business and partnership income, and property income for filers of tax returns (and their spouses in the case of a joint return). Thus, AF-CPS total money income consists of the following income types for those persons: (1) SOI wages and salaries; (2) SOI net income from nonfarm unincorporated business or partnership; (3) CPS net income from farm self-employment; (4) SOI property income (interest, dividends, rent, royalty, estate and trust); (5) CPS social security and railroad retirement benefits; (6) CPS public assistance; (7) CPS other government transfer payments (unemployment compensation, workmen's compensation, government pensions, veterans' benefits); (8) CPS other income (private pensions and annuities, alimony, contributions from persons outside the household, miscellaneous types). For nonfilers of tax returns, all income types come from the CPS.

SOI amounts are used in place of CPS amounts for types (1), (2), and (4) because the SOI estimates of those types are generally considered to be more accurate than the CPS estimates of those types. It is assumed that the resulting estimate of the distribution of income is more accurate than the CPS estimate, despite the inaccuracies on a single observation level produced by statistical matching. As noted earlier, those who are not inclined to accept this assumption can view

the comparison between the CPS and AF-CPS estimates as a sensitivity analysis. CPS amounts are used for farm income, type (3), because the SOI estimate, which contains many more losses and a much lower aggregate than the CPS estimate, is considered to be less appropriate for these comparisons. CPS amounts are used for the other four income types because, in general, those types are not fully included on tax returns.

As its name suggests, AF-CPS total money income minus total income tax is obtained by subtracting total Federal individual income tax from AF-CPS total money income. Federal individual total income tax consists of income tax after credits plus additional tax for tax preferences ("minimum tax") (Internal Revenue Service 1974). The tax amounts, along with the income amounts, are before audit.

CPS and AF-CPS aggregates, along with independently estimated control aggregates derived from the National Income and Product Accounts, are shown in table 7. As a result of replacing the three CPS income types with SOI types, AF-CPS total money income exceeds CPS income by roughly $35 billion, or 4.5 percent of the CPS aggregate. Thus, mean income also increased by 4.5 percent. This increase was the net result of a $24 billion increase in wages and salaries, a $3.5 billion decrease in nonfarm self-employment, and a $15 billion increase in property income.[27/] Even though the SOI nonfarm self-employment aggregate is lower than the CPS aggregate, the SOI distribution is considered here to be more accurate--a higher aggregate is not always assumed to imply a better estimate. AF-CPS total

Table 7.--Income Aggregates, 1972

(Billions of Dollars)

Item	Control Aggregate a/	CPS		AF-CPS	
		Aggregate	Percent of Control	Aggregate	Percent of Control
Wages and Salaries.....	$621.8	$607.1	98	$631.1	101
Nonfarm Self-Employment..	57.9	52.6	91	49.1	85
Farm Self-Employment.....	16.3	10.6	65	10.6	65
Property.....	80.6	34.2	42	49.4	61
Social Security and Railroad Retirement Benefits	39.8	37.8	95	37.8	95
Public Assistance.....	10.9	7.9	72	7.9	72
Other Government Transfers.....	27.7	19.3	70	19.3	70
Other Income.....	20.9 b/	14.2	68	14.2	68
Total Money Income.....	875.9 b/	783.8	89	819.4	94
Total Income Tax.....	91.7	-	-	91.1	99
Total Money Income minus Total Income Tax.....	784.2	-	-	728.3	93

a/ The controls for income types were preliminary estimates obtained from the Bureau of Economic Analysis, Department of Commerce, with some adjustments by Thomas Petska of the Office of Research and Statistics. The total income tax control was constructed by Thomas Petska.

b/ This control excludes alimony, child support, and regular contributions from persons outside the household. Satisfactory controls for those types are not available. Approximately $2-1/2 billion of those types was reported in the CPS and is included in the CPS and AF-CPS aggregates.

money income is still only 94 percent of the control; all income types except wages and salaries are below their controls. Subtracting the tax aggregate of $91 billion from the AF-CPS aggregate results in a decline in the aggregate (and mean) of 11.1 percent of the AF-CPS amount.

The size distributions using the three estimates are shown in table 8. The AF-CPS distribution shows more family units in all size classes above $11,999, while the CPS distribution shows more units in the classes from $0 to $11,999. Those differences are not unexpected, given the differences in mean amounts. The AF-CPS distribution shows more units with negative income because tax returns show far more (and larger) negative amounts of nonfarm self-employment and property incomes than the CPS does.

AF-CPS after-tax income shows the expected differences from the AF-CPS before-tax distribution--a substantially lower distribution. A word of caution about the after-tax distribution and the increase in the number of units with negative income is in order. Total income tax inlcudes tax liabilities on some income types which are not included in AF-CPS total money income; perhaps the most important example is income from capital gains. Thus, a unit which had income only from capital gains would have zero AF-CPS total money income and negative AF-CPS total money income minus total income tax if it had tax liability on those capital gains.

The relative distributions using the three estimates appear in table 9. The AF-CPS before-tax distribution shows more inequality than the CPS distribution--the AF-CPS share is higher for the top

Table 8.--Distribution of All Family Units, 1972

(Percent)

Size of Income	Estimate of Total Income		
	CPS Total Money Income	AF-CPS Total Money Income	AF-CPS Total Money Income minus Total Income Tax
Negative.....	.2	.3	.4
$0-$1,999.....	9.4	9.0	9.1
$2,000-$3,999.....	13.0	12.6	13.1
$4,000-$5,999.....	11.1	11.0	12.0
$6,000-$7,999.....	10.5	10.2	11.8
$8,000-$9,999.....	10.3	10.0	11.2
$10,000-$11,999....	9.7	9.3	10.4
$12,000-$13,999....	8.2	8.4	8.6
$14,000-$15,999....	6.9	7.1	6.7
$16,000-$17,999....	5.0	5.3	4.7
$18,000-$19,999....	3.9	4.1	3.3
$20,000-$24,999....	5.8	6.1	4.6
$25,000-$29,999....	2.7	2.9	1.9
$30,000-$49,999....	2.5	2.8	1.8
$50,000 and over...	.6	.8	.4
Total.....	100.0	100.0	100.0
Mean Income.....	$10,795	$11,286	$10,031

Table 9.--Shares of Aggregate Income, All Family Units, 1972

(Percent of total income)

Deciles	Estimate of Total Income		
	CPS Total Money Income	AF-CPS Total Money Income	AF-CPS Total Money Income minus Total Income Tax
Bottom.....	1.0	.9	.7
2.....	2.6	2.5	2.8
3.....	4.1	4.0	4.4
4.....	5.7	5.7	6.0
5.....	7.5	7.4	7.7
6.....	9.3	9.2	9.5
7.....	11.2	11.1	11.4
8.....	13.6	13.4	13.6
9.....	16.9	16.7	16.7
Top.....	28.1	29.1	27.2
All units....	100.0	100.0	100.0

decile and lower for all the other deciles. The AF-CPS after-tax distribution shows higher shares than the before-tax AF-CPS distribution for deciles two through eight, and a substantially lower share for the top decile. The decline in the share of the bottom quintile is related to the problem mentioned above of tax liabilities on income types not included in AF-CPS total money income.

VII. SUMMARY AND CONCLUSIONS

Statistical matching has been used in several cases to improve data on the size distribution of income, either by adding more variables to a file or by adding what are considered to be more accurate values for variables already present, or both. The statistical matching work carried out at the Office of Research and Statistics, Social Security Administration, with the cooperation of the Bureau of Economic Analysis, U.S. Department of Commerce, which is described in this paper includes both of these types of data improvement.

Using the Office of Research and Statistics example, three alternative estimates of the size distribution of total money income of family units are shown--an original CPS distribution, a combination of CPS and statistically matched in tax return amounts, and a distribution after statistically matched in Federal income tax liabilities were subtracted.

Although statistical matching has been used for more than a decade, not very much is known about the accuracy of such matches. Despite criticisms of statistical matching on a theoretical level (e.g., Sims 1972), there is some evidence (Ruggles, Ruggles, and Wolff 1977) that,

at least for some purposes and under some conditions, statistical matching can produce reasonable and useful estimates. An example using 1971 social security taxable earnings which appears in the appendix to this paper is further evidence that statistical matching can produce useful estimates.[28/] Other material included in the appendix provides more indirect evidence about the accuracy of the Office of Research and Statistics match and about the sensitivity of the results to the specification of the match.

My own conclusion is that statistical matching, when properly applied, is sufficiently accurate for many purposes, but that we need to learn much more about the limits to and the factors affecting the accuracy of statistical matching under various sets of conditions.

APPENDIX

THE "ACCURACY" OF THE OFFICE OF RESEARCH AND STATISTICS STATISTICAL MATCH

There are two questions about the results of a statistical match which are natural to ask: (1) How close are the values of matching variables from the two files in the matched records? (2) How accurately were the nonmatching variables of interest added? While the second question perhaps is more interesting, it is much more difficult to answer. Both of those questions are discussed briefly in this appendix. In most cases, results are shown for the initial match, the rematch and the high-income match. Thus the sensitivity of the results to the different specifications of those steps can be examined.[29/]

Matching Variables

In the discussion of the results for matching variables, both "net error" and "gross error" are examined. Net error refers to differences between EM and AF distributions and aggregates, allowing offsetting errors. Gross error refers to differences between EM and AF values in matched records, not allowing offsetting errors. Essentially, the EM estimates are assumed to be the "truth" in these comparisons, although the high-income match estimates are also compared to estimates from the full AF.[30/]

Eight matching variables which were not in the form of income amounts are discussed first (number of taxpayers, sex, race, type of 1972 social security taxable earnings, marital status, number of dependent and age and blind exemptions, and age). Net error for those

variables in general was quite small for the initial match, the rematch, and the high-income match. There was a tendency for the AF numbers for large groups (e.g., males, whites) to be slightly overestimated and for small groups to be slightly underestimated.

When gross error for these variables is examined, the percent of all records in which EM and AF values are equal (or in the same class) is very high except for age (table A-1). The percents for age would be expected to be relatively low because age is shown in classes and EM and AF values can be quite close but can still fall in different classes.

Net error for the six income amount matching variables is examined using numbers of recipients, aggregates, and mean incomes (table A-2). AGI, net AGI, wages and salaries, and 1972 social security taxable earnings show few differences either among the three steps or compared to the full AF estimates. Dividends in AGI and interest show larger differences. For those types, the AF aggregates from the matching steps are below the EM and full AF aggregates. The results for numbers of recipients in the rematch and high-income match are close to the EM and full AF estimates.

Gross error is examined using percent in the same size class and mean difference as a percent of the EM value for the six income variables discussed above and percent equal for five other variables showing presence of a specific schedule (table A-3). In general, percent in the same class is very high for all types except dividends in AGI and interest. For all records, the dividend percents

Table A-1.--Percent of Matched Records with EM and AF Values Identical

Matching Variable	Initial Match	Rematch	High-Income Match
Number of Taxpayers (2 categories)	100.0	99.8	99.8
Sex (2 categories)	100.0	98.8	98.7
Race (3 categories)	99.9	97.3	97.4
Type of 1972 Social Security Taxable Earnings (4 categories)	99.1	96.4	95.8
Martial Status (6 categories)	99.9	97.0	96.9
Number of Dependent Exemptions (10 categories)	98.4	89.3	88.6
Number of Age and Blind Exemptions (5 categories)	98.2	97.1	97.1
Age (8 classes)	68.5	66.9	66.6

Table A-2.--EM and AF Numbers of Recipients, Aggregates, and Means for Matching Variables

Matching Variable		Initial Match		Rematch		High-Income Match		Full AF	Statistics of Income
		EM	AF	EM	AF	EM	AF		
Total No. of Returns (in millions).....		74.5	74.5	74.6	74.6	75.0	75.0	76.5	77.6
Adjusted Gross Income.....	No. of Recip. a/	74.4	74.5	74.5	74.5	74.8	74.8	76.4	-
	Aggregate b/	723.9	720.7	724.6	723.8	723.5	722.2	738.5	746.0
	Mean c/	9,728	9,678	9,729	9,717	9,668	9,649	9,663	-
Net Adjusted Gross Income.	No. of Recip.	74.5	74.5	74.6	74.6	75.0	74.9	76.4	-
	Aggregate	575.4	573.2	576.0	578.0	575.3	576.6	584.3	579.5
	Mean	7,721	7,697	7,722	7,750	7,675	7,696	7,645	-
Size of 1972 Social Security Taxable Earnings.....	No. of Recip.	64.2	64.3	64.2	64.7	64.5	64.8	65.7	-
	Aggregate	374.5	378.7	374.9	382.3	375.9	382.7	388.1	-
	Mean	5,835	5,885	5,838	5,914	5,832	5,904	5,902	-
Wages and Salaries.....	No. of Recip.	67.2	67.7	67.2	67.5	67.5	67.8	69.2	70.0
	Aggregate	600.4	607.4	600.8	609.0	599.8	605.9	617.2	622.6
	Mean	8,938	8,974	8,940	9,022	8,884	8,935	8,918	8,891
Dividends in AGI.....	No. of Recip.	7.3	6.3	7.3	7.1	7.3	7.1	7.5	7.6
	Aggregate	15.2	13.7	15.2	13.7	14.9	14.4	16.6	16.8
	Mean	2,081	2,166	2,081	1,933	2,057	2,036	2,223	2,215
Interest.....	No. of Recip.	34.7	34.0	34.8	34.8	35.0	35.0	35.5	35.7
	Aggregate	25.6	23.0	25.7	23.8	25.7	24.0	27.1	27.4
	Mean	738	677	740	685	733	685	764	768

a/ Millions of returns.
b/ Billions of dollars.
c/ Means of nonzero amounts, in dollars.

Table A-3.--Percent of Matched Records with EM and AF Values in the Same Class and Mean Difference as a Percent of EM Mean

Matching Variable	All Records			EM Records with a Nonzero Amount		
	Initial Match	Rematch	High-Income Match	Initial Match	Rematch	High-Income Match
	PERCENT IN THE SAME CLASS a/					
Presence of Schedule C	98.0	99.9	99.0	83.5	99.9	94.1
Presence of Schedule E	97.6	99.9	98.8	87.0	99.9	95.5
Presence of Schedule D	97.8	96.9	95.9	82.6	80.8	75.9
Presence of Schedule SE	98.6	99.9	98.9	87.4	99.7	92.8
Presence of Schedule F	97.0	99.7 b/	99.4 b/	56.7	93.8 b/	90.8 b/
Adjusted Gross Income (19 classes) c/	92.2	93.8	94.0	92.3	93.8	94.0
Net Adjusted Gross Income (19 classes) c/	91.3	88.7	88.9	91.3	88.7	88.9
Size of 1972 Social Security Taxable Earnings (13 classes) c/	82.3	78.6	78.5	79.9	76.4	76.2
Wages and Salaries (14 classes) c/	84.9	84.9	84.3	85.3	84.4	84.0
Dividends in AGI (12 classes) c/	94.1	93.6	92.8	45.6	39.8	36.1
Interest (12 classes) c/	67.0	70.6	70.4	37.1	38.6	38.6
	MEAN DIFFERENCE AS A PERCENT OF EM MEAN d/					
Adjusted Gross Income	3.3	2.8	3.7	3.3	2.7	3.7
Net Adjusted Gross Income	4.2	5.2	6.3	4.2	5.2	6.3
Size of 1972 Social Security Taxable Earnings	7.4	10.5	10.9	7.1	9.3	9.8
Wages and Salaries	9.3	10.4	14.2	8.3	9.6	12.2
Dividends in AGI	31.1	77.7	111.1	29.1	73.8	96.5
Interest	84.0	93.0	95.6	79.7	90.7	92.8

a/ Computed using weighted numbers of records.
b/ After imputation of presence of Schedule F to some EM records.
c/ Income size classes were used.
d/ Computed using unweighted records for the initial match and rematch and weighted numbers for the high-income match.

are also very high. Mean difference as a percent of the EM mean is also small for all types except dividends in AGI and interest. It should be noted that many amounts of dividends and especially of interest are quite small; thus, the mean difference is not as large as a first glance at the table might suggest.

For the other five variables, the percents equal are quite high--usually higher for all records than for EM records with the schedule present. In some cases, differences among the results for the three steps are quite large, as would be expected from the differences in the specifications for those steps.

Nonmatching Variables

Examination of the accuracy with which nonmatching variables of interest were added in, of course, is much more difficult--e.g., ordinarily we do not know what the data from an exact match carried out without error would be. Thus, gross error ordinarily cannot be examined for those variables. However, net error can be examined, at least to some extent, using comparisons with the full AF and the SOI. When examining the comparisons, it should be noted that the full AF and SOI populations are slightly larger than the population represented in the EM-AF file.

For nonmatching AF tax return variables, net error is examined using numbers of returns showing an amount, aggregates, and means (table A-4). In general, the estimates were quite close to the full AF estimates, especially after the high-income match. Numbers of recipients of several income types were raised substantially in the rematch. After the high-income match, taxable pensions and annuities

Table A-4.--Numbers of Records, Aggregates, and Means for Selected AF Nonmatching Variables

Variable		Initial Match	Rematch	High-Income Match	Full AF	Statistics of Income d/
Total Number of Returns (millions)		74.5	74.6	75.0	76.5	77.6
Total Income Tax.....	No. of Returns a/	58.7	58.8	59.2	60.1	60.9
	Aggregate b/	91.1	91.0	91.1	92.7	93.6
	Mean c/	1,552	1,548	1,540	1,542	1,537
Total Deductions.....	No. of Returns	74.1	74.2	74.5	76.0	77.1
	Aggregate	160.4	162.3	162.1	164.7	166.4
	Mean	2,165	2,189	2,175	2,166	2,158
Taxes Paid Itemized Deduction....	No. of Returns	27.2	27.2	27.1	26.6	26.8
	Aggregate	36.7	37.4	36.7	35.9	36.2
	Mean	1,351	1,372	1,355	1,351	1,348
Business Income (Schedule C).....	No. of Returns	5.9	6.6	6.5	6.5	6.7
	Aggregate	33.9	34.1	33.7	33.5	34.5
	Mean	5,742	5,188	5,187	5,166	5,151
Rent.....	No. of Returns	5.6	6.1	6.1	6.2	6.3
	Aggregate	2.0	1.9	3.2	3.2	3.0
	Mean	358	309	528	524	472
Partnership Income.....	No. of Returns	1.9	2.1	2.1	2.1	2.2
	Aggregate	10.2	13.9	11.2	11.0	11.1
	Mean	5,372	6,620	5,372	5,150	4,997
Farm Income (Schedule F).....	No. of Returns	2.4	3.0	3.1	2.7	2.8
	Aggregate	4.4	4.4	4.4	4.0	4.1
	Mean	1,827	1,451	1,443	1,486	1,472
Royalty Income.....	No. of Returns	.4	.5	.5	.5	.5
	Aggregate	1.3	.6	.8	.9	.9
	Mean	2,823	1,161	1,567	1,673	1,741
Estate and Trust Income.....	No. of Returns	.6	.6	.6	.7	.7
	Aggregate	1.2	1.2	1.4	1.8	1.8
	Mean	2,200	1,947	2,325	2,543	2,635

Table A-4.--Numbers of Records, Aggregates, and Means for Selected AF Nonmatching Variables--Continued

Variable		Initial Match	Rematch	High-Income Match	Full AF	Statistics of Income d/
Net Capital Gains.....	No. of Returns	7.7	8.2	8.3	8.8	8.9
	Aggregate	14.8	13.6	15.9	16.6	17.1
	Mean	1,925	1,653	1,924	1,877	1,926
Small Business Corporation Income.	No. of Returns	.4	.4	.5	.5	.5
	Aggregate	2.2	1.5	2.1	2.3	2.1
	Mean	5,358	3,408	4,537	4,971	4,259
Taxable Pensions and Annuities....	No. of Returns	3.0	3.0	3.1	3.7	3.7
	Aggregate	9.0	8.7	8.9	10.8	11.0
	Mean	2,965	2,877	2,900	2,927	2,950

a/ Millions of returns.
b/ Billions of dollars.
c/ Mean for nonzero amounts, in dollars.
d/ Internal Revenue Service (1974).

and capital gains show the largest differences compared to the full AF. It should be noted that although these differences in some cases are substantial relative to the full AF estimates of the particular income type, the differences are very small relative to all returns and aggregate total income.

Two other estimates which utilize the estimated amounts of Federal total income tax will also be mentioned: mean income tax by size of AGI, and the distribution of AGI minus total income tax. Estimates of mean income tax by size of AGI were not sensitive to the different matching steps or to whether the EM or AF amount of AGI was used (table A-5). All estimates were quite close to the full AF and SOI estimates except in the high-AGI classes. In those classes the high-income match estimates were very close to the full AF and SOI estimates.

The size distribution of AGI minus total income tax was also not very sensitive to the matching steps or to whether the EM or AF amount of AGI was used, and the match estimates were quite close to the full AF estimate (table A-6).

Two other estimates involving nonmatching variables will also be discussed. One of the fundamental questions raised about statistical matching is how well it estimates the joint distributions of nonmatching variables of interest in the two files.[31/] Here we can examine this question, but only for 1971 social security taxable earnings. That variable appeared in both the EM and AF, but was not used as a matching variable. We can compare the estimate from the AF portion of the matched file with the "true" estimate using the EM amounts. The size distributions of 1971 social security taxable earnings from the EM and the AF in the initial match are shown for six years of school completed groups in table A-7.

Table A-5.--Mean Total Income Tax by Size of AGI for All Returns

(Dollars)

Size of AGI	Initial Match		Rematch		High-Income Match		Full AF	Statistics of Income
	EM AGI	AF AGI	EM AGI	AF AGI	EM AGI	AF AGI		
Negative or Zero....	0	0	0	0	9	9	12	28
$1-$1,999.....	1	1	1	1	1	1	1	1
$2,000-$3,999.....	83	84	85	85	87	87	85	85
$4,000-$5,999.....	312	316	316	319	322	325	312	311
$6,000-$7,999.....	586	586	594	593	600	599	576	577
$8,000-$9,999.....	850	849	849	847	859	856	842	839
$10,000-$14,999.....	1,300	1,306	1,304	1,309	1,304	1,309	1,302	1,304
$15,000-$19,999.....	2,102	2,112	2,113	2,118	2,111	2,116	2,142	2,146
$20,000-$24,999.....	3,137	3,153	3,150	3,160	3,145	3,156	3,165	3,173
$25,000-$29,999.....	4,296	4,355	4,313	4,349	4,306	4,307	4,317	4,325
$30,000-$49,999.....	6,904	7,097	7,045	7,093	7,121	7,070	7,095	7,048
$50,000-$99,999.....	17,544	17,879	17,751	17,843	17,595	17,595	17,640	17,632
$100,000 and over...	77,098	78,504	70,345	72,112	72,243	72,243	71,737	71,443
Total.....	1,223	1,223	1,220	1,220	1,216	1,216	1,212	1,206

Table A-6.--Size Distribution of AGI Minus Total Income Tax

(Percent Distributions)

Size of AGI Minus Total Income Tax	Initial Match		Rematch		High-Income Match		Full AF
	EM AGI	AF AGI	EM AGI	AF AGI	EM AGI	AF AGI	
Negative or Zero.....	.5	.5	.6	.5	.6	.6	.5
$1-$2,499.....	18.6	18.7	18.5	18.6	18.4	18.5	18.5
$2,500-$4,999.....	17.3	17.3	17.3	17.3	17.8	17.8	17.9
$5,000-$7,499.....	16.0	16.1	15.9	16.1	16.0	16.2	16.2
$7,500-$9,999.....	14.5	14.1	14.5	14.1	14.4	14.0	14.0
$10,000-$14,999.....	20.1	20.2	20.1	20.3	20.0	20.2	19.9
$15,000-$19,999.....	7.9	8.0	7.9	7.9	7.8	7.8	7.9
$20,000-$24,999.....	2.6	2.6	2.5	2.6	2.5	2.4	2.6
$25,000-$29,999.....	1.1	1.0	1.1	1.0	1.1	1.0	1.0
$30,000-$39,999.....	.8	.8	.8	.8	.7	.8	.8
$40,000-$49,999.....	.3	.3	.3	.3	.3	.3	.3
$50,000-$74,999.....	.3	.2	.3	.3	.3	.2	.3
$75,000-$99,999.....	.1	.1	.1	.0	.0	.1	.1
$100,000 and over.....	.1	.1	.1	.1	.1	.1	.1
Total.....	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Number of units (thousands).....	74,520	74,520	74,587	74,587	74,960	74,960	76,470
Mean.....	$8,491	$8,476	$8,495	$8,483	$8,437	$8,440	$8,450

Table A-7.--Percent Distribution of 1971 Social Security Taxable Earnings by Years of School Completed

Size of 1971 Social Security Taxable Earnings	Total		Years of School Completed (CPS)					
			8 years or less		9-11 years		12 years	
	EM Earnings	AF Earnings	EM Earnings	AF Earnings	EM Earnings	AF Earnings	EM Earnings	AF Earnings
Zero	18.0	18.2	22.3	22.2	24.3	23.5	13.9	14.5
$1-$999	9.6	9.3	6.3	6.4	14.0	13.7	8.6	8.5
$1,000-$1,999	7.4	7.6	7.7	7.5	6.7	7.8	6.9	7.4
$2,000-$2,999	5.6	5.7	5.6	5.8	4.8	5.4	5.6	5.8
$3,000-$3,999	5.5	5.7	6.6	7.4	5.0	5.3	6.2	6.2
$4,000-$4,999	5.6	5.4	8.2	6.4	5.1	5.1	6.3	6.1
$5,000-$5,999	5.8	6.0	7.2	7.4	5.7	5.2	6.8	6.8
$6,000-$6,999	5.9	6.1	7.4	7.2	6.2	5.7	6.7	6.9
$7,000-$7,799	4.8	4.9	5.6	6.2	4.5	4.7	5.9	5.6
$7,800	25.3	24.6	18.6	19.0	18.0	18.1	26.2	25.2
$7,801-$8,999	3.6	3.5	2.6	2.6	2.9	2.8	4.2	3.9
$9,000-$11,999	2.1	2.2	1.5	1.4	2.1	2.0	1.9	2.4
$12,000 and over	.8	.8	.4	.5	.6	.6	.7	.6
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Mean	$4,392	$4,374	$4,035	$4,082	$3,666	$3,662	$4,723	$4,644
Number of Units (in thousands)	74,520		11,296		13,323		27,209	

Table A-7.--Percent Distribution of 1971 Social Security Taxable Earnings by Years of School Completed -- Continued

Size of 1971 Social Security Taxable Earnings	Years of School Completed (CPS)					
	13-15 years		16 years		17 years or more	
	EM Earnings	AF Earnings	EM Earnings	AF Earnings	EM Earnings	AF Earnings
Zero	15.0	17.2	17.9	16.7	21.9	20.1
$1-$999	12.7	12.4	7.2	6.4	5.8	4.4
$1,000-$1,999	10.9	9.8	5.5	5.6	4.9	4.7
$2,000-$2,999	7.7	7.2	4.9	4.7	2.8	3.3
$3,000-$3,999	5.7	5.1	3.4	4.5	2.6	2.7
$4,000-$4,999	5.0	5.0	3.3	3.6	1.4	2.8
$5,000-$5,999	4.8	5.6	3.8	4.2	2.2	3.5
$6,000-$6,999	4.9	5.2	3.2	4.8	2.6	3.4
$7,000-$7,799	3.6	3.7	3.2	4.2	3.0	2.7
$7,800	23.5	22.5	38.8	36.2	43.1	42.2
$7,801-$8,999	3.2	3.6	4.3	4.0	4.1	4.9
$9,000-$11,999	2.1	2.1	2.7	3.2	3.7	3.2
$12,000 and over	.8	.6	1.7	1.6	1.9	2.1
Total	100.0	100.0	100.0	100.0	100.0	100.0
Mean	$4,115	$4,058	$5,021	$5,130	$5,198	$5,325
Number of Units (in thousands)	11,786		6,404		4,501	

The EM and AF total distributions are quite close. Given that, we can say that the AF distribution for each education category would resemble the AF distribution for all units if the statistical match did not capture any of the relationship (e.g., if the match were random). It can be seen that the AF distribution for each education category resembles the EM distribution for that category far more than it resembles the AF distribution for all units. Whether the estimates are "close enough" depends upon the use to which they would be put. Of course, these results are not necessarily representative of the general accuracy of statistical matching.

It is also useful to use 1971 social security taxable earnings to examine the accuracy of one example of the type of estimate which might be made from a statistically matched file of this kind. Again, one nonmatching variable from each file is used, the EM estimate is assumed to be the "true" estimate, and data from the initial match are used. Estimated 1971 social security employee tax as a percent of 1972 CPS total person income was chosen as the example (table A-8).[32/] The amount of tax was estimated in a crude way by using 5.2 percent of the 1971 EM and AF social security taxable earnings in each record.[33/] The EM and AF estimates are very close--except for the negative and $50,000 and over classes, which are very small, the estimates differ by no more than 0.1 percent. My conclusion is that, at least in this case, the estimate from the statistical match is more than adequate.

Table A-8.--Estimated 1971 Social Security Payroll Tax as a Percent of CPS 1972 Person Income

Size of 1972 CPS Person Income	EM Estimate	AF Estimate
Negative	5.7 a/	6.1 a/
$1-$999	8.1	8.1
$1,000-$1,999	4.7	4.7
$2,000-$2,999	3.7	3.8
$3,000-$3,999	3.6	3.6
$4,000-$4,999	3.6	3.7
$5,000-$5,999	3.7	3.8
$6,000-$6,999	3.8	3.8
$7,000-$7,999	3.7	3.7
$8,000-$9,999	3.5	3.4
$10,000-$11,999	3.0	3.0
$12,000-$14,999	2.6	2.6
$15,000-$19,999	2.0	2.0
$20,000-$24,999	1.4	1.4
$25,000-$49,999	.9	.9
$50,000 and over	.5	.9
Total	2.6	2.6

a/ Estimated tax as a percent of the absolute value of CPS income.

FOOTNOTES

1/ At the present time, developmental work is under way at the Department of Health and Human Services and the Bureau of the Census on the Survey of Income and Program Participation, a household survey which is planned to collect detailed income data (Ycas 1979). Although some preliminary interviewing has already been done, that survey is not expected to become a regular source of size distribution data for several more years.

2/ See Budd, Radner, and Hinrichs (1973) for an example of the effects of correcting income tax data for audit.

3/ Of course, it is often difficult to specify which data source is "best" for a particular income type, especially when possible inaccuracies in matching are taken into account. Also, when estimates of income types from different data sources are combined, the estimated joint distributions of those income types should be used with caution.

4/ See Herriot and Spiers (1976) and several other papers referred to in Kilss and Scheuren(1978) for examples of comparisons carried out using the 1973 CPS-IRS-SSA Exact Match Study.

5/ Other work with that file was presented at the 1980 American Statistical Association meetings by Frederick Scheuren and H. Lock Oh. That work focused on the accuracy of the assignment of amounts to nonrespondents.

6/ Exact matching has been used to construct longitudinal size distribution data from annual data. For example, see David, Gates, and Miller (1974). Data were also augmented in that example.

7/ See Office of Federal Statistical Policy and Standards (1980) or Radner (1979b) for more complete descriptions of the matches mentioned in this section.

8/ Exact matching could not be used because the data consisted of samples with few persons in common. Also, no available data contained personal identifying information.

9/ BEA has recently completed a statistical match between the statistically matched file described in this paper and the 1972 Consumer Expenditure Survey, in connection with making estimates of the size distribution of family personal income.

10/ See Office of Federal Statistical Policy and Standards (1980) or Radner (1979b) for more detailed discussions of statistical matching methods.

11/ In a strict definition of a constrained match, sample weights in the nonbase set before and after matching must be identical; a looser definition allows small changes in the sample weights in the nonbase set, for example, through reweighting prior to the matching. In either case, in a constrained match the nonbase set data are used as a control; in an unconstrained match the nonbase set is not used as a control, but merely as a population to be drawn from. The "high-income" match described in the following section is a constrained match only under the looser definition.

12/ See Scheuren et al. (1975, pp. 102-3) for the definition of a good CPS-SER exact match.

13/ The only tax return file which could be used in the construction of the EM was a file which contained only a few items from the return. Those items appear in table 1.

14/ A good pair of matching variables is a pair which is defined the same way (or almost the same way) in both files, has the same (or almost the same) error pattern (e.g., reporting error) in both files, and is highly associated with important nonmatching variables.

15/ The initial match and rematch essentially were viewed as approximations of an exact match using the EM as the base. The high-income match basically was not viewed as an approximation of an exact match.

16/ Relative importance was specified without the use of statistical tests.

17/ For some EM records, some AF records within the appropriate cells and ranges were defined to be ineligible for matching to allow for the fact that the AF was a stratified sample. Thus, for some EM records, the AF was subsampled before matching.

18/ After each test run, cross-tabulations of matching variables and totals for several AF nonmatching variables were examined. The weights were changed to try to improve those test results.

19/ The specifications for the rematch were decided upon jointly by the author and Edward Budd, Jean Salter, and Robert Yuskavage of BEA.

20/ The rematch also included 84 records which had not been matched in the initial match because they had not been considered good exact matches at that time.

21/ In order to adjust for an error in the Schedule F indicator in the EM, the presence of Schedule F was imputed to some EM records. This imputation was carried out prior to the consistency check.

22/ For many EM records, two AF matches (a best match and a second best match) were chosen in the initial match. The second match was chosen primarily to be used if the best match was unsatisfactory for any reason. For example, for a few high-income returns, it appeared that the best AF match was also the true match. To avoid confidentiality problems, in those cases the best match was replaced by another AF record, usually the second best AF match.

23/ See Budd, Radner, and Hinrichs (1973, p. 23) for a more detailed description of this type of matching procedure.

24/ A family unit can be a family or an unrelated individual. See U.S. Bureau of the Census (1973b) for definitions of those terms.

25/ A family unit file was constructed from the matched EM-AF file by adding nonfilers to that file and summing the incomes and tax liabilities of all family members.

26/ All estimates shown in this section utilize a family sample weight constructed for the EM file. Only family units which are considered to be good exact matches are included in these estimates.

27/ The SOI wage and salary aggregate from the match exceeds the control aggregate by about $9 billion. Possible explanations for this excess include inaccuracies in the statistical matching, sampling error in the AF-CPS aggregate, and inaccuracies in the control.

28/ Whether statistical matching is the best way of obtaining any given estimate remains an unanswered question.

29/ Results from the rematch and high-income match refer to the entire file, not just records used in those steps.

30/ When discussing the results for matching variables, it should be noted that the high-income match was not meant to be solely an approximation of the EM data. Thus, for that step, differences between EM and statistically matched in values do not necessarily indicate error.

31/ Here we are not discussing the joint distribution conditional on the matching variables, as mentioned by Sims (1972, 1974), but the estimated joint distribution for all units. The latter distribution is more relevant here.

32/ CPS income for 1972 was used because CPS income for 1971 was not available. Although this example is not of analytic interest, it is as close as we could come, with the variables available, to checking something which is of analytic interest.

33/ The same rate was applied to both wages and salaries and self-employment income. That crude procedure was considered to be adequate for the purpose of comparing the EM and AF estimated effective tax rates in this example.

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 Marjorie Odle (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.

Barr, Richard S., and J. Scott Turner (1978). "A New, Linear Programming Approach to Microdata File Merging," in 1978 Compendium of Tax Research sponsored by the Office of Tax Analysis, U.S. Department of the Treasury.

Barr, Richard S., and J. Scott Turner (1980). "Merging the 1977 Statistics of Income and the March 1978 Current Population Survey." Analysis, Research and Computation, Inc., Austin, Texas (July).

Beebout, Harold, Pat Doyle, and Allen Kendall (1976). "Estimation of Food Stamp Participation and Cost for 1977: A Microsimulation Approach (Final Report)." MPR Working Paper #E-48. Mathematica Policy Research, Inc. (July).

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 Daniel B. Radner (1969). "The OBE Size Distribution Series: Methods and Tentative Results for 1964." American Economic Review (May) LIX: 435-49.

Budd, Edward C., and Daniel B. Radner (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., Daniel B. Radner, and John C. Hinrichs (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).

Bureau of Labor Statistics (1978). "Consumer Expenditure Survey: Interview Survey, 1972-73," Bulletin 1997, U.S. Department of Labor.

David, Martin H., William A. Gates, and Roger F. Miller (1974). Linkage and Retrieval of Microeconomic Data. Lexington Books, D.C. Heath and Company.

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

Goldsmith, Selma F. (1958). "Size Distribution of Personal Income." Survey of Current Business (April) 38: 10-19.

Herriot, Roger A., and Emmett F. Spiers (1976). "Measuring the Impact on Income Statistics of Reporting Differences Between the Current Population Survey and Administrative Sources." Proceedings of the 1975 Meetings of the American Statistical Association, Social Statistics Section.

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

Internal Revenue Service (1980). "Internal Revenue Service Preliminary Report: Statistics of Income--1978, Individual Income Tax Returns."

Kilss, Beth, and Fritz Scheuren (1978). "The 1973 CPS-IRS-SSA Exact Match Study: Past, Present, and Future," in Policy Analysis with Social Security Research Files, Research Report No. 52, Office of Research and Statistics, Social Security Administration.

Office of Federal Statistical Policy and Standards (1980). "Report on Exact and Statistical Matching Techniques." Statistical Policy Working Paper 5 (June). U.S. Department of Commerce.

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.

Radner, Daniel B. (1978). "Age and Family Income," in Policy Analysis with Social Security Research Files, Research Report No. 52, Office of Research and Statistics, Social Security Administration.

Radner, Daniel B. (1979a). "Federal Income Taxes, Social Security Taxes, and the U.S. Distribution of Income, 1972." ORS Working Paper No. 7, Office of Research and Statistics, Social Security Administration (April).

Radner, Daniel B. (1979b). "The Development of Statistical Matching in Economics." 1978 Proceedings of the American Statistical Association, Social Statistics Section, 503-8.

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

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

Ruggles, Nancy, Richard Ruggles, and Edward Wolff (1977). "Merging Microdata: Rationale, Practice and Testing." Annals of Economic and Social Measurement (Fall) 6: 407-28.

Scheuren, Frederick, Benjamin Bridges, and Beth Kilss (1973). "Subsampling the Current Population Survey: 1963 Pilot Link Study," Studies from Interagency Data Linkages, No. 1, Office of Research and Statistics, Social Security Administration.

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

Sims, Christopher A. (1972). "Comments." Annals of Economic and Social Measurement (July) 1: 343-46.

Sims, Christopher A. (1974). "Comment." Annals of Economic and Social Measurement (April) 2: 395-8.

Springs, Ricardo, and Harold Beebout (1976). "The 1973 Merged SPACE/AFDC File: A Statistical Match of Data from the 1970 Decennial Census and the 1973 AFDC Survey." Mathematica Policy Research, Inc. (March 31).

Steinberg, Joseph (1973). "Some Observations on Linkage of Survey and Administrative Record Data." Studies from Interagency Data Linkages, Office of Research and Statistics, Social Security Administration.

U.S. Bureau of the Census (1970). "Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Record Check of Accuracy of Income Reporting," Series ER-60, No. 8.

U.S. Bureau of the Census (1973a). Census of Population: 1970, Subject Reports, Final Report PC (2)-8A, "Sources and Structure of Family Income."

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

U.S. Bureau of the Census (1978). Current Population Reports, Series P-60, No. 110, "Money Income and Poverty Status in 1975 of Families and Persons in the United States and the Northeast Region, by Divisions and States (Spring 1976 Survey of Income and Education)."

U.S. Bureau of the Census (1980). Current Population Reports, Series P-60, No. 123, "Money Income of Families and Persons in the United States: 1978."

Wolff, Edward (1977). "Estimates of the 1969 Size Distribution of Household Wealth in the U.S. from a Synthetic Database." Paper presented at the Conference on Research in Income and Wealth, December, Williamsburg, Virginia.

Ycas, Martynas A. (1979). "An Introduction to the Income Survey Development Program," U.S. Department of Health, Education, and Welfare. Mimeographed.